Machine Translated by Google

SPRINGER
REFERENCE

Yadolah Dodge

The Concise Encyclopedia of Statistics

Springer

Machine Translated by Google

Die prägnante Enzyklopädie der Statistik

Machine Translated by Google

Yadolah Dodge

Die prägnante Enzyklopädie der Statistik

Mit 247 Tabellen

123

Machine Translated by Google

Autor

Yadolah Dodge
Honorarprofessor
Universität Neuenburg
Schweiz
yadolah.dodge@uninc.ch

Ein CiP-Katalogeintrag für dieses Buch ist bei der Library of Congress Control erhältlich

ISBN: 978-0-387-32833-1

Diese Publikation ist auch erhältlich als:
Printpublikation unter ISBN 978-0-387-31742-7and
Print- und elektronisches Bundle unter der ISBN 978-0-387-33828-6

Dieses Werk unterliegt dem Urheberrecht. Alle Rechte vorbehalten, unabhängig davon, ob es sich um das gesamte Material oder Teile davon handelt, insbesondere das Recht der Übersetzung, des Nachdrucks, der Wiederverwendung von Abbildungen, der Wiedergabe, der Ausstrahlung, der Wiedergabe auf Mikrofilmen oder auf andere Weise sowie der Speicherung in Datenbanken. Die Vervielfältigung dieser Veröffentlichung oder Teilen davon ist nur im Rahmen der Bestimmungen des Urheberrechtsgesetzes vom 9. September 1965 in der jeweils gültigen Fassung gestattet und bedarf stets der Zustimmung des Springer-Verlags. Zuwiderhandlungen können nach dem deutschen Urheberrecht strafbar sein.

Springer ist Teil von Springer Science+Business Media

springer.com

© 2008 Springer Science + Business Media, LLC.

Die Verwendung eingetragener Namen, Marken usw. in dieser Veröffentlichung bedeutet, auch ohne ausdrückliche Angabe, nicht, dass diese Namen von den einschlägigen Schutzgesetzen und -vorschriften ausgenommen und daher für die allgemeine Verwendung frei sind.

Gedruckt auf säurefreiem Papier SPIN: 10944523 2109–543210

Machine Translated by Google

Zum Gedenken an meine geliebte Frau K, meine

fürsorgliche Mutter,

meinen hart arbeitenden Vater

und

an meine beiden freundlichen und warmherzigen Söhne Ali und Arash

Machine Translated by Google

Vorwort

Wir hoffen, mit diesem prägnanten Band den Bedürfnissen einer großen wissenschaftlichen Gemeinschaft gerecht zu werden, die bisher hauptsächlich durch umfangreiche enzyklopädische Referenzen bedient wurde. Anstatt eine umfassende Berichterstattung über unser Thema anzustreben, haben wir uns auf die wichtigsten Themen konzentriert, diese aber so ausführlich erklärt, wie es der Platz erlaubte. Das Ergebnis ist ein kompaktes Werk, das unserer Überzeugung nach keine zentralen Themen auslässt.

Einträge haben eine starre Struktur, um das Auffinden von Informationen zu erleichtern. Jeder hier vorgestellte Begriff enthält eine Definition, Geschichte, mathematische Details, Einschränkungen bei der Verwendung der Begriffe, gefolgt von Beispielen, Referenzen und relevanter Literatur zur weiteren Lektüre. Die Referenz ist alphabetisch geordnet, um einen schnellen Zugriff auf die grundlegenden Werkzeuge der statistischen Methodik und die Biografien berühmter Statistiker zu ermöglichen, darunter auch einige aktuelle, die weiterhin zur Statistikwissenschaft beitragen, wie etwa Sir David Cox, Bradley Efron und TW Anderson, um nur einige zu nennen ein paar. Die Kriterien für die Auswahl dieser Statistiker, ob lebend oder abwesend, sind natürlich eher persönlicher Natur und es ist durchaus möglich, dass einige dieser berühmten Persönlichkeiten, die einen Eintrag verdienen, abwesend sind. Ich entschuldige mich aufrichtig für solche unbeabsichtigten Auslassungen.

Darüber hinaus wurde versucht, die wesentlichen Informationen über statistische Tests, Konzepte und Analysemethoden in einer Sprache darzustellen, die für Praktiker und Studenten sowie die breite Community, die Statistiken in den Bereichen Medizin, Ingenieurwesen, Physik, Biowissenschaften und Sozialwissenschaften nutzt, zugänglich ist und Betriebswirtschaft/Volkswirtschaft.

Die ersten Schritte zum Schreiben dieses Buches wurden 1983 unternommen. 1993 wurde die erste französischsprachige Version vom Dunod-Verlag in Paris veröffentlicht. Später, im Jahr 2004, wurde die aktualisierte und längere Version auf Französisch von Springer France veröffentlicht und 2007 erschien eine Studentenausgabe der französischen Ausgabe bei Springer.

In dieser Enzyklopädie werden, ebenso wie im *Oxford Dictionary of Statistical Terms, das* 2003 für das International Statistical Institute veröffentlicht wurde, für jeden Begriff ein oder mehrere Verweise angegeben, in einigen Fällen auf eine frühe Quelle, in anderen auf eine neuere Veröffentlichung. Obwohl bei der Auswahl der Referenzen mit einiger Sorgfalt vorgegangen wurde, ist die Festlegung historischer Prioritäten bekanntermaßen schwierig und die historischen Zuordnungen sind nicht als maßgeblich anzusehen.

Weitere Informationen zu Begriffen, die in dieser Enzyklopädie nicht vorkommen, finden Sie in kurzen Artikeln in den folgenden Enzyklopädien und Wörterbüchern:

Machine Translated by Google

International Encyclopedia of Statistics, Hrsg. William Kruskal und Judith M. Tanur (The Freie Presse, 1978).

Encyclopedia of Statistical Sciences, Hrsg. Samuel Kotz, Norman L. Johnson und Cambell Reed (John Wiley and Sons, 1982).

The Encyclopedia of Biostatistics, Hrsg. Peter Armitage und Ted Colton (Chichester: John Wiley und Söhne, 1998).

The Encyclopedia of Environmetrics, Hrsg. AH El-Sharaawi und WW Paregoric (John Wiley und Söhne, 2001).

The Encyclopedia of Statistics in Quality and Reliability, Hrsg. F. Ruggeri, RS Kenett und FW Faltin (John Wiley and Sons, 2008).

Wörterbuch – Encylopédique en Statistique, Yadolah Dodge, Springer 2004

Zwischen der Veröffentlichung der ersten Fassung des aktuellen Buches auf Französisch im Jahr 1993 und der späteren Auflage im Jahr 2004 bis zur aktuellen Fassung wurden am Manuskript zahlreiche Korrekturen vorgenommen. Besonderes Augenmerk wurde auf die Auswahl geeigneter Übersetzungen der Begriffe gelegt, um sowohl in der englischen als auch in der französischen Sprache eine fundierte Bedeutung zu erreichen. Sollte dies in einigen Fällen nicht geschehen sein, entschuldige ich mich. Ich wäre den Lesern sehr dankbar für Kommentare zu Ungenauigkeiten, Korrekturen und Vorschlägen zur Aufnahme neuer Begriffe oder zu anderen Themen, die die nächste Ausgabe verbessern könnten. Bitte senden Sie Ihre Kommentare an den Springer-Verlag.

Ich möchte vielen Menschen danken, die mir in diesen vielen Jahren geholfen haben, dieses Manuskript in seine heutige Form zu bringen. Beginnend mit meinen ehemaligen Assistenten von 1983 bis 2004, Nicole Rebetez, Sylvie Gonano-Weber, Maria Zegami, Jurg Schmid, Severine Pfaff, Jimmy Brignony Elisabeth Pasteur, Valentine Rousson, Alexandra Fragnieire und Theiry Murrier.
Meinen Kollegen Joe Whittaker von der University of Lancaster, Ludevic Lebart von France Telecom und Bernard Fisher von der Universität Marseille für das Lesen von Teilen des Manuskripts. Besonderer Dank geht an Gonna Serbinenko und Thanos Kondylis für ihre bemerkenswerte Zusammenarbeit bei der Übersetzung einiger Begriffe aus der französischen Version Englisch.In Zusammenarbeit mit Thanos, meinem ehemaligen Doktoranden. Student, war eine wunderbare Erfahrung. Tomys Kollege Shahriar Huda, dessen hilfreiche Kommentare, Kritiken und Korrekturen wesentlich zu diesem Buch beigetragen haben. Abschließend danke ich dem Springer-Verlag, insbesondere John Kimmel, Andrew Spencer und Oona Schmid, für ihre akribische Sorgfalt bei der Erstellung dieser Enzyklopädie.

Januar 2008

Yadolah Dodge
Honorarprofessor
Universität Neuenburg
Schweiz

Machine Translated by Google

Über den Autor

Professor Yadolah Dodge gründete 1989 das Master-in-Statistik-Programm der Universität Neuenburg in der Schweiz und erwarb 1970 seinen Master in Angewandter Statistik an der Utah State University und seinen Ph.D. in Statistik mit Nebenfach Biometrie an der Oregon State University im Jahr 1973. Er hat zahlreiche Artikel veröffentlicht und mehrere Bücher in englischer und französischer Sprache verfasst, mitverfasst und herausgegeben, darunter *Mathematical Programming in Statistics* (John Wiley 1981, Classic Edition 1993), *Analysis of Experiments with Missing Data* (John Wiley) . 1985), *Alternative Methoden der Regression* (John Wiley 1993), *Premier Pas en Statistique* (Springer 1999), *Adaptive Regression* (Springer 2000), *The Oxford Dictionary of Statistical Terms* (2003), *Statistique: Dictionnaire encyclopédique* (Springer 2004) und *Optimierung Applikation* (Springer 2005). Professor Dodge ist gewähltes Mitglied des International Statistical Institute (1976) und Fellow der Royal Statistical So

Machine Translated by Google

A

Akzeptanzregion

Der Akzeptanzbereich ist das **Intervall** innerhalb der **Stichprobenverteilung** der **Teststatistik ,** das mit der **Nullhypothese** H0 aus **dem Hypothesentest übereinstimmt.**
Es handelt sich um den Komplementärbereich zum **Ablehnungsbereich.**
Dem Akzeptanzbereich ist eine **Wahrscheinlichkeit** von 1 ÿ ÿ zugeordnet, wobei ÿ das **Signifikanzniveau** des Tests ist.

MATHEMATISCHE ASPEKTE

Siehe **Ablehnungsbereich.**

BEISPIELE

Siehe **Ablehnungsbereich.**

WEITERLESEN

- **Kritischer Wert**
- **Hypothesentest**
- **Ablehnungsbereich**
- **Signifikanzniveau**

Genauigkeit

Die allgemeine Bedeutung von Genauigkeit ist die Nähe eines **Werts** oder einer **Statistik** zu einem Referenzwert. Genauer gesagt misst sie die Nähe des **Schätzers** zum unbekannten **Parameter** ÿ zum wahren Wert von ÿ.

Die Genauigkeit eines **Schätzers** kann anhand des **Erwartungswerts** der quadratischen Abweichung zwischen *T* und ÿ gemessen werden , mit anderen Worten:

$$E\,(T \ddot{y}\ \ddot{y}\,)2\ .$$

Genauigkeit sollte nicht mit dem Begriff Präzision verwechselt werden, der den Grad der Genauigkeit einer Messung angibt und normalerweise durch die Anzahl der Dezimalstellen nach dem Komma angegeben wird ma.

WEITERLESEN

Voreingenommenheit

- **Schätzer**
- **Parameter**
- **Statistiken**

Algorithmus

Ein Algorithmus ist ein Prozess, der aus einer Abfolge genau definierter Schritte besteht, die zur Lösung eines bestimmten Problemtyps führen.
Dieser Prozess kann iterativ sein, das heißt, er wird mehrmals wiederholt. Im Allgemeinen handelt es sich um einen numerischen Prozess.

GESCHICHTE

Der Begriff Algorithmus stammt aus der lateinischen Aussprache des Namens des Mathematikers al-Khwarizmi aus dem 9. Jahrhundert, der in Bagdad lebte und der Vater der Algebra war.

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Wort Algorithmus hat in den letzten Jahren durch das Aufkommen von Computern eine andere Bedeutung bekommen. Im Bereich der Informatik bezeichnet er einen Prozess, der so beschrieben wird, dass er in einem Computerprogramm verwendet werden kann. Das Hauptziel von Statistiksoftware besteht darin, eine Programmiersprache zu entwickeln, die in der Lage ist, statistische Algorithmen zu integrieren, sodass diese Algorithmen dann in einer für den Benutzer verständlichen Form dargestellt werden können. Der Vorteil dieses Ansatzes besteht darin, dass der Benutzer die vom Algorithmus erzeugten Ergebnisse versteht und auf die Präzision der Lösungen vertraut. Zu den verschiedenen statistischen Rezensionen, die sich mit Algorithmen befassen, gehören das *Journal of Algorithms* von der *Aca demic Press* (New York), der Teil des *Journal of the Royal Statistical Society Series C (Applied Statistics) ,* der sich auf Algorithmen konzentriert, und *Computational Statistics* vom *Physica-Verlag* (Heidelberg) und *Random Structures and Algorithms,* herausgegeben von *Wiley* (New York), verdienen besondere Erwähnung.

BEISPIELE

Wir stellen hier einen Algorithmus vor, der den Absolutwert einer Zahl ungleich Null berechnet; mit anderen Worten |x|.
Verfahren:

Schritt 1. Identifizieren Sie das Vorzeichen des angegebene Nummer.

Schritt 2. Wenn das Vorzeichen negativ ist, fahren Sie mit Schritt 3 fort. Wenn das Vorzeichen positiv ist, geben Sie den Absolutwert der Zahl als an Nummer selbst:

$$|x| = x$$

und stoppen Sie den Vorgang.

Schritt 3. Geben Sie den Absolutwert der angegebenen Zahl als Gegenzahl an:

$$|x| = \ddot{y}x$$

und stoppen Sie den Vorgang.

WEITERLESEN

- **Statistiksoftware**
- **Yates-Algorithmus**

VERWEISE

Chambers, JM: Computergestützte Methoden zur Datenanalyse. Wiley, New York (1977)

Khwarizmi, Musa ibn Meusba (9. Jh.). Jabr wa-al-muqeabalah. Die Algebra von Mohammed ben Musa, Rosen, F. (Hrsg. und Übersetzung). Georg Olms Verlag, Hildesheim (1986)

Rashed, R.: La naissance de l'algèbre. In: Noël, E. (Hrsg.) Le Matin des Mathémati ciens. Belin-Radio France, Paris (1985)

Alternative Hypothese

Eine Alternativhypothese ist die Hypothese, die sich von der getesteten Hypothese unterscheidet.

Die Alternativhypothese wird üblicherweise mit H1 bezeichnet.

GESCHICHTE Siehe **Hypothese** und **Hypothesentests.**

MATHEMATISCHE ASPEKTE

Beim **Hypothesentest** eines **Parameters** einer **Grundgesamtheit** wird die **Nullhypothese** wie folgt dargestellt:

$$H0 : \ddot{y} = \ddot{y}0 ,$$

Machine Translated by Google

wobei ÿ der Parameter der Population ist
der geschätzt werden soll, und ÿ0 ist der
angenommene **Wert** dieses Parameters. Die
Alternativhypothese kann dann drei verschiedene annehmen
Formen:

1. H1 : ÿ>ÿ0
2. H1 : ÿ<ÿ0
3. H1 : ÿ = ÿ0

In den ersten beiden Fällen der **Hypothesentest**
heißt das **einseitige,** während im dritten
In diesem Fall spricht man von der **Zweiseitigkeit.**

Die Alternativhypothese kann auch drei annehmen
verschiedene Formen während des **Hypothesentests**
von Parametern zweier Populationen. Wenn das
Die Nullhypothese behandelt die beiden Parameter ÿ1
und ÿ2 gleichermaßen, dann:

H0 : ÿ1 = ÿ2 oder

H0 : ÿ1 ÿ ÿ2 = 0 .

Die Alternativhypothese könnte dann lauten

- H1 : ÿ1 > ÿ2 oder H1 : ÿ1 ÿ ÿ2 > 0
- H1 : ÿ1 < ÿ2 oder H1 : ÿ1 ÿ ÿ2 < 0
- H1 : ÿ1 = ÿ2 oder H1 : ÿ1 ÿ ÿ2 = 0

Beim Vergleich von mehr als zwei
Populationen, so die **Nullhypothese**
dass die **Werte** aller **Parameter** sind
identisch. Wenn wir *k* Populationen vergleichen
wollen , lautet die Nullhypothese wie folgt:

H0 : ÿ1 = ÿ2 = ··· = *ÿk* .

Die Alternativhypothese wird dann wie folgt formuliert:

H1: Die Werte von ÿi(i = *1,...,k)* sind nicht alle
identisch.

Das bedeutet, dass nur ein **Parameter** benötigt wird
einen anderen Wert haben als die anderen
Parameter, um die **Nullhypothese** abzulehnen und
die Alternativhypothese zu akzeptieren.

BEISPIELE

Wir werden die Alternative prüfen
Hypothesen für drei Beispiele **von Hypothesen**
testen:

1. *Hypothesentest zum Prozentsatz von*
   eine Bevölkerung
   Ein Wahlkandidat möchte wissen, ob er
   wird mehr als 50 % der Stimmen erhalten.
   Die **Nullhypothese für** dieses Problem kann
   wie folgt geschrieben werden:

   H0 : ÿ = *0,5* ,

   Dabei ist ÿ der zu schätzende **Prozentsatz** der
   Bevölkerung .
   Wir führen auf der rechten **Seite einen einseitigen**
   Test durch, der es uns ermöglicht, die Dose zu beantworten
   Didates Frage. Die Alternativhypothese lautet
   daher:

   H1 : ÿ > *0,5* .

2. *Hypothesentest zum Mittelwert eines Pops*
   ulation
   Ein Bolzenhersteller möchte die Präzision testen
   einer neuen Maschine, die Schrauben herstellen sollte
   8 mm Durchmesser.
   Wir können die folgende **Nullhypothese**
   verwenden:

   H0 : ÿ = 8 ,

   wobei ÿ der **Mittelwert** der **Grundgesamtheit** ist
   das ist abzuschätzen.
   Zur Kontrolle führen wir einen **zweiseitigen Test** durch
   ob der Bolzendurchmesser zu klein ist oder
   zu groß.
   Die Alternativhypothese lässt sich wie folgt
   formulieren:

   H1 : ÿ = 8 .

3. *Hypothesentest anhand eines Vergleichs von*
   die Mittelwerte zweier Populationen
   Eine Versicherungsgesellschaft entschied sich für die Ausstattung
   seine Büros mit Mikrocomputern. Es will

Machine Translated by Google

Diese Computer von zwei verschiedenen Firmen zu kaufen, solange zwischen ihnen kein wesentlicher Unterschied in der Haltbarkeit besteht die beiden Marken. Es testet daher die Zeit

Das geht vor der ersten Panne weiter eine **Auswahl** von Mikrocomputern von jedem Marke.

Nach der **Nullhypothese** ist die **Mittelwert** der verstrichenen Zeit vor dem ersten Die Aufteilung ist für jede Marke gleich:

$$H0 : ÿ1 ÿ ÿ2 = 0 .$$

Dabei sind ÿ1 und ÿ2 die jeweiligen Mittelwerte der beiden **Populationen.**
Da wir nicht wissen, was das bedeutet sei das höchste, wir führen eine **zweiseitige** durch **prüfen.** Daher die Alternativhypothese wird sein:

$$H1 : ÿ1 ÿ ÿ2 = 0 .$$

WEITERLESEN

- **Varianzanalyse**
- **Hypothese**
- **Hypothesentest**
- **Nullhypothese**

REFERENZ

Lehmann, EI, Romann, SP: Testing Statis tical Hypothesis, 3. Aufl. Springer, Neu York (2005)

Analyse binärer Daten

Die Untersuchung, wie die Erfolgswahrscheinlichkeit hängt von erklärenden Variablen und der Gruppierung von Materialien ab.
Die Analyse binärer Daten umfasst auch **Anpassungstests** einer Binärstichprobe Variablen auch zu einer theoretischen Verteilung als das Studium von 2 × 2- **Kontingenztafeln** und deren anschließende Analyse. In Letzterem In diesem Fall weisen wir insbesondere auf **Unabhängigkeitstests** hin zwischen Attributen und **Homogenitätstests.**

GESCHICHTE

Siehe **Datenanalyse.**

MATHEMATISCHE ASPEKTE

Sei *Y* eine binäre **Zufallsvariable** und X1, *X2,...,Xk* seien ergänzende Binärvariablen. Die **Abhängigkeit** von *Y* von den Variablen X1, *X2,...,Xk* wird also durch die folgenden Modelle dargestellt (deren Koeffizienten sind geschätzt über die **maximale Wahrscheinlichkeit):**

1. *Lineares Modell:* P(Y = 1) wird ausgedrückt als eine lineare Funktion (in den Parametern) *von Xi.*
2. *Log-lineares Modell:* log P(Y = 1) ist ausgedrückt als lineare Funktion (im Parameter) von *Xi.*
3. *Logistisches Modell:* Log $\frac{P(Y=1)}{P(Y=0)}$ Ist ausgedrückt als lineare Funktion (im Parameter) von *Xi.*

Die Modelle 1 und 2 sind einfacher zu interpretieren. Noch Letzteres hat den Vorteil, dass die zu erklärende Größe alle möglichen Werte annimmt der linearen Modelle. Es ist auch wichtig Achten Sie auf die Extrapolation des Modells außerhalb der Domäne, in der es angewendet wird.
Es ist möglich, dass unter den Unabhängigen Variablen (X1, *X2,...,Xk),* es gibt kategorische Variablen (z. B. binäre). In diesem In diesem Fall ist es notwendig, das Nichtbinäre zu behandeln kategoriale Variablen auf folgende Weise:
Sei *Z* eine Zufallsvariable mit *m* cate blutrünstig. Wir zählen die Kategorien ab 1 auf zu *m* und wir definieren *m* ÿ 1 Zufallsvariable ables Z1, *Z2,...,Zmÿ1. Zi* nimmt also den Wert 1 an, wenn *Z* zu der Kategorie gehört, die durch diesen Index dargestellt wird. Die Variable *Z* wird also durch diese *m* ÿ 1 Variablen ersetzt, die Koeffizienten davon drücken den Einfluss von aus

Machine Translated by Google

die betrachtete Kategorie. Die Referenz (wird verwendet, um die Situation einer **Kollinearität** zu vermeiden **ity)** wird (zum Zwecke des Vergleichs mit anderen Kategorien) einen Parameter von haben null.

WEITERLESEN

- **Binärdaten**
- **Datenanalyse**

REFERENZEN

Cox, DR, Snell, EJ: Die Analyse binärer Daten. Chapman & Hall (1989)

Analyse kategorialer Daten

Die Analyse **kategorialer Daten** umfasst die folgenden Methoden:

(a) Eine Studie zum **Anpassungstest;**

(b) Das Studium einer **Kontingenztabelle** und ihre anschließende Analyse, die darin besteht, Beziehungen zwischen den Attributen (sofern vorhanden) zu entdecken und zu untersuchen;

(c) Ein **Homogenitätstest** einiger Populationen, bezogen auf die Verteilung einer binären qualitativen kategorialen Variablen;

(d) Eine Untersuchung der **Unabhängigkeitshypothese .**

GESCHICHTE

Der Begriff „Kontingenz“, der im Zusammenhang mit Kreuztabellen **kategorialer Daten** verwendet wird, wurde wahrscheinlich erstmals **von Pearson,** Karl (1904) verwendet . Der **Chi-Quadrat-Test** wurde 1937 von Barlett, MS, vorgeschlagen.

MATHEMATISCHE ASPEKTE

Siehe **Anpassungsgüte** und **Kontingenztabelle.**

WEITERLESEN

- **Daten**
- **Datenanalyse**
- **Kategoriale Daten**
- **Chi-Quadrat-Anpassungstest**
- **Kontingenztabelle**
- **Korrespondenzanalyse**
- **Anpassungstest**
- **Homogenitätstest**
- **Test der Unabhängigkeit**

VERWEISE

Agresti, A.: Kategorische Datenanalyse. Wiley, New York (1990)

Bartlett, MS: Eigenschaften von Suffizienz und statistische Tests. Proz. Roy. Soc. London. Ser. A **160,** 268–282 (1937)

Cox, DR, Snell, EJ: Analysis of Binary Data, 2. Aufl. Chapman & Hall, London (1990)

Haberman, SJ: Analyse qualitativer Daten. Bd. I: Einführungsthemen. Academic, New York (1978)

Pearson, K.: Zur Theorie der Kontingenz und ihrer Beziehung zu Assoziation und Normalität Korrelation. Drapers' Company Research Memoirs, Biometric Ser. I., S. 1–35 (1904)

Analyse von Residuen

Eine Residuenanalyse wird verwendet, um die Gültigkeit des statistischen Modells zu testen und die Annahmen zum Fehlerterm zu kontrollieren . Es kann auch zur Ausreißererkennung verwendet werden.

GESCHICHTE

Die Residuenanalyse geht auf Euler (1749) und Mayer (1750) zurück

Machine Translated by Google

Jahrhundert, die im Bereich der Astronomie mit dem Problem der **Schätzung** von **Parametern** aus **Beobachtungen** konfrontiert wurden. Die meisten Methoden zur Analyse von Residuen basieren auf den Arbeiten von Anscombe (1961) und Anscombe und Tukey (1963). Im Jahr 1973 präsentierte Anscombe außerdem eine interessante Diskussion über die Gründe für den Einsatz grafischer Analysemethoden. Cook und Weisberg (1982) widmeten der Analyse von Residuen ein komplettes Buch. Auch Draper und Smith (1981) befassten sich in einem Kapitel ihrer Arbeit *Applied Regression Analysis mit diesem Problem.*

MATHEMATISCHE ASPEKTE

Betrachten Sie ein allgemeines Modell **der multiplen linearen Regression:**

$$Yi = \ddot{y}0 + \ddot{y}jXij + \ddot{y}i \quad , \; j=1 \text{ bis } p\ddot{y}1 \qquad i = 1,...,n,$$

wobei *ÿi* der nicht beobachtbare **Zufallsfehler** ist Begriff.

Die **Hypothesen** für die **Fehler** *ÿi* lauten im Allgemeinen wie folgt: • Die Fehler sind unabhängig; • Sie sind normalverteilt (sie folgen einer **Normalverteilung);** • Ihr **Mittelwert** ist gleich Null; • Ihre **Varianz** ist konstant und gleich

ÿ2.

Die **Regressionsanalyse** liefert eine Schätzung für Yi, bezeichnet als *Yˆi.* Wenn das gewählte **Modell** angemessen ist, sollte die Verteilung der **Residuen** oder „beobachteten **Fehler“** *ei = Yi* ÿ Yˆi diese Hypothesen bestätigen.

Die zur Analyse von Residuen verwendeten Methoden sind hauptsächlich grafisch. Zu diesen Methoden gehören: 1. Darstellung der Residuen durch ein Häufigkeitsdiagramm (z. B. ein **Streudiagramm).**

2. Darstellung der Residuen als Funktion der Zeit (sofern die chronologische Reihenfolge bekannt ist).
3. Darstellung der Residuen als Funktion der Schätzwerte **Yˆi .**
4. Darstellung der Residuen als Funktion der **unabhängigen Variablen** *Xij.*
5. Erstellen eines Q-Q- **Diagramms** der Residuen.

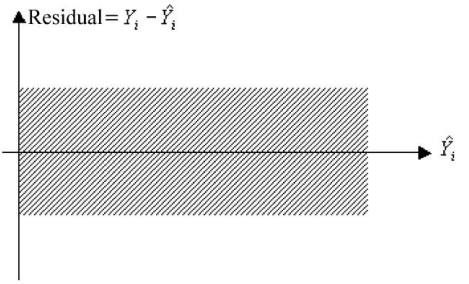
DOMÄNEN UND EINSCHRÄNKUNGEN

Um die Analyse zu validieren, müssen einige der Hypothesen zutreffen (wie zum Beispiel die Normalität der Residuen bei Schätzungen, die auf dem **mittleren Quadrat basieren).**

Betrachten Sie eine Darstellung der **Residuen** als Funktion der geschätzten **Werte** *Yˆi.* Dies ist einer der am häufigsten verwendeten grafischen Ansätze zur Überprüfung der Gültigkeit eines **Modells.** Es besteht darin, Folgendes zu platzieren: • Die **Residuen** *ei = Yi* ÿ Yˆi in zunehmender Form **Befehl;**

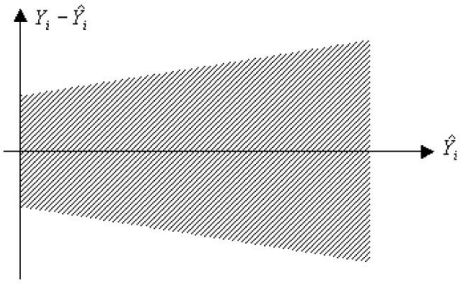
• Die Schätzwerte Yˆi auf der Abszisse.

Wenn das gewählte **Modell** geeignet ist, werden die **Residuen** gleichmäßig auf einem horizontalen Punkteband verteilt.



Wenn die **Hypothesen** für die **Residuen** jedoch nicht verifiziert werden, kann die Form des Diagramms davon abweichen. Die drei folgenden Abbildungen zeigen die Formen, die man erhält, wenn: 1. Die **Varianz** ÿ2 nicht konstant ist. In diesem Fall ist es notwendig, eine **Transformation** der **Daten** *Yi* durchzuführen, bevor die **Regressionsanalyse in Angriff genommen wird.**

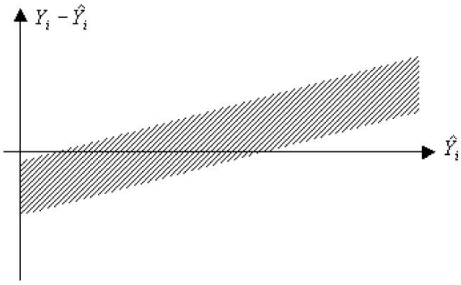
Machine Translated by Google



2. Das gewählte **Modell** ist unzureichend (z

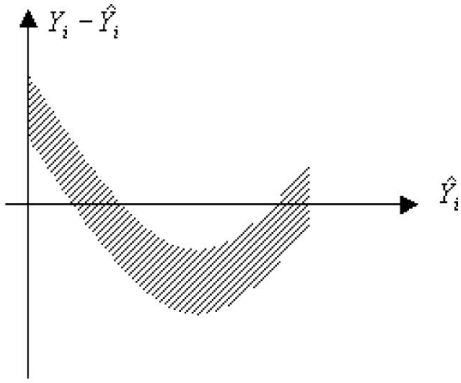
Das Modell ist beispielsweise linear, aber der konstante Term wurde weggelassen, als er erforderlich war.

Aufsatz).



3. Das gewählte **Modell** ist unzureichend

(Es wird eine parabolische Tendenz beobachtet).



Verschiedene **Statistiken** wurden vorgeschlagen

um numerische Messungen zu ermöglichen

ergänzen die visuellen Techniken

oben dargestellt, zu denen auch die von Anscombe

(1961) und Anscombe and gegebenen gehören

Türkei (1963).

BEISPIELE

Im neunzehnten Jahrhundert ein schottischer Physiker

namens Forbe wollte James D. schätzen

die Höhe über dem Meeresspiegel durch Messung der

Siedepunkt von Wasser. Er wusste, dass die Höhe

anhand der Atmosphäre bestimmt werden konnte

phärischer Druck; Anschließend untersuchte er die Beziehung

zwischen Druck und Siedepunkt von

Wasser. Forbe schlug das für eine **gewisse Zeit** vor

der beobachteten **Werte,** eine Darstellung des Logarithmus von

der Druck als Funktion des Siedepunktes

Wasser sollte eine gerade Linie ergeben. Seit

der Logarithmus dieser Drücke ist klein und

Da die Werte kaum variieren, haben wir diese Werte multipliziert

um 100 darunter.

X Siedepunkt	Y 100 · log (Druck)
194,5	131,79
194.3	131,79
197,9	135.02
198.4	135,55
199,4	136,46
199,9	136,83
200,9	137,82
201.1	138,00
201.4	138.06
201.3	138.05
203.6	140.04
204.6	142,44
209,5	145,47
208.6	144,34
210.7	146,30
211.9	147,54
212.2	147,80

Das einfache lineare Regressionsmodell hierfür

Das Problem ist:

$$Yi = \ddot{y}0 + \ddot{y}1Xi + i\,, \qquad i = 1,\ldots, 17 \quad .$$

Machine Translated by Google

Mit der Methode der **kleinsten** Quadrate können wir finden die folgende Schätzfunktion:

$$\hat{Y}i = \ddot{y}42{,}131 + 0{,}895Xi$$

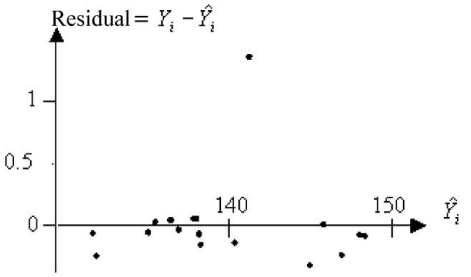
wobei Yˆi der geschätzte **Wert** der **Variablen Y** ist für ein gegebenes X.

Für jeden dieser 17 Werte von *Xi* haben wir ein geschätzter Wert *Yˆi*. Wir können das berechnen **Reste:**

$$ei = Yi \ddot{y} \hat{Y}i .$$

Diese Ergebnisse werden im Folgenden vorgestellt Tisch:

	Xi	Yi	$\hat{Y}_i$	$ei = Yi - \hat{Y}_i$
1	194,5	131,79	132,037	ÿ0,247
2	194,3	131,79	131,857	ÿ0,067
3	197,9	135,02	135,081	ÿ0,061
4	198,4	135,55	135,529	0,021
5	199,4	136,46	136,424	0,036
6	199,9	136,83	136,872	ÿ0,042
7	200,9	137,82	137,768	0,052
8	201,1	138,00	137,947	0,053
9	201,4	138,06	138,215	ÿ0,155
10	201,3	138,05	138,126	ÿ0,076
11	203,6	140,04	140,185	ÿ0,145
12	204,6	142,44	141,081	1.359
13	209,5	145,47	145,469	0,001
14	208,6	144,34	144,663	ÿ0,323
15	210,7	146,30	146,543	ÿ0,243
16	211,9	147,54	147,618	ÿ0,078
17	212,2	147,80	147,886	ÿ0,086



Darstellung der **Residuen** als Funktion von Schätzwerte Yˆi geben den vorherigen **an** Graph.

Aus dieser Grafik geht hervor, dass mit Ausnahme von eine **Beobachtung** (die 12.), wo der Wert des **Residuums** scheint auf einen **Ausreißer** hinzuweisen **ähm,** die Residuen sind sehr dünn verteilt horizontaler Streifen. In diesem Fall tun es die **Residuen** keinen Anlass geben, an der Gültigkeit zu zweifeln des gewählten **Modells.** Durch die Analyse der standardisierten Residuen können wir feststellen, ob ob die 12. Beobachtung ein **Ausreißer** ist oder nicht.

WEITERLESEN

- **Anderson-Darling-Test**
- **Kleinsten Quadrate**
- **Multiple lineare Regression**
- **Ausreißer**
- **Regressionsanalyse**
- **Restwert**
- **Streudiagramm**
- **Einfache lineare Regression**

VERWEISE

Anscombe, FJ: Untersuchung von Residuen. Proz. 4. Berkeley Symp. Mathematik. Statist. Wahrscheinlich. 1, 1–36 (1961)

Anscombe, FJ: Graphen in der statistischen Analyse. Bin. Stat. **27,** 17–21 (1973)

Anscombe, FJ, Tukey, JW: Analyse von Reste. Technometrie 5, 141–160 (1963)

Cook, RD, Weisberg, S.: Residuen und EinflussinRegression.Chapman&Hall, London (1982)

Cook, RD, Weisberg, S.: Eine Einführung zu Regression Graphics.Wiley, New York (1994)

Cook, RD, Weisberg, S.: Angewandte Regression einschließlich Computer und Grafik. Wiley, New York (1999)

Machine Translated by Google

Draper, NR, Smith, H.: Applied Regression Analysis, 3. Aufl. Wiley, New York (1998)

Euler, L.: Recherchiert über die Frage, die die Bewegung des Saturns und des Mondes bestimmt hat Jupiter, ein Teil, hat den Preis gemeldet Im Jahr 1748, von der Académie Royale des Wissenschaften von Paris. Republiziert im Jahr 1960, dat Leonhardi Euleri, Opera Omnia, 2ème Serie. Turici, Basel, 25, S. 47–157 (1749)

Mayer, T.: Abhandlung über die Umwälzung desMonds um seineAchse und die schein bloße Bewegung der Mondflecken. Kosmographische Nachrichten und Sammlungen auf das Jahr 17481, 52–183 (1750)

Varianzanalyse

Die Varianzanalyse ist eine Technik, die besteht darin, die Gesamtvariation von zu trennen **Der Datensatz** wird in logische Komponenten unterteilt, die bestimmten Variationsquellen der Reihe nach zugeordnet sind um den **Mittelwert** mehrerer **Populationen zu vergleichen.** Diese Analyse hilft uns auch beim Testen bestimmte **Hypothesen** zu den Parametern des Modells aufzustellen oder die Komponenten der **Varianz abzuschätzen.** Die Quellen von vari ationen werden global in einer Komponente namens **Fehlervarianz,** manchmal auch genannt, zusammengefasst Mittelquadrat innerhalb der Behandlung und ein anderes Komponente, die als „Wirkung“ oder Behandlung bezeichnet wird, manchmal auch als Zwischenbehandlung bezeichnet quadratischer Mittelwert.

GESCHICHTE

Die Varianzanalyse geht auf **Fisher, RA** (1925) zurück. Er legte die ersten Grundprinzipien auf diesem Gebiet fest. Analyse von Varianz wurde erstmals in den Bereichen Biologie und Landwirtschaft angewendet.

MATHEMATISCHE ASPEKTE

Die Varianzanalyse vergleicht **die mittels** drei oder mehr **Zufallsstichproben** und ermittelt, ob es einen signifikanten Unterschied zwischen den **Populationen** gibt aus dem die Proben entnommen werden. Das Technik kann nur angewendet werden, wenn der Zufall Stichproben sind unabhängig, wenn die Grundgesamtheit Verteilungen sind ungefähr normal und alle haben die gleiche Varianz ÿ2.

Nachdem festgestellt wurde , **dass die Nullhypothese** geht davon aus, dass die **Mittel** gleich sind, während die **Die alternative Hypothese** bestätigt dies zumindest Einer davon ist anders, wir beheben einen **signifikanten eben.** Anschließend nehmen wir zwei **Schätzungen** vor unbekannte **Varianz** ÿ2:

- Die erste, mit s2 bezeichnet $_{E,}$ entspricht dem Mittelwert der Varianzen jeder Stichprobe;
- Die zweite, s2 $_{Tr,}$ basiert auf der Variation zwischen den Mittelwerten der Stichproben.

Wenn die **Nullhypothese** bestätigt ist, gilt im Idealfall: Diese beiden Schätzungen werden gleich sein und die F Das Verhältnis *(F* = s2 wie im **Fisher-Test** verwendet Tr/s2 E, definiert als Quotient der zweiten Schätzung von ÿ2 zur ersten) ist gleich 1.

Der **Wert** des *F*- Verhältnisses, der im Allgemeinen beträgt mehr als 1 aufgrund der Abweichung von der **Probenahme,** muss mit dem Wert in verglichen werden die **Fisher-Tabelle** entspricht dem Fest **signifikantes Niveau.** Die Entscheidungsregel besteht entweder die Nullhypothese abzulehnen, wenn die Der berechnete Wert ist größer oder gleich dem tabellarischer Wert, sonst sind die **Mittelwerte** gleich, was zeigt, dass **die Proben** aus der stammen gleiche **Bevölkerung.**

Betrachten Sie das folgende **Modell:**

$$Yij = ÿ + ÿi + ÿij ,$$

$$i = 1, 2,...,t, j = 1, 2,...,ni .$$

Hier

Yij repräsentiert die **Beobachtung,** *die j* empfängt die **Behandlung** ich,

Machine Translated by Google

ÿ ist **der allgemeine Mittelwert** für alle Behandlungen,

ÿi ist die tatsächliche Wirkung der Behandlung *i* auf die Überwachung,

ÿij ist der experimentelle Fehler für die Beobachtung *Yij*.

In diesem Fall wird die **Nullhypothese** ausgedrückt auf die folgende Weise:

$$H0 : \ddot{y}1 = \ddot{y}2 = \cdots = \ddot{y}t ,$$

was bedeutet, dass die *t*- Behandlungen identisch sind tisch.

Die **Alternativhypothese** formuliert in auf folgende Weise:

H1 : die **Werte** von ÿi(i = *1, 2,...,t)* sind nicht alle identisch .

Folgende Formeln werden verwendet:

$$SSTr = \sum_{i=1}^{T} ni(\bar{Y}i. \ddot{y} \bar{Y}..)^2 , \quad s^2_{Tr} = \frac{SSTr}{t \ddot{y} 1} ,$$

$$SSE = \sum_{i=1}^{T} \sum_{j=1}^{ni} (Yij \ddot{y} \bar{Y}i.)^2 , \quad s^2_E = \frac{SSE}{N \ddot{y} t} ,$$

Und

$$SST = \sum_{i=1}^{T} \sum_{j=1}^{ni} (Yij \ddot{y} \bar{Y}..)^2$$

oder

$$SST = SSTr + SSE .$$

Wo

$\bar{Y}i. = \sum_{j=1}^{ni} \frac{Yij}{ni}$ ist der **Mittelwert** von der i-te Satz

$\bar{Y}.. = \frac{1}{N} \sum_{i=1}^{T} \sum_{j=1}^{ni}$ *Yij* ist der globale Mittelwert alles übernommen **Beobachtungen** und

$N = \sum_{i=1}^{T} ni$ ist die Gesamtzahl von Beobachtungen.

und schließlich der Wert des F-Verhältnisses

$$F = \frac{s^2_{Tr}}{s^2_E} .$$

Es ist üblich, die Informationen zusammenzufassen aus der Varianzanalyse in einer Analyse Varianztabelle:

Quelle von Variation	Abschlüsse von Freiheit	Die Summe von Quadrate	Bedeuten von Quadrate	F
Unterbehandeln gen	*t* ÿ 1	SSStr	s^2_{Tr}	$\frac{s^2_{Tr}}{s^2_E}$
Innerhalb behandeln gen	*N* ÿ *t*	*SSE*	s^2_E	
Insgesamt	*N* ÿ 1	*SST*		

DOMÄNEN UND EINSCHRÄNKUNGEN

Eine Varianzanalyse ist immer mit einem **Modell verbunden.** Daher gibt es bei jedem Unterschied eine unterschiedliche Varianzanalyse Fall. Betrachten Sie zum Beispiel den Fall, in dem die Varianzanalyse wird auf **faktorielle Experimente** mit einem oder mehreren **Faktoren angewendet,** und diese faktoriellen Experimente sind damit verknüpft mehrere **Versuchspläne.**

Wir können nicht nur die Anzahl unterscheiden **Faktoren** im **Experiment ,** sondern auch die Art von **Hypothesen** im Zusammenhang mit den Auswirkungen der **Behandlungen. Wir** haben dann **ein Modell** mit festem Effekte, ein Modell mit variablen Effekten und ein Modell mit gemischten Effekten. Jedes von diesen erfordert eine spezifische Analyse, aber welches Modell auch immer verwendet wird, die Grundannahmen von Additivität, Normalität, Homoskedastizität und Unabhängigkeit muss respektiert werden. Das bedeutet Das:

1. Die experimentellen **Fehler** des **Modells** sind **Zufallsvariablen ,** die unabhängig sind von einander;

Machine Translated by Google

2. Alle Fehler folgen einer **normalen Verteilung** **Bution** mit einem **Mittelwert** von Null und einem unbekannte **Varianz** ÿ2.

Alle **Versuchspläne** können analysiert werden mittels Varianzanalyse. Die gebräuchlichsten Designs sind **vollständig zufällig Designs, randomisierte Blockdesigns** und **Lateinische quadratische Designs.**

Eine Varianzanalyse kann auch einfach oder mehrfach **linear** durchgeführt werden **Rückschritt.**

Wenn während einer Varianzanalyse die Null Hypothese (der Fall der Gleichheit der Mittel) ist abgelehnt, ein **Test mit dem geringsten signifikanten Unterschied** wird verwendet, um die **Populationen** zu identifizieren, die vorhanden sind signifikant unterschiedlich bedeutet, was eine Varianzanalyse nicht leisten kann.

BEISPIELE

Siehe **bidirektionale Varianzanalyse, einseitige Varianzanalyse, lineares Vielfaches Regression** und **einfache lineare Regression.**

WEITERLESEN

Versuchsplanung
Faktor
Fisher-Verteilung
Fisher-Tisch
Fisher-Test
Test auf den geringsten signifikanten Unterschied
Multiple lineare Regression
Einseitige Varianzanalyse
Regressionsanalyse
Einfache lineare Regression
Zwei-Wege-Varianzanalyse

VERWEISE

Fisher, RA: Statistische Methoden für Forschungsmitarbeiter. Oliver & Boyd, Edinburgh (1925)

Rao, CR: Fortgeschrittene statistische Methoden in biometrischer Forschung. Wiley, New York (1952)

Scheffé, H.: Die Varianzanalyse. Wiley, New York (1959)

Anderson, Oskar

Anderson, Oskar (1887–1960) war ein wichtiges Mitglied der Continental School für Statistik; seine Beiträge berührten Ein breites Themenspektrum, darunter Korrelation, Zeitreihenanalyse, nichtparametrische Methoden und Stichprobenerhebung, sowie Ökonometrie und statistische Anwendungen in Sozialwissenschaften.

Anderson erhielt Oskar einen Bachelor-Abschluss mit Auszeichnung vom Kasaner Gymnasium und studierte anschließend Mathematik und Physik für ein Jahr an der Universität Kasan. Er trat dann in die Fakultät für Wirtschaftswissenschaften ein das Polytechnische Institut von St. Petersburg, wo er Mathematik, Statistik und studierte Wirtschaft.

Die Veröffentlichungen von Anderson und Oskar verbinden die Traditionen der Continental School der Statistik mit den Konzepten des Englischen Biometrische Schule, insbesondere in zwei von seine Werke: „Einführung in die mathematische Statistik“ und „Probleme der statis tischen Methodenlehre in den Sozialwissenschaften senschaften“.

1949 gründete er mit ihm die Zeitschrift *Mitteilungsblatt für Mathematische Statistik* Kellerer, Hans und Münzner, Hans.

Einige Hauptwerke von Anderson, Oskar:

1935 Einführung in die Mathematische Statistik. Julius Springer, Wien

1954 Probleme der statistischen Methodik denlehre in den Sozialwissenschaf zehn. Physica-Verlag, Würzberg

Machine Translated by Google

Anderson, Theodore W.

Anderson, Theodore Wilbur wurde am geboren
am 5. Juni 1918 in Minneapolis, im
Bundesstaat Minnesota in den USA. Er wurde
ein Doktor der Mathematik im Jahr 1945 an der
University of Princeton, und im Jahr 1946 er
wurde Mitglied der Abteilung für
Mathematische Statistik an der Universität
Columbia, wo er zum Professor ernannt wurde
im Jahr 1956. Im Jahr 1967 wurde er zum Professor ernannt
für Statistik und Wirtschaftswissenschaften an der Stanford Uni
Vielseitigkeit. Er war nacheinander: Fellow of the
Guggenheim Foundation zwischen 1947 und
1948; Herausgeber der *Annals of Mathematical*
Statistiken von 1950 bis 1952; Präsident der
Institut für Mathematische Statistik im Jahr 1963;
und Vizepräsident des American Statis
Er war von 1971 bis 1973 Mitglied der Technischen Vereinigung
ein Mitglied der American Academy of Arts
und Wissenschaften der National Academy of
Naturwissenschaften, des Instituts für Mathematik
Statistik und der Royal Statistical Society. Andersons
wichtigster Beitrag
zur Statistik gehört sicherlich zum Bereich der
multivariaten Analyse. 1958 veröffentlichte er die
Buch mit dem Titel *„An Introduction to Multivari“.*
ate Statistische Analyse. Dieses Buch war das
Nachschlagewerk in diesem Bereich seit über vierzig Jahren
Jahre. Es wurde sogar ins Russische übersetzt.

Einige der wichtigsten Werke und Artikel von Theodore Wilbur Anderson:

1952 (mit Darling, DA) Asymptotische Theorie bestimmter Anpassungsgütekriterien basierend auf stochastischen Prozessen. Ann. Mathematik. Stat. 23, 193–212.

1958 Eine Einführung in multivariate Sta tistische Analyse. Wiley, New York.

1971 Die statistische Analyse der Zeit Serie. Wiley, New York.

1989 Lineare latente Variablenmodelle und Kovarianzstrukturen. J. Economet rics, 41, 91–119.

1992 (mit Kunitoma, N.) Asymptotisch Verteilungen der Regressions- und Autoregressionskoeffizienten mit Martin-Gale-Differenzstörungen. J. Multivariate Anal., 40, 221–243.

1993 Anpassungstests für Spektralverteilungen. Ann. Stat. 21, 830–847.

WEITERLESEN

Anderson-Darling-Test

Anderson-Darling-Test

Der Anderson-Darling-Test ist ein Anpassungstest, der
es ermöglicht, die Hypothese zu kontrollieren, dass die
Verteilung einer Zufallsvariablen vorliegt
Die in einer Probe beobachtete Menge folgt einer
bestimmten theoretischen Verteilung. Insbesondere ermöglicht es
Wir testen, ob die empirische Verteilung
Der erhaltene Wert entspricht einer **Normalverteilung.**

GESCHICHTE

Anderson, Theodore W. und Darling DA
verwendete zunächst die Anderson-Darling-Statistik,
mit A2 bezeichnet , um die Konformität einer Verteilung
mit perfekt spezifizierten Parametern zu testen
(1952 und 1954). Später, in den 1960er Jahren
und insbesondere die 1970er Jahre, einige andere Autoren
(hauptsächlich Stephens) haben den Test an einen breiteren Rahmen angepasst
Reihe von Distributionen, bei denen einige der
Parameter sind möglicherweise nicht bekannt.

MATHEMATISCHE ASPEKTE

Betrachten wir die **Zufallsvariable** X,
was der Normalverteilung mit folgt
eine Erwartung ÿ und eine Varianz ÿ2, und
hat eine Verteilungsfunktion FX(x; ÿ), wobei ÿ
ist ein Parameter (oder eine Reihe von Parametern), der

Machine Translated by Google

bestimmen, *FX.* Wir nehmen weiterhin ÿ an
bekannt sein.

Eine Beobachtung einer Stichprobe der Größe *n* wurde ausgegeben
aus der VariablenX ergibt sich eine Verteilungsfunktion
Fn(x). Die Anderson-Darling-Statistik,
mit A2 bezeichnet , ergibt sich dann aus der
gewichteten Summe der quadrierten Abweichungen FX(x4ÿÿ)ÿ und ÿ2 sind beide unbekannt und werden jeweils
Fn(x):

$$A2 = \frac{1}{N} \sum_{i=1}^{N} (FX(x;\ÿ) \ÿ Fn(x))2 .$$

Ausgehend von der Tatsache, dass A2 ein Zufall ist
Variable, die einer bestimmten Verteilung folgt
über das Intervall [0; +ÿ[, es ist möglich
Testen Sie für ein a priori festgelegtes
Signifikanzniveau, ob Fn(x) die Realisierung des ist
ZufallsvariableFX(X; ÿ); das heißt, obX
folgt der Wahrscheinlichkeitsverteilung mit dem
Verteilungsfunktion FX(x; ÿ).

2 Berechnung der Statistik

Ordnen Sie die Beobachtungen x1, *x2,...,xn* an
Die Probe wird von X in aufsteigender Reihenfolge ausgegeben, d. h.
x1 < x2 < ··· < *xn.* Beachten Sie, dass *zi* =
FX(xi; ÿ), *(i = 1, 2,...,n).* Berechnen Sie dann:
A2 von:

$$A2 = \ÿ1 \frac{}{N} \sum_{i=1}^{N} (2i \ÿ 1)(\ln (zi) + \ln(1 \ÿ zn+1ÿi)) \ÿ n .$$

Für die hier bevorzugte Situation *(X* folgt
die Normalverteilung mit Erwartungswert ÿ
und Varianzÿ2) können wir abhängig von den
bekannten Parametern ÿ vier Fälle aufzählen
und ÿ2 *(F* ist die Verteilungsfunktion der
Standardnormalverteilung):

1. ÿ und ÿ2 sind bekannt, also FX(x;(ÿ, ÿ2))
   ist perfekt spezifiziert. Dann natürlich
   Es gilt *zi = F(wi)* , wobei *wi* = 2. ÿ2 $\frac{xi\ÿ\ÿ}{\ÿ}$.
ist bekannt, aber ÿ ist unbekannt und wird unter
   Verwendung von $\bar{x}$ = geschätzt $\frac{1}{N}$ lch xi , der Mittelwert von
   die Probe. Dann sei *zi = F(wi),* wobei
   wi = $\frac{xi\ÿ\bar{x}}{\ÿ}$.
3. ÿ ist bekannt, aber ÿ2 ist unbekannt und wird mit
   s2 = *(xi* ÿ u)2 geschätzt . $\frac{1}{N}$ In
   In diesem Fall sei *zi = F(wi),* wobei *wi* =
   $\frac{x(i)\ÿ\ÿ}{S}$.
   unter Verwendung von *x* und s2 = $\bar{(xi}$ ÿ x)2)
   $\frac{1}{n\ÿ1}$ (geschätzt . Dann sei *zi = F(wi),*
   wobei *wi* = Für $\frac{xi\ÿ\bar{x}}{S}$.

A2 wurden asymptotische Verteilungen gefunden
von Anderson und Darling für den ersten Fall,
und von Stephens für die nächsten beiden Fälle. Für
Im letzten Fall ermittelte Stephens eine asymptotische
Verteilung für die Transformation: Aÿ =
A2(1,0 +). $\frac{0,75}{N}$ + $\frac{2,25}{n2}$
Daher können wir, wie unten gezeigt, konstruieren
eine Tabelle, die je nach Fall und
das Signifikanzniveau (10 %, 5 %, 2,5 % oder 1 %)
unten), die Grenzwerte von A2 (und Aÿ
für den Fall 4), jenseits dessen die Normalität
Hypothese wird abgelehnt:

	Signifikanzniveau			
Fall:	*0,1*	**0,050**	**0,025**	**0,01**
1: A2 =	1,933	2,492	3,070	3,857
2: A2 =	0,894	1,087	1,285	1,551
3: A2 =	1,743	2,308	2,898	3,702
4: Aÿ =	0,631	0,752	0,873	1,035

DOMÄNEN UND EINSCHRÄNKUNGEN

Da die Verteilung *von A2* asymptotisch ausgedrückt
wird, muss der Test die Stichprobengröße *n* haben
groß. Wenn dies nicht der Fall ist, dann zunächst einmal
In zwei Fällen ist die Verteilung *von A2* nicht bekannt
und es ist notwendig, eine Transformation vom Typ
A2 ÿÿ Aÿ durchzuführen, woraus Aÿ
bestimmt werden kann. Wenn *n* > 20, können wir
Vermeiden Sie eine solche Transformation und damit die Daten
in der obigen Tabelle sind gültig.

Der Anderson-Darling-Test hat den Vorteil, dass er
auf einen weiten Bereich angewendet werden kann

Machine Translated by Google

von Verteilungen (nicht nur eine **Normalverteilung ,** sondern auch Exponential-, Logistik- und Gammaverteilungen, unter anderem). Das ermöglicht es uns, eine große Auswahl an Alternativen auszuprobieren Verteilungen, wenn der erste Test die Null ablehnt Hypothese für die Verteilung eines Zufalls Variable.

BEISPIELE

Die folgenden Daten veranschaulichen die Anwendung des Anderson-Darlingtests für die Normalitätshypothese:

Betrachten Sie ein Beispiel der Höhen (in cm) von 25 männliche Studenten. Die folgende Tabelle zeigt die Beobachtungen in der Stichprobe, und auch *wi* und *zi*. Wir können auch *x* und *s* daraus berechnen diese Daten: $x \equiv 177{,}36$ und $s = 4{,}98$.

Unter der Annahme, dass *F* eine Standardnormalverteilungsfunktion ist, gilt:

Obs:	xi	$wi = \frac{x}{xi\ s}$	$zi = F\ wi$
1	169	ÿ1,678	0,047
2	169	ÿ1,678	0,047
3	170	ÿ1,477	0,070
4	171	ÿ1,277	0,100
5	173	ÿ0,875	0,191
6	173	ÿ0,875	0,191
7	174	ÿ0,674	0,250
8	175	ÿ0,474	0,318
9	175	ÿ0,474	0,318
10	175	ÿ0,474	0,318
11	176	ÿ0,273	0,392
12	176	ÿ0,273	0,392
13	176	ÿ0,273	0,392
14	179	0,329	0,629
15	180	0,530	0,702
16	180	0,530	0,702
17	180	0,530	0,702
18	181	0,731	0,767
19	181	0,731	0,767
20	182	0,931	0,824
21	182	0,931	0,824
22	182	0,931	0,824
23	185	1.533	0,937
24	185	1.533	0,937
25	185	1.533	0,937

Wir erhalten dann A2 ÿ= 0,436, was ergibt

$$Aÿ = A2 \cdot 1{,}0 + \frac{0{,}75}{25} + \frac{0{,}25}{625}$$

$$= A2 \cdot (1{,}0336)\ ÿ = 0{,}451\ .$$

Da wir Fall 4 haben und ein Signifikanzniveau auf 1 % festgelegt ist, beträgt der berechnete Wert von Aÿ viel weniger als der in der Tabelle angegebene Wert (1.035). Daher die Normalitätshypothese kann auf einem Signifikanzniveau von nicht abgelehnt werden 1 %.

WEITERLESEN

- **Anpassungstest**
- **Histogramm**
- **Nichtparametrische Statistik**
- **Normalverteilung**
- **Statistiken**

VERWEISE

Anderson, TW, Darling, DA: Asymptotische Theorie bestimmter Anpassungsgütekriterien basierend auf stochastischen Prozessen. Ann.Math. Stat. **23,** 193–212 (1952)

Anderson, TW, Darling, DA: Ein Test von Güte der Anpassung. Marmelade. Stat. Assoc. **49,** 765–769 (1954)

Durbin, J., Knott, M., Taylor, CC: Komponenten der Cramer-Von-Mises-Statistik, II. J. Roy. Stat. Soc. Ser. B **37,** 216–237 (1975)

Stephens, MA: EDF-Statistiken für das Gute der Passform und einige Vergleiche. Marmelade. Stat. Assoc. **69,** 730–737 (1974)

Machine Translated by Google

Arithmetisches Mittel

Das arithmetische **Mittel** ist ein **Maß für cen trale Tendenz.** Es ermöglicht uns, das Zentrum der **Häufigkeitsverteilung** einer **quantitativen Variablen** zu charakterisieren , indem wir alle Beobachtungen mit der gleichen Gewichtung berücksichtigen (im Gegensatz zum **gewichteten arithmetischen Mittel).**

Sie wird berechnet, indem die Beobachtungen summiert und dann durch die Anzahl der Beobachtungen dividiert werden.

GESCHICHTE

Das arithmetische Mittel ist eines der ältesten Methoden zur Kombination von Beobachtungen um einen eindeutigen ungefähren Wert anzugeben. Es scheint erstmals im dritten Jahrhundert v. Chr. von babylonischen Astronomen verwendet worden zu sein. Das arithmetische Mittel wurde von den Astronomen verwendet, um die Positionen der Sonne, des Mondes und der Planeten zu bestimmen. Laut Plackett (1958) stammt das Konzept des arithmetischen Mittels vom griechischen Astronomen Hipparchos.

Im Jahr 1755 schlug Thomas Simpson in einem Brief an den Präsidenten der Royal Society offiziell die Verwendung des arithmetischen Mittels vor.

MATHEMATISCHE ASPEKTE

Sei x1, *x2,...,xn* eine Menge von *n* Größen oder *n* Beobachtungen, die sich auf eine **quantitative Variable** X beziehen.

Das arithmetische Mittel $\bar{x}$ von x1, *x2,...,xn* ist die Summe dieser Beobachtungen geteilt durch die Anzahl *n* der Beobachtungen:

$$\bar{x} = \frac{\sum_{i=1}^{N} xi}{N} .$$

Wenn die Beobachtungen in der Reihenfolge angeordnet sind

Form einer **Häufigkeitsverteilung,** der Arith

Im Folgenden wird der metrische Mittelwert berechnet Weg:

$$\bar{x} = \frac{\sum_{i=1}^{k} xi \cdot fi}{\sum_{i=1}^{k} fi} ,$$

Dabei sind *xi* die verschiedenen Werte der **Variablen,** *fi* die mit diesen Werten verbundenen Häufigkeiten, *k* die Anzahl der verschiedenen Werte und die Summe der Häufigkeiten gleich der Anzahl der Beobachtungen:

$$\sum_{i=1}^{k} fi = n .$$

Um den **Mittelwert** einer Häufigkeitsverteilung zu berechnen, bei der Werte der quantitativen Variablen *X* in Klassen gruppiert sind, gehen wir davon aus, dass alle Beobachtungen, die zu einer bestimmten Klasse gehören, den zentralen Wert von annehmen der Klasse, unter der Annahme, dass die Beobachtungen innerhalb der Klassen gleichmäßig verteilt sind (wenn diese **Hypothese** nicht zutrifft, ist das erhaltene arithmetische Mittel nur eine Annäherung).
Daher haben wir in diesem Fall:

$$\bar{x} = \frac{\sum_{i=1}^{k} xi \cdot fi}{\sum_{i=1}^{k} fi} ,$$

Dabei sind *xi* die Klassenzentren, *fi* die mit jeder Klasse verbundenen Häufigkeiten und *k* die Anzahl der Klassen.

Eigenschaften des arithmetischen

Mittels • Die algebraische Summe der Abweichungen zwischen jedem Wert der Menge und dem arithmetischen Mittel dieser Menge ist gleich 0:

$$\sum_{i=1}^{N} (xi - \bar{x}) = 0 .$$

Machine Translated by Google

• Die Summe der quadratischen Abweichungen von jedem Der Wert einer gegebenen Zahl „*a*“ ist am kleinsten wenn „*a*“ das arithmetische Mittel ist:

$$\sum_{i=1}^{N} (x_i - a)^2 \geq \sum_{i=1}^{N} (x_i - \bar{x})^2 .$$

Nachweisen:

Wir können schreiben:

$$x_i - a = (x_i - \bar{x}) + (\bar{x} - a) .$$

Finden der Quadrate beider Mitglieder von die Gleichheit, sie zusammenfassen und dann Vereinfachung ergibt:

$$\sum_{i=1}^{N} (x_i - a)^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2 + n \cdot (\bar{x} - a)^2 .$$

Da $n \cdot (\bar{x} - a)^2$ nicht negativ ist, gilt geprüft, dass:

$$\sum_{i=1}^{N} (x_i - a)^2 \geq \sum_{i=1}^{N} (x_i - \bar{x})^2 .$$

• Das arithmetische Mittel $\bar{x}$ einer **Stichprobe** *(x1,...,xn)* wird normalerweise berücksichtigt sei ein **Schätzer** des **Mittelwerts** μ des **Population ,** aus der die Stichprobe stammte vergriffen.

• Angenommen, dass *xi* unabhängige Zufallsvariablen mit der **gleichen Verteilung** sind **Funktion** für den Mittelwert μ und die **Varianz** σ^2, das können wir zeigen

1. $E[\bar{x}] = \mu$,
2. $\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{N}$,

wenn diese Momente existieren.

Da die **mathematische Erwartung** von $\bar{x}$ gleich μ, das arithmetische Mittel ist ein Schätzer **ohne Verzerrung** des Mittelwerts der Bevölkerung.

• Wenn das Ergebnis aus der **Zufallsstichprobe** resultiert ohne Ersatz einer endlichen Bevölkerung mit einem Mittelwert μ, der Identität

$$E[\bar{x}] = \mu$$

ist immer noch gültig, aber die Varianz von $\bar{x}$ muss sein angepasst um einen Faktor, der davon abhängt Größe *N* der Bevölkerung und die Größe *n* von die Probe:

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{N} \cdot \frac{N - n}{N - 1} ,$$

wobei σ^2 die Varianz der Grundgesamtheit ist.

Beziehung zwischen dem arithmetischen Mittel und andere Maße der zentralen Tendenz

• Das arithmetische Mittel bezieht sich auf zwei Hauptmaße der zentralen Tendenz: das **Modus** Mo und der **Median** Md.

Wenn die Verteilung symmetrisch und unimodal ist:

$$\bar{x} = \mathrm{Md} = \mathrm{Mo} .$$

Wenn die Verteilung unimodal ist, gilt normalerweise:

$\bar{x} - \mathrm{Md} \approx \mathrm{Mo}$, wenn die Verteilung ist nach rechts gestreckt,

$\bar{x} - \mathrm{Md} \approx \mathrm{Mo}$, wenn die Verteilung ist nach links gestreckt.

Für ein unimodales, leicht asymmetrisches Modell Verteilung, diese drei Maßnahmen der zentrale Tendenz oft ungefähr die folgende Beziehung erfüllen:

$$(\bar{x} - \mathrm{Mo}) = 3 \cdot (\bar{x} - \mathrm{Md}) .$$

• Wenn wir für eine unimodale Verteilung eine Menge positiver Zahlen betrachten, beträgt der **geometrische Mittelwert** *G* auf die gleiche Weise

Machine Translated by Google

immer kleiner oder gleich dem **arithmetischen Mittel** $\bar{x}$ und immer größer als oder gleich dem **harmonischen Mittel** H. Also wir haben:

$$H \leq G \leq \bar{x}.$$

Diese drei Mittel sind nur dann identisch, wenn alle der Zahlen sind gleich.

DOMÄNEN UND EINSCHRÄNKUNGEN

Das arithmetische Mittel ist ein einfaches Maß des zentralen **Werts** einer Menge quantitativer Beobachtungen. Das Ermitteln des Mittelwerts kann manchmal zu einer schlechten Dateninterpretation führen:

Wenn die monatlichen Gehälter (in Euro) von 5 Personen sind 3000, 3200, 2900, 3500 und 6500, das arithmetische Mittel der Das Gehalt $\frac{19100}{5} = 3820$. Das **bedeutet** gibt uns eine Vorstellung von der Größe des Gehälter beprobt, da es liegt zwischen dem Größten und dem Kleinsten eins. Allerdings sind es 80 % der Gehälter kleiner als der Mittelwert, also in diesem Fall es ist keine besonders gute Darstellung eines typischen Gehalts.

Dieser Fall zeigt, dass wir aufpassen müssen zur Form der Verteilung und der Relia

Fähigkeit der Beobachtungen, bevor wir die verwenden arithmetisches Mittel als **zentrales Maß**

Tendenz zu einem bestimmten Wertesystem. Wenn ein In der Verteilung kommt es zu absurden Beobachtungen, Das arithmetische Mittel könnte einen nicht repräsentativen Wert für die zentrale Tendenz liefern. Wenn einige Beobachtungen berücksichtigt werden

Obwohl es weniger zuverlässig ist als andere, könnte es nützlich sein

um sie weniger wichtig zu machen. Das kann sein erfolgt durch Berechnung einer **gewichteten Arithmetik**

Mittelwert oder durch Verwendung des **Medians,** was nicht der Fall ist stark von irgendwelchen absurden Beobachtungen beeinflusst.

BEISPIELE

Im Unternehmen A beziehen neun Mitarbeiter folgende Monatsgehälter (in Euro):

3000 3200 2900 3440 5050

4150 3150 3300 5200

Das arithmetische Mittel dieser monatlichen Werte Gehälter ist:

$$\bar{x} = \frac{(3000 + 3200 + \cdots + 3300 + 5200)}{9}$$

$$= \frac{33390}{9} = 3710 \text{ Euro.}$$

Wir untersuchen nun einen Fall, in dem die Daten vorhanden sind

in Form einer **Frequenzverteilung dargestellt.**

Die folgende **Häufigkeitstabelle** gibt die Anzahl der Tage, an denen 50 Mitarbeiter beschäftigt waren

während einer Abwesenheit wegen Krankheit während eines Zeitraums von einem Jahr Jahr:

xi : Krankheitstage	fi : Anzahl Mitarbeiter
0	7
1	12
2	19
3	8
4	4
Gesamt	50

Versuchen wir, die mittlere Anzahl zu berechnen Tage, aufgrund derer die Mitarbeiter abwesend waren Erkrankung.

Die Gesamtzahl der Krankheitstage für die 50 Mitarbeiter entsprechen der Summe des Produkts jedes *xi* durch seine jeweilige **Häufigkeit** *fi:*

$$\sum_{i=1}^{5} xi \cdot fi = 0 \cdot 7 + 1 \cdot 12 + 2 \cdot 19 + 3 \cdot 8 + 4 \cdot 4 = 90.$$

Machine Translated by Google

Die Gesamtzahl der Mitarbeiter beträgt:

$$\sum_{i=1}^{5} fi = 7 + 12 + 19 + 8 + 4 = 50 .$$

L Das arithmetische Mittel der Krankenzahl

Tage pro Mitarbeiter beträgt dann:

$$\bar{x} = \frac{\sum_{i=1}^{5} xi \cdot fi}{\sum_{i=1}^{5} fi} = \frac{90}{50} = 1{,}8$$

was bedeutet, dass im Durchschnitt die
50 Mitarbeiter waren 1,8 Tage krankgeschrieben
ness pro Jahr.
Im folgenden Beispiel sind die Daten
in Klassen gruppiert.
Wir wollen das arithmetische Mittel von berechnen
der tägliche Gewinn aus dem Verkauf von 50 Sorten
Lebensmittelgeschäft. Die Häufigkeitsverteilung für die
Lebensmittel sind in der folgenden Tabelle aufgeführt:

Klassen (Gewinne in Euro)	Mitte Punkte xi	Frequenzen fi (Nummer von Lebensmitteln)	xi · fi
500–550	525	3	1575
550–600	575	12	6900
600–650	625	17	10625
650–700	675	8	5400
700–750	725	6	4350
750–800	775	4	3100
Gesamt		50	31950

Das arithmetische Mittel der Gewinne beträgt:

$$\bar{x} = \frac{\sum_{i=1}^{6} xi \cdot fi}{\sum_{i=1}^{6} fi} = \frac{31950}{50} = 639 \quad ,$$

was bedeutet, dass im Durchschnitt jeder von
Die 50 Lebensmittelgeschäfte erwirtschaften einen täglichen Gewinn von
639 Euro.

WEITERLESEN

- **Geometrisches Mittel**
- **Harmonische Mittel**
  - **Bedeuten**
- **Maß für die zentrale Tendenz**
- **Gewichtetes arithmetisches Mittel**

VERWEISE

Plackett, RL: Studien zur Geschichte von
Wahrscheinlichkeit und Statistik. VII. Das Prinzip
des arithmetischen Mittels. Biometrie **45,**
130–135 (1958)

Simpson, T.: Ein Brief an den rechten Ehrenwerten
George Earl of Macclesfield, Präsident von
Die Royal Society nutzte den Vorteil, den Mittelwert
einer Reihe von Beobachtungen zu ermitteln
in der praktischen Astronomie. Philos.Trans.Roy.
Soc. London. **49,** 82–93 (1755)

Simpson, T.: Ein Versuch, den Vorteil aufzuzeigen,
der sich aus der Mittelwertbildung aus einer Reihe
von Beobachtungen in der praktischen Astronomie
ergibt. In: Verschiedene Traktate über einige
Kuriose und sehr interessante Themen
in Mechanik, Physikalischer Astronomie und
Spekulative Mathematik. Nourse, London (1757).
S. 64–75

Arithmetisches Dreieck

Zur Bestimmung dient das Rechendreieck
Binomialkoeffizienten *(a* + b)n bei der Berechnung
der Anzahl möglicher Kombinationen von *k* Objekten
aus insgesamt *n* Objekten
(Ck n).

GESCHICHTE

Die Idee, die Anzahl der Kombinationen von *k* Objekten
aus insgesamt *n* Objekten zu ermitteln
wird in Indien seit dem 9. Jahrhundert erforscht
Jahrhundert. Tatsächlich gibt es Spuren davon in der

Machine Translated by Google

Meru Prastara , geschrieben von Pingala in der Zeit 200 v. Chr.

Zwischen dem vierzehnten und fünfzehnten Jahrhundert turies, al-Kashi, ein Mathematiker aus der Die iranische Stadt Kashan schrieb *The Key to Arithmetik.* In dieser Arbeit nennt er **Binomial** Koeffizienten „Exponentenelemente“.

In seinem *Werk Traité du Triangle Arithmétique* veröffentlicht 1665, Pascal, Blaise (1654) definierte die Zahlen im „arithmetischen Tri „Winkel“ und daher wird dieses Dreieck auch „Winkel“ genannt Pascals Dreieck.

Wir sollten auch beachten, dass das Dreieck war populär gemacht durch Tartaglia, Niccolo Fontana im Jahr 1556, weshalb Italiener es oft als bezeichnen Tartaglias Dreieck, obwohl Tartaglia habe das arithmetische Dreieck nicht wirklich studiert.

MATHEMATISCHE ASPEKTE

Das arithmetische Dreieck hat Folgendes form:

1

1 1

121

13 31

14 6 41

1 5 10 10 5 1

1 6 15 20 15 6 1

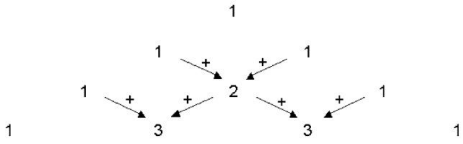
...

Jedes Element ist ein **Binomialkoeffizient**

$$Ck_N = \frac{N!}{k!(n \ddot{y} k)!} = \frac{n \cdot (n \ddot{y} 1) \cdot \quad \cdots \quad \cdot (n \ddot{y} k + 1)}{1 \cdot 2 \cdot \ldots \quad \cdot k} .$$

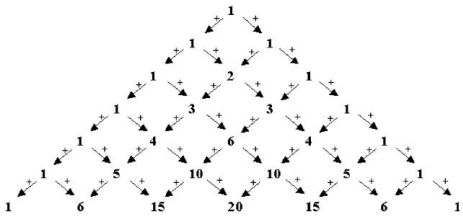
Dieser Koeffizient entspricht dem Element *k* der Geraden *n + 1, k = 0,...,n.*

Jede bestimmte Zahl erhält man durch Addition Fügen Sie die benachbarten Zahlen in der vorherigen Zeile zusammen.



Zum Beispiel:

$$C4_6 = C3_5 + C4_5 = 10 + 5 = 15 .$$



Allgemeiner ausgedrückt gilt die **Beziehung:**

$$Ck_N + Ck_{\tilde{N}}1 = Ck_{n+1} ,$$

Weil:

$$Ck_N + Ck_{\tilde{N}}1 = \frac{N!}{(n \ddot{y} k)! \cdot k!} + \frac{N!}{(n \ddot{y} k \ddot{y} 1)! \cdot (k + 1)!} = \frac{N! \cdot [(k + 1) + (n \ddot{y} k)]}{(n \ddot{y} k)! \cdot (k + 1)!} = \frac{(n + 1)!}{(n \ddot{y} k)! \cdot (k + 1)!} = Ck_{n+1} .$$

WEITERLESEN

Binomial

Binomialverteilung

Kombination

Kombinatorische Analyse

VERWEISE

Pascal, B.: Traité du Triangle Arithmétique (veröffentlicht posthum 1665), Paris (1654)

Pascal, B.: Œuvres, Bde. 1–14. Brun schvicg, L., Boutroux, P., Gazier, F. (Hrsg.)

Machine Translated by Google

LesGrandsEcrivainsdeFrankreich.Hachette,
Paris (1904–1925)

Pascal, B.: Mesnard, J. (Hrsg.) Œuvres complètes.
Bd. 2. Desclée de Brouwer, Paris
(1970)

Rashed, R.: La naissance de l'algèbre. In:
Noël, E. (Hrsg.) Le Matin des Mathémati ciens.
Belin-Radio France, Paris (1985)
Kerl. 12).

Youschkevitch, AP: Les mathématiques
Araber (VIIIème-XVème siècles). Teilweise
Übersetzung von Cazenave, M., Jaouiche, K.
Vrin, Paris (1976)

ARMA-Modelle

ARMA-Modelle (manchmal auch Box-Jenkins-Modelle genannt) sind autoregressive Bewegungen Durchschnittsmodelle, die in der Zeitreihenanalyse verwendet werden. Der autoregressive Teil, bezeichnet als *AR,* besteht aus einer endlichen Linearkombination von frühere Beobachtungen. Der Teil des gleitenden Durchschnitts, *MA,* besteht aus einer endlichen Linearität Kombination in *t* der vorherigen Werte für ein weißes Rauschen (eine Folge voneinander unabhängiger und identisch verteilter Zufallssignale). Variablen).

MATHEMATISCHE ASPEKTE

1. *AR*- Modell (autoregressiv)

   In einem autoregressiven Prozess der Ordnung p, Die aktuelle Beobachtung *yt* wird erzeugt durch ein gewichteter Mittelwert der vergangenen Beobachtungen bis zur p-ten Periode. Dies geschieht in folgender Form:

   AR(1): $yt = ÿ1ytÿ1 + ÿt$,

   AR(2): $yt = ÿ1ytÿ1 + ÿ2ytÿ2 + ÿt$,

   ⋮

   $AR(p): yt = ÿ1ytÿ1 + ÿ2ytÿ2 + ...$

   $+ ÿpytÿp + ÿt$,

   wobei ÿ1, *ÿ2,..., ÿp* die positiven oder sind zu schätzende negative Parameter und *ÿt* ist der Fehlerfaktor, der einer Normalverteilung folgt.

2. *MA*- Modell (gleitender Durchschnitt)

   In einem gleitenden Durchschnittsprozess der Ordnung q, Jede Beobachtung *yt* wird zufällig durch ein **gewichtetes arithmetisches Mittel** bis generiert die qte Periode:

   MA(1): $yt = ÿt ÿ ÿ1ÿtÿ1$

   MA(2): $yt = ÿt ÿ ÿ1ÿtÿ1 ÿ ÿ2ÿtÿ2$

   ...

   $MA(p): yt = ÿt ÿ ÿ1ÿtÿ1 ÿ ÿ2ÿtÿ2$

   $\cdots \quad ÿ \; ÿqÿtÿq$,

   wobei ÿ1, *ÿ2,..., ÿq* positive oder negative Parameter sind und *ÿt* der Gaußsche Zufallsfehler ist.

   Das *MA*- Modell stellt eine Zeitreihe dar zufällig um seinen Mittelwert schwankt Art und Weise, aus der der Begriff hervorgeht „gleitender Durchschnitt“, weil er glättet die Reihe und subtrahiert das weiße Rauschen, das durch die Zufälligkeit des Elements erzeugt wird.

3. *ARMA* -Modell (autoregressives Bewegen). Durchschnittsmodell)

   ARMA- Modelle stellen Prozesse dar, die aus einer Kombination vergangener Werte generiert werden und vergangene Fehler. Sie werden durch die definiert folgende Gleichung:

   ARMA(p, q):

   $yt = ÿ1ytÿ1 + ÿ2ytÿ2 + ...$

   $+ ÿpytÿp + ÿt ÿ ÿ1ÿtÿ1 ÿ ÿ2ÿtÿ2$

   $\cdots \quad ÿ \; ÿqÿtÿq$,

   mit *ÿp = 0, ÿq* = 0 und *(ÿt, t* ÿ Z) ist ein schwaches weißes Rauschen.

Machine Translated by Google

WEITERLESEN

- Zeitfolgen
- Gewichtetes arithmetisches Mittel

VERWEISE

Box, GEP, Jenkins, GM: Zeitreihen
Analyse: Prognose und Kontrolle (Reihe
inTimeSeriesAnalysis).HoldenDay, San
Francisco (1970)

Anordnung

Arrangements sind ein Konzept, das in **der kombinatorischen Analyse zu finden ist.** Die Zahl der Arrangements ist die Zahl von Möglichkeiten, *k* Objekte aus *n* Objekten zu zeichnen wo die Reihenfolge ist, in der sich die Objekte befinden gezogen wird berücksichtigt (im Gegensatz zu **Kombinationen).**

GESCHICHTE

Siehe **kombinatorische Analyse.**

MATHEMATISCHE ASPEKTE

1. *Arrangements ohne Wiederholungen*

Eine Anordnung ohne Wiederholung bezieht sich zu der Situation, in der die Objekte gezeichnet werden werden bei der nächsten Ziehung nicht wieder eingesetzt. Jedes Objekt kann dann nur gezeichnet werden einmal während der *k*- Zeichnungen.

Die Anzahl der Anordnungen von *k* Objekten unter *n* ohne Wiederholung ist gleich:

$$Ak_N = \frac{N!}{(n \ddot{y} k)!} .$$

2. *Arrangements mit Wiederholungen*

Arrangements mit Wiederholung treten auf, wenn Jeder herausgezogene Gegenstand wird wieder hineingelegt für die nächste Zeichnung. Jedes Objekt kann dann *r* mal aus *k* Zeichnungen gezogen werden , *r* = *0, 1,...,k.*

Die Anzahl der Anordnungen von *k* Objekten unter *n* mit Wiederholungen ist gleich *n* bis die Leistung k:

$$Ak_N = nk \quad .$$

BEISPIELE

1. *Arrangements ohne Wiederholungen*

Stellen Sie sich eine Urne mit sechs Kugeln vor, die von 1 bis 6 nummeriert sind. Wir ziehen vier Kugeln heraus aus der Urnenfolge, und wir wollen wissen, wie viele Zahlen möglich sind Bilden Sie sich aus den Zahlen der gezogenen Kugeln. Uns interessiert dann die Anzahl Absprachen (da wir berücksichtigen die Reihenfolge der Bälle) ohne Wiederholung (da jede Kugel nur herausgezogen werden kann einmal) von vier Objekten unter sechs. Wir erhalten:

$$Ak_N = \frac{N!}{(n \ddot{y} k)!} = \frac{6!}{(6 \ddot{y} 4)!} = 360$$

mögliche Vereinbarungen. Daher ist es Es ist möglich, 360 verschiedene Zahlen zu bilden indem Sie vier Zahlen aus den Zahlen 1,2,3,4,5,6 ziehen, wenn jede Zahl dies kann kommen in der vierstelligen Zahl nur einmal vor gebildet.

Lassen Sie uns als zweites Beispiel das untersuchen Arrangements ohne Zweierwiederholungen Buchstaben aus den Buchstaben A, B und C. Mit *n* = 3 und *k* = 2 ergibt sich:

$$Ak_N = \frac{N!}{(n \ddot{y} k)!} = \frac{3!}{(3 \ddot{y} 2)!} = 6 .$$

Wir erhalten dann:

AB, AC, BA, BC, CA, CB.

2. *Arrangements mit Wiederholungen*

Betrachten Sie die gleiche Urne wie zuvor beschrieben. Wir führen vier aufeinanderfolgende Ziehungen durch, aber dieses Mal legen wir jede gezogene Kugel ab zurück in der Urne.

Machine Translated by Google

Wir wollen wissen, wie viele vierstellig sind Nummern (oder Absprachen) sind möglich, wenn Es werden vier Zahlen gezogen.

In diesem Fall untersuchen wir die Anzahl der Arrangements mit Wiederholung (seit Jede Kugel wird zuvor wieder in die Urne gelegt die nächste Zeichnung). Wir erhalten

$$Ak_N = nk = 64 = 1296$$

verschiedene Arrangements. Es ist möglich zu Bilden Sie 1296 vierstellige Zahlen aus dem Zahlen 1,2,3,4,5,6, wenn jede Zahl kann mehr als einmal im vierstelligen Bereich vorkommen Nummer.

Als zweites Beispiel nehmen wir wieder die drei Buchstaben A, B und C und bilden ein Anordnung zweier Buchstaben mit Wiederholungen. Mit $n = 3$ und $k = 2$ gilt:

$$Ak_N = nk = 32 = 9 .$$

Wir erhalten dann:

AA, AB, AC, BA, BB, BC, CA, CB, CC.

WEITERLESEN

Kombination

Kombinatorische Analyse

Permutation

VERWEISE

Siehe **kombinatorische Analyse.**

Zurechenbares Risiko

Das zurechenbare Risiko ist die Differenz zwischen dem **Risiko ,** dem Einzelpersonen ausgesetzt sind, die einem bestimmten Faktor ausgesetzt sind, und dem Risiko, dem Personen ausgesetzt sind, die dies nicht tun ihr ausgesetzt. Dies ist das Gegenteil, um **Risiken zu vermeiden.** Es misst die absolute Wirkung von eine Ursache (d. h. das übermäßige **Risiko** oder Fälle von Erkrankung).

GESCHICHTE

Siehe **Risiko.**

MATHEMATISCHE ASPEKTE

Per Definition haben wir:

zurechenbares Risiko = Risiko für die Exponierten

ÿ Risiko für diejenigen, die nicht exponiert sind .

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Konfidenzintervall eines zuschreibbaren Werts Das Risiko entspricht dem Konfidenzintervall des Unterschieds zwischen den Proportionen pE und pNE, wobei pE und pNE darstellen die Risiken, denen die betroffenen Personen ausgesetzt sind bzw. nicht dem untersuchten Faktor ausgesetzt. Nehmen Sie an , dass nE und *nNE* jeweils die Größe des belichteten bzw. nichtbelichteten Objekts sind Populationen. Dann für ein Konfidenzniveau von (1 ÿ ÿ), ist gegeben durch:

$$(pE \text{ ÿ } pNE) \pm z\ddot{y} \sqrt{\frac{pE\cdot(1\ddot{y}pE)}{nE} + \frac{pNE\cdot(1\ddot{y}pNE)}{nNE}},$$

wobei zÿ der aus der Normalen erhaltene Wert ist **Tabelle** (z. B. für ein Konfidenzintervall). von 95 %, ÿ = 0,05 und zÿ = 1,96). Das Konfidenzintervall für (1 ÿ ÿ) für ein vermeidbares

Das Risiko hat Grenzen, die gegeben sind durch:

$$(pNE \text{ ÿ } pE) \pm z\ddot{y} \sqrt{\frac{pE\cdot(1\ddot{y}pE)}{nE} \cdot \frac{pNE\cdot(1\ddot{y}pNE)}{nNE}} .$$

Hier müssen nE und *nNE* groß sein. Wenn das Konfidenzintervall Null enthält, können wir nicht entscheiden auf das Fehlen eines zurechenbaren Risikos schließen lassen.

BEISPIELE

Als Beispiel betrachten wir eine Studie über die Brustkrebsrisiko bei Frauen durch Rauchen ing:

Machine Translated by Google

Gruppe	Vorfall Rate (/100000 /Jahr)	Auf Risiko zurückzuführen vom Rauchen (ÿ(/100000 /Jahr)
Nonex gestellt	57,0	*57,0* ÿ *57,0* = 0
Passiv Raucher	126.2	126,2 ÿ *57,0* = *69,2*
Aktiv Raucher	138.1	138,1 ÿ *57,0* = *81,1*
Gesamt	114,7	114,7 ÿ *57,0* = *57,7*

Die Risiken, die auf passive und Aktivraucher sind 69 bzw. 81 Jahre alt (/100000 Jahr). Mit anderen Worten, wenn die Exposition gegenüber Tabak wurde entfernt, die Inzidenzrate für aktive Raucher (138/ 100.000 pro Jahr) könnten um reduziert werden 81/100.000 pro Jahr und das für Passivraucher (126/100.000 pro Jahr) um 69/100000 pro Jahr. Die Inzidenzraten in beide Kategorien von Rauchern würden werden entspricht dem Satz für nicht exponierte Frauen (57/100000 pro Jahr). Beachten Sie, dass die Inzidenzrate für nicht exponierte Frauen nicht hoch ist Null, aufgrund des Einflusses anderer Faktoren abgesehen vom Rauchen.

Gruppennr.	Indiv. beobachtet über zwei Jahre	Fälle attrib. Zu Rauchen (für zwei Jahre Zeitraum)	Fälle attrib. Zu Rauchen (pro Jahr)
Nonex gestellt	70160	0,0	0,0
Passiv Rauchers	110860	76,7	38.4
Aktiv Rauchers	118636	96,2	48.1
Gesamt	299656	172,9	86,5

Wir können die Anzahl der Fälle berechnen Brustkrebs, der auf Tabakexposition zurückzuführen ist, indem man die Zahl der pro Jahr beobachteten Personen mit der zurechenbaren Zahl multipliziert Risiko. Durch Teilen der Anzahl der Vorfälle Durch das Rauchen im Zweijahreszeitraum um zwei ergibt sich die Zahl der Fälle pro Jahr auf das Rauchen zurückzuführen, und das können wir Bestimmen Sie dann das durch Rauch verursachte Risiko in der Bevölkerung, bezeichnet als PAR, wie gezeigt im folgenden Beispiel. Die vorherige Tabelle zeigt die Details der Berechnung.

Wir beschreiben hier die Berechnung für Passivraucher. In der zweijährigen Studie Es wurden 110.860 Passivraucher beobachtet. Das Risiko ist auf das Passivrauchen zurückzuführen betrug 69,2/100.000 pro Jahr. Das bedeutet, dass die Anzahl der Fälle, die auf Rauch zurückzuführen sind ing über den Zweijahreszeitraum beträgt (110860 · 69,2)/100000 = *76,7*. Wenn wir die Zahl der auf Passivrauchen zurückzuführenden Fälle pro Jahr berechnen wollen, müssen wir dividieren den letzten Wert um 2, was 38,4 ergibt. Darüber hinaus können wir das zurechenbare Risiko berechnen zum Rauchen pro Jahr einfach durch Teilen Anzahl der Fälle, die auf das Rauchen zurückzuführen sind den Zweijahreszeitraum (172,9) durch die Zahl der Personen, die in diesen zwei Jahren studiert wurden (299656 Personen). Wir erhalten dann das Risiko 57,7/100.000 pro Person auf das Rauchen zurückzuführen Jahr. Wir stellen fest, dass wir das gleiche Ergebnis erzielen können indem man die Differenz zwischen der Summe bildet Inzidenzrate (114,7/100000 pro Jahr, vgl die Beispiele unter den Einträgen **für Vorfälle Rate, Prävalenzrate)** und die Inzidenz Rate der nicht exponierten Gruppe (57,0/100000 pro Jahr).

Das Brustkrebsrisiko ist darauf zurückzuführen Rauchen in der Bevölkerung (PAR) ist das Verhältnis der Anzahl der Fälle von Brustdrüsenerkrankungen cer, der auf die Exposition gegenüber Tabak zurückzuführen ist und die Zahl der Brustkrebsdiagnosen

Machine Translated by Google

Nasein der Bevölkerung (siehe Tabelle oben).
Das zurechenbare Risiko in der Bevölkerung beträgt
22,3 % (38,4/172) für Passivrauchen und
28 % (48,1/172) für aktives Rauchen. Für beide
Expositionsformen liegt sie bei 50,3 % *(22,3 % +).*
28 %). Also die Hälfte der Fälle von Brustkrebs
jedes Jahr in dieser Population diagnostiziert
auf Rauchen (aktiv oder passiv) zurückzuführen.

Gruppenfall	attrib. Zu Rauchen (für einen Zeitraum von zwei Jahren)	Fall attrib. Zu Rauchen (pro Jahr)	PAR
Nonex gestellt	0,0	20	0,0
Passiv Rauch ers	38.4	70	22.3
Aktiv Rauch ers	48.1	82	28.0
Gesamt	86,5	172	50.3

WEITERLESEN

- Vermeidbares Risiko
- Ursache und Wirkung in der Epidemiologie
- Inzidenzrate
- Quoten und Quotenverhältnis
- Prävalenzrate
- Relatives Risiko
- Risiko

VERWEISE

Cornfield, J.: Eine Methode zur Schätzung von Vergleichsraten anhand klinischer Daten. Anwendungen bei Lungen-, Brust- und Krebserkrankungen Gebärmutterhals. J. Natl. Krebsinst. **11,** 1269–75 (1951)

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Kleiner Brauner, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Morabia, A.: L'Épidémiologie Clinique. Ausgaben „Que sais-je?“. Presses Univer Sitaires de France, Paris (1996)

Autokorrelation

Die Autokorrelation, mit *ÿk bezeichnet,* ist ein Maß für die Korrelation einer bestimmten **Zeitreihe** mit derselben Zeitreihe, die um *k* Verzögerungen verzögert ist (der Abstand zwischen den Beobachtungen, die sind so korreliert). Man erhält es durch Division die **Kovarianz** zwischen zwei Beobachtungen, getrennt durch *k* Verzögerungen, einer Zeitreihe (Auto-Kovarianz) durch die **Standardabweichung** von *yt* und ytÿk. Wenn die Autokorrelation berechnet wird Für alle Werte von *k* erhalten wir die **Autokorrela tionsfunktion.** Für eine Zeitreihe, die das nicht tut Wenn sich die Autokorrelationsfunktion mit der Zeit ändert, nimmt sie exponentiell auf 0 ab.

GESCHICHTE

Die erste Forschung zur Autokorrelation, die partielle Autokorrelation und das Korrelogramm wurde in den 1920er und 1930er Jahren von aufgeführt Yule, George, der die Theorie von entwickelt hat autoregressive Prozesse.

MATHEMATISCHE ASPEKTE

Wir definieren die Autokorrelation von Zeitreihen *Yt* von:

$$ÿk = \frac{\text{cov}\ (yt,\ \text{ytÿk})}{ÿytÿytÿk}$$

Machine Translated by Google

$$= \frac{\sum_{t=k+1}^{T} (yt \ddot{y} \bar{y})(yt \ddot{y} k \ddot{y} \bar{y})}{\sum_{t=k+1}^{T} (yt \ddot{y} \bar{y})^2 \sum_{t=k+1}^{T} (ytk \ddot{y} \bar{y})^2}.$$

Hier ist $\bar{y}$ der Mittelwert der berechneten Reihe $T \ddot{y} k$ Verzögerungen, wobei T die Anzahl der Beobachtungen ist.

Wir finden heraus, dass:

$$\ddot{y}0 = 1 \text{ und}$$

$$\ddot{y}k = \ddot{y}\ddot{y}k .$$

Es ist möglich, die Autokorrelation abzuschätzen (bezeichnet mit $\ddot{y}k$), vorausgesetzt, die Anzahl der Beobachtungen ist groß genug ($T > 30$). folgende Formel:

$$\ddot{y}k = \frac{\sum_{t=k+1}^{T} (yt \ddot{y} \bar{y})(yt \ddot{y} k \ddot{y} \bar{y})}{\sum_{t=1}^{T} (yt \ddot{y} \bar{y})^2}.$$

Die partielle Autokorrelationsfunktion für Eine Verzögerung von k Lags ist definiert als die Autokorrelation zwischen yt und ytÿk, der Einfluss anderer Variablen wird um k Lags verschoben (ytÿ1, $yt\ddot{y}2,...,yt\ddot{y}k+1$).

Hypothesentest

Bei der Analyse der Autokorrelationsfunktion einer Zeitreihe kann es hilfreich sein, dies zu wissen die Terme $\ddot{y}k$, die sich deutlich unterscheiden von 0. Das Testen der Hypothese erfolgt dann wie folgt folgt:

$$H0 : \ddot{y}k = 0$$

$$H1 : \ddot{y}k = 0 .$$

Für eine große Stichprobe ($T > 30$) der Koeffizient $\ddot{y}k$ tendiert asymptotisch zu einer **Normalverteilung** mit einem Mittelwert von 0 und einer Standardabweichung ation von $\frac{1}{\ddot{y}T}$. **Der Studententest** basiert auf dem Vergleich eines empirischen t und eines theoretischen t.

Das Konfidenzintervall für den Koeffizientenÿk ist gegeben durch:

$$\ddot{y}k = 0 \pm t\ddot{y}/2 \quad \frac{1}{\ddot{y}T} .$$

Wenn der berechnete Koeffizient $\ddot{y}k$ nicht fällt Innerhalb dieses Konfidenzintervalls unterscheidet es sich deutlich von 0 auf der Ebene ÿ (im Allgemeinen ÿ = 0,05 und tÿ/2 = 1,96).

DOMÄNEN UND EINSCHRÄNKUNGEN

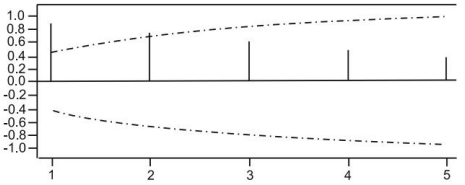
Die Funktion der partiellen Autokorrelation wird hauptsächlich bei Untersuchungen von Zeitreihen verwendet und Genauer gesagt, wenn wir eine anpassen möchten ARMA-Modell. Diese Funktionen werden auch verwendet in der **räumlichen Statistik,** allerdings im Kontext der räumlichen Autokorrelation, bei der wir die Korrelation einer Variablen mit sich selbst untersuchen im Weltraum. Wenn das Vorhandensein eines Phänomens vorliegt Eine bestimmte räumliche Region beeinflusst die Wahrscheinlichkeit, dass das Phänomen dort auftritt In benachbarten Regionen zeigt das Phänomen eine räumliche Autokorrelation. In diesem Fall, Eine positive Autokorrelation tritt auf, wenn die Benachbarte Regionen weisen tendenziell identische Eigenschaften oder ähnliche Werte auf (Beispiele). umfassen homogene Regionen und reguläre Farbverläufe). Es kommt zu einer negativen Autokorrelation wenn die benachbarten Regionen unterschiedliche Qualitäten aufweisen oder zwischen starken und wechselnden Regionen wechseln schwache Werte für das Phänomen. Autocor-Relationsmaße hängen von der Skalierung ab die Variablen, die in der Analyse verwendet werden sowie auf dem Raster, das die Beobachtungen registriert.

BEISPIELE

Wir nehmen als Beispiel den landesweiten Durchschnittslohn in der Schweiz von 1950 bis 1994, alle zwei Jahre gemessen.

Machine Translated by Google

Wir berechnen die Autokorrelationsfunktion zwischen den Daten; Wir möchten eine positive Autokorrelation finden. Die folgenden Zahlen zeigen das Vorhandensein dieser Autokorrelation.



Wir stellen fest, dass die Korrelation signifikant ist Spitzenwerte zwischen der Beobachtung zum Zeitpunkt *t* und die Beobachtung bei ÿ1 und auch zwischen die Beobachtung zum Zeitpunkt *t* und die Beobachtung zum Zeitpunkt *t* ÿ2. Diese Datenkonfiguration ist typisch für einen autoregressiven Prozess. Zuerst Fortwo Werte, wir können sehen, dass diese Autokorrelation ist signifikant, da die Student-Statistik *t* für die *T* = 23 Beobachtungen ergibt sich:

$$\ddot{y}k = 0 \pm 1{,}96 \quad \frac{1}{\ddot{y}\ 23} \cdot$$

Jahr	Nationaler Durchschnittsloh
50	11999
52	12491
54	13696
56	15519
58	17128
60	19948
62	23362
64	26454
66	28231
68	30332
70	33955
72	40320
74	40839
76	37846
78	39507
80	41180

Jahr	Nationaler Durchschnittsloh
82	42108
84	44095
86	48323
88	51584
90	55480
92	54348
94	54316

Quelle: Bundesamt für Statistik

WEITERLESEN

- **Schülertest**
- **Zeitfolgen**

VERWEISE

Bourbonnais, R.: Econométrie, Manuel et exercices corrigés, 2. Aufl. Dunod, Paris (1998)

Box, GEP, Jenkins, GM: Zeitreihen Analyse: Prognose und Kontrolle (Reihe inTimeSeriesAnalysis).HoldenDay, San Francisco (1970)

Chatfield, C.: Die Analyse von Zeitreihen: Eine Einführung, 4. Aufl. Chapman & Halle (1989)

Vermeidbares Risiko

Das vermeidbare Risiko (das natürlich ist „vermeidbar, wenn wir die Auswirkungen der Exposition gegenüber einem bestimmten Phänomen neutralisieren“) ist das Gegenteil des **zurechenbaren Risikos.** Mit anderen Worten, Es ist der Unterschied zwischen der **Risikobegegnung** von nicht exponierten Personen und so weiter denen Personen ausgesetzt sind, die dem ausgesetzt sind Phänomen.

GESCHICHTE

Siehe **Risiko.**

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Per Definition haben wir:

vermeidbares Risiko = Risiko, wenn es nicht ausgesetzt wird

ÿ Risiko bei Exposition .

DOMÄNEN UND EINSCHRÄNKUNGEN

Das vermeidbare Risiko wurde in Ordnung gebracht um die Notwendigkeit zu vermeiden, ein Negativ zu definieren **Zurechenbares Risiko.** Es ermöglicht uns zu rechnen die Anzahl der Patienten, die sein müssen behandelt, weil:

$$\text{Anzahl der Patienten behandelt werden} = \frac{1}{\text{Vermeidbares Risiko}}.$$

Siehe auch **zurechenbares Risiko.**

BEISPIELE

Betrachten Sie als Beispiel eine Studie über die Wirksamkeit eines Arzneimittels zur Behandlung einer Krankheit. Die 223 in die Studie einbezogenen Patienten sind Alle sind gefährdet, an der Krankheit zu erkranken, aber sie habe dies noch nicht getan. Wir trennen sie in zwei Gruppen: Patienten in der ersten Gruppe (114 Patienten) erhielten das Medikament; jene in der zweiten Gruppe (109 Patienten) wurden gegeben ein Placebo. Die Studiendauer betrug zwei Jahre. Insgesamt werden 11 Fälle der Krankheit diagnostiziert in der ersten Gruppe und 27in der Placebogruppe.

Gruppe	Fälle von Erkrankung (A)	Anzahl Patienten ein die Gruppe (B)	Risiko für Die zwei Jahre Zeitraum (A / B In %)
1. Gruppe	11	114	9,6 %
2. Gruppe	27	109	*24,8 %*

Das vermeidbare Risiko aufgrund des Medikaments beträgt also *24,8ÿ 9,6 = 15,2 %* pro zwei Jahre.

WEITERLESEN

- **Zurechenbares Risiko**
- **Ursache und Wirkung in der Epidemiologie**
- **Inzidenzrate**
- **Quoten und Quotenverhältnis**
- **Prävalenzrate**
- **Relatives Risiko**
- **Risiko**

VERWEISE

Cornfield, J.: Eine Methode zur Schätzung von Vergleichsraten anhand klinischer Daten. Anwendungen bei Lungen-, Brust- und Krebserkrankungen Gebärmutterhals. J. Natl. Krebsinst. **11,** 1269–75 (1951)

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Kleiner Brauner, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Morabia, A.: L'Épidémiologie Clinique. Ausgaben „Que sais-je?“. Presses Univer Sitaires de France, Paris (1996)

Machine Translated by Google

Balkendiagramm

Balkendiagramm ist eine Art **quantitatives Diagramm.** Es besteht aus einer Reihe vertikaler oder horizontaler Elemente

Balken gleicher Breite, aber mit Längen relativ zu den dargestellten Größen.

Balkendiagramme werden verwendet, um die **Kategorien** einer **kategorialen qualitativen Variablen** zu vergleichen oder um Datensätze **aus** verschiedenen Jahren oder verschiedenen Orten für eine bestimmte **Variable zu vergleichen.**

GESCHICHTE

Siehe **grafische Darstellung.**

MATHEMATISCHE ASPEKTE

Es muss eine vertikale und eine horizontale Achse geben definiert, um einen vertikalen Balken zu konstruieren Diagramm.

Die horizontale Achse ist in verschiedene Kategorien unterteilt ; Die vertikale Achse zeigt den Wert jeder Kategorie.

Um ein horizontales Balkendiagramm zu erstellen, werden die Achsen einfach invertiert.
Die Balken müssen daher alle die gleiche Breite haben

Es werden nur ihre Längen verglichen.

Zur besseren Verständlichkeit der Grafik können Schattierungen, Schraffuren oder Farben eingesetzt werden.

Domänen und Einschränkungen Ein

Balkendiagramm kann auch zur Darstellung negativer Kategoriewerte verwendet werden. Um dies zu ermöglichen, muss die **Skala** der Achse, die die Kategoriewerte anzeigt, unter Null liegen.

Es gibt verschiedene Arten von Balkendiagrammen. Das oben beschriebene Diagramm wird als einfaches Balkendiagramm bezeichnet. Zum Vergleich mehrerer **Variablen wird ein Diagramm mit mehreren Balken verwendet.**

Ein zusammengesetztes Balkendiagramm ist ein Diagramm mit mehreren Balken, in dem die verschiedenen Datensätze **gestapelt** sind übereinander. Diese Art von Diagramm wird verwendet, wenn die verschiedenen Datensätze miteinander verknüpft werden können in eine Gesamtpopulation zusammengefasst, und wir möchten die Änderungen in den Datensätzen und der Gesamtpopulation im Zeitverlauf vergleichen.

Es gibt eine andere Möglichkeit, die Teilmengen einer Gesamtbevölkerung darzustellen. In diesem Fall stellt die Gesamtpopulation 100 % dar und der für jede Teilmenge angegebene Wert ist ein **Prozentsatz** der Gesamtmenge (siehe auch **Kreisdiagramm).**

BEISPIELE

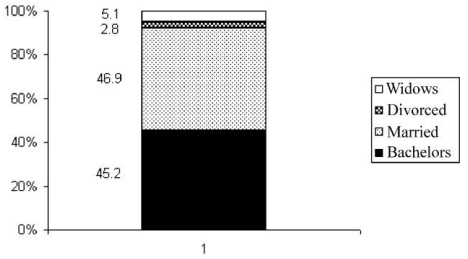
Lassen Sie uns ein *Balkendiagramm erstellen, das in Prozentsätze* für die **Daten** in der folgenden **Häufigkeitstabelle unterteilt ist:**

Machine Translated by Google

Familienstand in einer Stichprobe des Australiers weibliche Bevölkerung am 30.06.1981 (in Tausende)

Ehe Status	Frei quenz	Relativ frei quenz	Prozent Alter
Bachelor	6587.3	0,452	45.2
Verheiratet	6836,8	0,469	46.9
Geschieden	403.5	0,028	2.8
Witwe	748,7	0,051	5.1
Gesamt	14576.3	1.000	100,0

Quelle: ABS (1984) Australian Pocket Year Buch. Australian Bureau of Statistics, Canberra, S. 11



WEITERLESEN

Grafische Darstellung

Quantitatives Diagramm

Barnard, George A.

Barnard, George Alfred wurde 1915 geboren, in Walthamstow, Essex, England. Er erhielt einen Abschluss in Mathematik von Cambridge Universität im Jahr 1936. Zwischen 1942 und 1945 Er arbeitete im Versorgungsministerium als wissenschaftlicher Berater. Barnard trat von 1945 bis 1966 der Mathematikabteilung des Imperial College London bei. Von 1966 bis 1975 Er war Professor für Mathematik an der Universität Universität von Essex und von 1975 bis zu seinem Nach seiner Pensionierung im Jahr 1981 wurde er Professor für Statistik Tics an der University of Waterloo, Kanada.

Barnard, George Alfred erhielt zahlreiche Auszeichnungen, darunter eine Goldmedaille von der Royal Statistical Society und vom Institute of Mathematics and its Applications. In 1987 wurde er zum Ehrenmitglied ernannt das Internationale Statistische Institut. Er starb im August 2002.

Einige Artikel von Barnard, George Alfred:

1954 Stichprobenkontrolle und Statistik Entscheidungen. J. Roy. Stat. Soc. Ser. B 16, 151–174.

1958 ThomasBayes – Eine biografische Notiz. Biometrie 45, 293–315.

1989 Über angebliche Machtgewinne von unten p-Werte. Stat. Med., 8, 1469–1477.

1990 Müssen klinische Studien groß sein? Der Interpretation von p-Werten und der Kombination von Testergebnissen. Stat. Med., 9, 601–614.

Satz von Bayes

Wenn wir die Menge der „Gründe“ betrachten, die Wenn ein **Ereignis** eintritt, gibt der Satz von Bayes eine Formel für die **Wahrscheinlichkeit** an , mit der das Ereignis eintritt direkte Folge eines bestimmten Grundes. Daher kann der Satz von Bayes als Formel für die **bedingte Wahrscheinlichkeit** eines Ereignisses interpretiert werden .

GESCHICHTE

Der Satz von Bayes ist nach **Bayes benannt, Thomas** und wurde Mitte des 18. Jahrhunderts entwickelt. Allerdings Bayes veröffentlichte den Satz zu seinen Lebzeiten nicht; stattdessen wurde es von Price, R. präsentiert. am 23. Dezember 1763, zwei Jahre später seinem Tod an die Royal Society of London,

Machine Translated by Google

in dem Bayes während der Zeit Mitglied war
letzten zwanzig letzten Jahre seines Lebens.

MATHEMATISCHE ASPEKTE

Sei {A1, *A2,...,Ak}eine* Partition des **Probenraums .**
Wir gehen davon aus, dass jedes Ereignis
A1,...,Ak hat eine **Wahrscheinlichkeit ungleich Null.** Lass *E*
sei ein Ereignis mit *P(E)* > 0.
Für jedes i(1 ÿ *i* ÿ k) gilt also der Satz von Bayes
(für den diskreten Fall) ergibt:

$$P(Ai|E) = \frac{P(Ai) \cdot P(E|Ai)}{\sum_{j=1}^{k} P(Aj) \cdot P(E|Aj)}.$$

Im kontinuierlichen Fall, in dem *X* eine **Zufallsvariable**
mit **der Dichtefunktion** *f(x) ist,*
soll auch eine A-priori-Dichtefunktion sein,
Der Satz von Bayes gibt die Dichte a posteriori an
entsprechend

$$f(x|E) = \frac{f(x) \cdot P(E|X = x)}{\int_{-\infty}^{\infty} f(t) \cdot P(E|X = t)dt}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Satz von Bayes war Gegenstand vieler Untersuchungen
Kontroverse im Zusammenhang mit der Fähigkeit, es zu verwenden
wenn die Werte der Wahrscheinlichkeiten verwendet werden
Die **Wahrscheinlichkeitsfunktion** a post riori zu
bestimmen, ist im Allgemeinen nicht präzise festgelegt
Weg.

BEISPIELE

Drei Urnen enthalten rote, weiße und schwarze Kugeln:

- Urne A enthält 5 rote Kugeln, 2 weiße Kugeln
  und 3 schwarze Kugeln;
- Urne B enthält 2 rote Kugeln, 3 weiße Kugeln
  und 1 schwarze Kugel;
- Urne C enthält 5 rote Kugeln, 2 weiße Kugeln
  und 5 schwarze Kugeln.

Wir wählen zufällig eine Urne aus und ziehen eine Kugel
zufällig: es ist weiß. Wir möchten feststellen
die **Wahrscheinlichkeit ,** dass es aus Urne A entnommen wurde.
Sei A1 dem **Ereignis** entsprechen, bei dem wir
„Wähle Urne A“, A2 sei das Ereignis, bei dem wir
„Wähle Urne B“ und A3 sei das Ereignis, bei dem
wir „wählen Urne C.“ {A1, A2, A3}bildet eine Partition
des **Probenraums.**
Sei *E* das Ereignis, bei dem „der Ball genommen wird“.
weiß“, was eine streng positive Wahrscheinlichkeit hat.

Wir haben:

$$P(A1) = P(A2) = P(A3) = \frac{1}{3},$$
$$P(E|A1) = \frac{2}{10}, P(E|A2) = \text{und } \frac{3}{6},$$
$$P(E|A3) = \frac{2}{12}.$$

Die Formel von Bayes ermöglicht es uns, die zu bestimmen
Wahrscheinlichkeit, dass die gezogene weiße Kugel kommt
aus der Urne A:

$$P(A1|E) = \frac{P(A1) \cdot P(E|A1)}{\sum_{i=1}^{3} P(Ai) \cdot P(E|Ai)}$$
$$= \frac{\frac{1}{3} \cdot \frac{2}{10}}{\frac{1}{3} \cdot \frac{2}{10} + \frac{1}{3} \cdot \frac{3}{6} + \frac{1}{3} \cdot \frac{2}{12}} = \frac{3}{13}.$$

WEITERLESEN

Bedingte Wahrscheinlichkeit

Wahrscheinlichkeit

REFERENZ

Bayes, T.: Ein Aufsatz zur Lösung eines Problems in
der Zufallslehre. Philos.
Trans. Roy. Soc. London. **53,** 370–418
(1763). Veröffentlicht auf Veranlassung von
Price, R., 2 Jahre nach seinem Tod.
Neuveröffentlichung mit einer Biographie von Barnard,
George A. im Jahr 1958 und in Pearson, ES,
Kendall, MG: Studien zur Geschichte von
Statistik und Wahrscheinlichkeit. Griffin, London, S.
131–153 (1970)

Machine Translated by Google

Bayes, Thomas

Bayes, Thomas (1702–1761) war der älteste Sohn von Bayes, Joshua, der einer der ersten sechs nonkonformistischen Geistlichen war, die ordiniert wurden in England und war Mitglied der Royal Society. Er wurde privat von Professoren unterrichtet, wie es in nonkonformistischen Familien üblich war. Im Jahr 1731 wurde er Pfarrer die Presbyterianische Kapelle in Tunbridge Wells, einer Stadt etwa 150 km südwestlich von Tunbridge Wells London. Aufgrund einiger religiöser Veröffentlichungen wurde er 1742 zum Fellow der Royal Society gewählt.

Sein Interesse an Mathematik war bekannt zu seinen Zeitgenossen, obwohl er keine technischen Veröffentlichungen verfasst hatte, da er von **De Moivre, A.,** einem der Begründer der Wahrscheinlichkeitstheorie, unterrichtet worden war . Im Jahr 1763 sortierte Price, R. die von Bayes hinterlassenen Papiere und ließ sein Hauptwerk veröffentlichen:

1763 Ein Aufsatz zur Lösung eines Problems in der Zufallslehre.
Philos. Trans. Royal Soc. London, 53, S. 370–418. Neuveröffentlichung mit einer Biografie von Barnard, GA (1958).
In: Pearson, ES und Kendall, M (1970).
Studien zur Geschichte der Statistik und Wahrscheinlichkeit. Griffin, London, S. 131–153.

WEITERLESEN

Satz von Bayes

Bayesianische Statistik

Die Bayesian-Statistik ist ein großer Bereich der Statistik, der sich aufgrund eines Axioms unterscheidet Atisierung der Statistiken, die ihr eine gewisse innere Kohärenz verleiht.

Die Grundidee besteht darin, die **Wahrscheinlichkeit** eines **Ereignisses** so zu interpretieren , wie sie üblicherweise verwendet wird; also als die damit verbundene Unsicherhei Im Gegensatz dazu betrachtet der klassische Ansatz die Wahrscheinlichkeit eines Ereignisses als Grenze der relativen Häufigkeit (einen formaleren Ansatz finden Sie unter **Wahrscheinlichkeit).**
Der bekannteste Aspekt der Bayes'schen **Inferenz** ist die Wahrscheinlichkeit der Berechnung der **gemeinsamen Wahrscheinlichkeitsverteilung** (oder **Dichtefunktion) f(ÿ,** *X = x1,...,X = xn)* eines oder mehrerer Parameter **ÿ** (ein Parameter oder ein Vektor von Parameter) nach Beobachtung der Daten *x1,...,xn,* die unabhängig von einer Zufallsvariablen *X* abgetastet wurden , von der **ÿ** abhängt. (Es ist erwähnenswert, dass wir damit auch die Wahrscheinlichkeitsverteilung für eine neue Beobachtung xn+1 berechnen können .)
Bayesianische Statistiken behandeln die unbekannten Parameter nicht als Zufallsvariablen mögliche Variabilität (in Wirklichkeit gelten die unbekannten Parameter als fest), sondern aufgrund unserer Unkenntnis oder Unsicherheit darüber.

Die Posteriorverteilung f *(ÿ* | X = x1 , ... , _ |ÿ).

Posterior ÿ Prior *x* Wahrscheinlichkeit

Der zweite Ausdruck bereitet keine Probleme, da es sich um eine Funktion handelt, die wir häufig in der klassischen Statistik verwenden und die als Like Lihood bekannt ist (siehe **Maximum Likelihood).**
Im Gegensatz dazu geht der erste Teil von einer Prior-Verteilung für ÿ aus. Wir verwenden oft den Anfangsbuchstaben Verteilung von **ÿ ,** um mögliche Zusatzinformationen über die interessierenden Parameter einzubeziehen. In Ermangelung dieser Informationen mation verwenden wir eine Referenzfunktion, die den Mangel an Informationen (die dann am „objektivsten“ oder „nichtinformativsten“ sind) imitiert

Machine Translated by Google

Funktion, die der allgemeinen, aber nicht genauen Verwendung folgt).

Sobald die Verteilung f(ÿ |*x1,...,xn*) berechnet ist, stehen alle Informationen zu den interessierenden Parametern zur Verfügung. Daher können wir plausible Werte für den unbekannten Parameter (den **Mittelwert,** den **Median** oder ein anderes **Maß der zentralen Zehn)** berechnen

(Denz), seine **Standardabweichung,**

Konfidenzintervalle oder führen Sie Hypothesentests für seinen Wert durch.

GESCHICHTE Siehe Satz **von Bayes, Thomas** und **Bayes.**

MATHEMATISCHE ASPEKTE

Sei *D* der Datensatz *X = x1,...,X = xn,* der unabhängig aus einer Zufallsvariablen *X* unbekannter Verteilung entnommen wurde. Wir werden

Betrachten Sie den einfachen Fall, in dem es nur einen interessanten Parameter, ÿ, gibt, der von X abhängt.

Dann kann ein Bayesianisches Standardverfahren wie folgt ausgedrückt

werden: 1. Identifizieren Sie die bekannten Größen *x1,...,xn.*

2. Geben Sie ein Modell für die Daten an; mit anderen Worten eine parametrische Familie *f (x |* ÿ) von Verteilungen, die die Generierung von Daten beschreiben.
3. Geben Sie die Unsicherheit bezüglich ÿ durch eine anfängliche Verteilungsfunktion *f* (ÿ) an.
4. Anschließend können wir die Verteilung berechnen

f (ÿ |*D)* (Endverteilung *genannt)* unter Verwendung **des Bayes-Theorems.**

Die ersten beiden Punkte sind allen statistischen **Schlussfolgerungen gemeinsam.**

Der dritte Punkt ist problematischer. In Ermangelung zusätzlicher Informationen über ÿ besteht die Idee darin, eine *Referenzverteilung f* (ÿ) zu berechnen, indem eine Funktion maximiert wird, die die fehlenden Informationen über angibt Parameter ÿ. Sobald dieses Problem gelöst ist, lässt sich der vierte Punkt mithilfe des Bayes-Theorems leicht lösen.

Der Satz von Bayes kann in seiner kontinuierlichen Form ausgedrückt werden durch:

$$f(D) = f(D)\, f \frac{f(D|\ddot{y}) \cdot f(\ddot{y})\, f(\ddot{y}|}{(D|\ddot{y}) \cdot f(\ddot{y})\, f} = \frac{(D|\ddot{y})f(\ddot{y})\,d\ddot{y}}{}.$$

Da die *xi* unabhängig sind, können wir schreiben:

$$f(\ddot{y}|D) = \frac{\prod_{i=1}^{N} f(xi|\ddot{y}) \cdot f(\ddot{y})}{\int \prod_{i=1}^{N} f(xi|\ddot{y})f(\ddot{y})\,d\ddot{y}}.$$

Jetzt haben wir die Möglichkeit, die **Höhle** zu berechnen **Sitätsfunktion** einer neuen (unabhängigen) Beobachtung xn+1, gegeben *x1,...,xn:*

$$f(X = xn+1|D)\, f(X = x1,...,X = xn+1)\, f(X = x1,...,X = xn) = \frac{f(X = x1,...,X = xn+1|\ddot{y}) \cdot f(\ddot{y})\,d\ddot{y}}{\int f(D|\ddot{y}) \cdot f(\ddot{y})\,d\ddot{y}} = \frac{\prod_{i=1}^{n+1} f(X = xi|\ddot{y}) \cdot f(\ddot{y})\,d\ddot{y}}{\int \prod_{i=1}^{N} f(X = xi|\ddot{y}) \cdot f(\ddot{y})\,d\ddot{y}} = f(X = xn+1|\ddot{y}) \cdot f(\ddot{y}|D)\,d\ddot{y}.$$

Wir werden nun kurz die Methoden erläutern, die es uns ermöglichen:

• Finden Sie einen Wert für den geschätzten Parameter, der wahrscheinlicher ist als die anderen; • Finden Sie ein Konfidenzintervall für ÿ und; • Führen Sie Hypothesentests durch.

Diese Methoden stehen in engem Zusammenhang mit der Entscheidungstheorie, die in der Bayes'schen Statistik eine erhebliche Rolle spielt.

Machine Translated by Google

Punktschätzung des Parameters

Um eine **Punktschätzung** für den Parameter ÿ zu erhalten, geben wir eine Verlustfunktion l(ÿˆ, ÿ) an ergibt sich aus der Verwendung von ÿˆ (dem geschätzten Wert) anstelle des wahren Wertes (der Unbekannten) ÿ. Dann minimieren wir diese Funktion von ÿˆ. Für Wenn beispielsweise ÿ real ist und „der Verlust quadratisch ist“ (das heißt, l(ÿˆ, ÿ) = (ÿˆÿÿ)2), dann ist der Rest Die Schätzung von ÿ wird der Mittelwert der Berechnung sein kultivierte Verteilung.

Vertrauensintervalle

Das Konzept eines **Konfidenzintervalls** ist in der Bayes'schen Statistik durch das Konzept ersetzt einer ÿ-glaubwürdigen Region (wobei ÿ das „Konfidenzniveau“ ist), die einfach als definiert wird Intervall *I* so, dass:

$$\int_{I(\ddot{y})} f(\ddot{y}\ |D)\, d\ddot{y} = \ddot{y}\ .$$

Oft verlangen wir auch, dass die Breite des Intervall wird minimiert.

Hypothesentest

Der allgemeine Ansatz zum **Testen von Hypothesen:**

$$H0 : \ddot{y} \ddot{y} I \qquad \text{versus } H1 : \ddot{y} / \ddot{y} I \quad .$$

hängt mit der Entscheidungstheorie zusammen. Basierend auf, Wir definieren eine Verlustfunktion *l(a0,* ÿ), die akzeptiert H0 (wobei der wahre Wert des Parameters ist ÿ) und eine Verlustfunktion *l(a1,* ÿ), die H0 ablehnt. Wenn der Wert der wahre Wert ist, der durch erhalten wird Akzeptieren von H0, das heißt,

$$\int l(a0, \ddot{y})\, d\ddot{y}\ ,$$

kleiner ist als derjenige, der durch Ablehnung von H0 erhalten wird, dann können wir H0 akzeptieren. Unter Verwendung dieser Einschränkung lehnen wir die Einschränkungen ab, die ÿ durch die Nullhypothese auferlegt werden (ÿ ÿ I).

BEISPIELE

Im folgenden Beispiel geht es um die Schätzung der Parameter ÿ aus der **Bernoulli-Verteilung** *X* mit Hilfe von *n* unabhängig Beobachtungen *x1,...,xn,* die den Wert 1 in annehmen im Erfolgsfall und 0 im Misserfolgsfall. Sei *r* die Anzahl der Erfolge und *n* ÿ *r* sei die Anzahl der Fehlschläge unter den Beobachtungen. Wir haben dann:

$$L(\ddot{y}) = P\,(X = x1,...,X = xn\ |\ddot{y}) = \prod_{i=1}^{N} P\,(X = xi\ |\ddot{y}) = \ddot{y}r \cdot (1 \ddot{y} \ddot{y})^{n\ddot{y}r} \quad .$$

Eine Schätzung der **maximalen Wahrscheinlichkeit** von ÿ, bezeichnet mit ÿˆmle maximiert diese Funktion. Dazu betrachten wir den Logarithmus dieser Funktion, mit anderen Worten die Log-Likelihood:

$$\log L\,(\ddot{y}) = r\log\,(\ddot{y}) + (n \ddot{y} r)\log\,(1 \ddot{y} \ddot{y}).$$

Wir maximieren dies, indem wir seine Ableitung durch setzen ÿ gleich Null:

$$\frac{\ddot{y}\ (\log L)}{\ddot{y}\ddot{y}} = \frac{R}{\ddot{y}} \quad \frac{n \ddot{y} r}{1 \ddot{y} \ddot{y}} = 0\ .$$

Dies entspricht r(1 ÿ ÿ) = *(n* ÿ r) ÿ, und es vereinfacht sich zu ÿˆmle = $\frac{R}{N}$. Der Schätzer denn die maximale Wahrscheinlichkeit von ÿ ist dann sim Zahlen Sie den Anteil der beobachteten Erfolge. Nun kehren wir zur Bayes'schen Methode zurück. In diesem Fall ist die Bernoulli-Verteilung die Die A-priori-Referenzverteilung des Parameters ÿ wird ausgedrückt durch:

$$f(\ddot{y}\ ) = c \cdot (h\,(\ddot{y}))^{\frac{1}{2}}\ ,$$

wobei *c* eine geeignete Konstante ist (z. B f(ÿ)dÿ = 1) und wobei h(ÿ) heißt

Machine Translated by Google

Fisher-Informationen von X:

$$h(\ddot{y}) = \ddot{y}E\left(\frac{\ddot{y}2L}{\ddot{y}\ddot{y}\,2}\right)$$

$$= \ddot{y}E\,\ddot{y}\,\ddot{y}\,2\left(\frac{X}{\ddot{y}} - \frac{1\,\ddot{y}\,X}{(1\,\ddot{y}\,\ddot{y})^2}\right)$$

$$= \frac{E[X]}{\ddot{y}\,2} + \frac{1\,\ddot{y}\,E[X]}{(1\,\ddot{y}\,\ddot{y})^2}$$

$$= \frac{\ddot{y}}{\ddot{y}\,2\,(1\,\ddot{y}\,\ddot{y})} + \frac{1\,\ddot{y}\,\ddot{y}}{(1\,\ddot{y}\,\ddot{y})^2}$$

$$= \frac{1}{\ddot{y}} + \frac{1}{1\,\ddot{y}\,\ddot{y}} = \frac{1}{\ddot{y}\,(1\,\ddot{y}\,\ddot{y})}.$$

Die Verteilung wird dann sein:

$$f(\ddot{y}) = c \cdot \ddot{y}\ddot{y}1\,2 \cdot \bar{}(1\,\ddot{y}\,\ddot{y})^{\ddot{y}1_2}.$$

Die Verteilungsfunktion von ÿ, gegeben Beobachtungen *x1,...,xn,* können dann ausgedrückt werden durch:

$$f(\ddot{y}\,|X = x1,...,X = xn)$$

$$= \frac{1}{D}\sum_{i=1}^{N} P(X = xi\,|\ddot{y}) \cdot f(\ddot{y})$$

$$= \frac{1}{-} \cdot \ddot{y}r \cdot (1\,\ddot{y}\,\ddot{y})^{n\ddot{y}r} \cdot c \cdot \ddot{y}\ddot{y}1\bar{}2 \cdot (1\,\ddot{y}\,\ddot{y})\ddot{y}1\,2\bar{}$$

$$= \frac{\text{Gleichstrom}}{-} \cdot \ddot{y}r\ddot{y}1\bar{}2 \cdot (1\,\ddot{y}\,\ddot{y})^{n\ddot{y}k\ddot{y}1\,2}$$

$$= \frac{\text{Gleichstrom}}{D} \cdot \ddot{y}(r+1\,2)\ddot{y}1 \cdot (1\,\ddot{y}\,\ddot{y})(n\ddot{y}r+1\,2\bar{})\ddot{y}1.$$

Das ist eine Beta- **Verteilung** mit Parametern

$$\ddot{y} = r + \frac{1}{2} \quad \text{und } \ddot{y} = n\,\ddot{y}\,r + \frac{1}{2},$$

und mit einer Konstante von

$$\frac{c}{D} = \frac{(\ddot{y} + \ddot{y})}{(\ddot{y})\,(\ddot{y})},$$

Wo ist die Gammafunktion (siehe **Gamma Verteilung).**

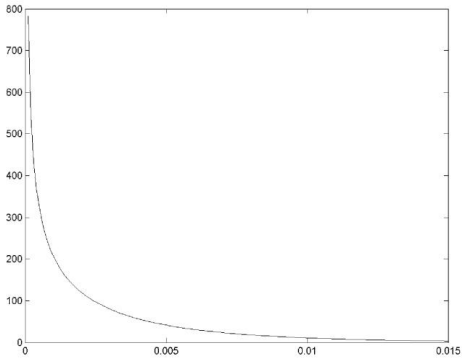
Wir betrachten nun einen konkreten Fall, in dem Wir wollen den Anteil HIV-positiver Studierender abschätzen. Wir testen 200 Studenten und Keiner von ihnen ist HIV-positiv. Der Anteil der HIV-positiven Studierenden wird daher mit maximaler Wahrscheinlichkeit auf 0 geschätzt. Konfidenzintervalle sind nicht sehr nützlich In diesem Fall, weil (wenn wir dem Üblichen folgen Ansatz) werden sie berechnet durch:

$$0 \pm 1{,}96 \cdot \sqrt{p(1\,\ddot{y}\,p)}.$$

In diesem Fall ist *p* der Anteil der beobachteten Bei Erfolgen verringert sich das Konfidenzintervall auf 0.

Nach der Bayes'schen Methodik erhalten wir die Verteilung, nur basierend auf den Daten, die die Unsicherheit über die zu schätzenden Parameter beschreibt. Je größer *n* ist, desto sicherer sind wir bei ÿ; die letzte Referenz Die Verteilung für ÿ ist dann konzentrierter um den wahren Wert von ÿ.

In diesem Fall finden wir als endgültige Referenzverteilung eine Betaverteilung mit Parametern ÿ = *0,5* und ÿ = *200,5,* was zusammenfasst Die Informationen über ÿ (die Werte, die den Spitzen in der Verteilung entsprechen, sind am wahrscheinlichsten). Es ist oft nützlich, solche Ergebnisse grafisch darzustellen:



Daraus können wir Folgendes ersehen:

- Die Wahrscheinlichkeit, dass der Anteil von Zinsen sind kleiner als 0,015, also fast 1;

Machine Translated by Google

• Die Wahrscheinlichkeit, dass der Zinsanteil kleiner als 0,005 ist, beträgt ungefähr 0,84, und; • Der Median, der aufgrund der starken Asymmetrie der Verteilung als bestes Maß für die zentrale Tendenz gewählt wird, beträgt 0,011.

Wir stellen fest, dass es einen qualitativen Unterschied zwischen dem klassischen Ergebnis (der Anteil kann als Null geschätzt werden) und der bayesschen Lösung des Problems gibt, der es uns ermöglicht, die Wahrscheinlichkeitsverteilung des Parameters zu berechnen und die Unsicherheit darüber mathematisch umzuwandeln. Diese Methode sagt uns insbesondere, dass bei bestimmten Informationen die korrekte Schätzung für den Zinsanteil 0,011 beträgt. Beachten Sie, dass die Bayes'sche Schätzung (wie die Unsicherheit) von der Anzahl der beobachteten Fälle abhängt und im klassischen Fall der Beobachtung von 0 HIV-Fällen bei 2, 200 oder 20,0 Schülern entspricht.

WEITERLESEN

Satz von Bayes

Bayes, Thomas

Bedingte Wahrscheinlichkeit

Inferenz

Funktion der Gelenkdichte

Maximale Wahrscheinlichkeit

Wahrscheinlichkeit

REFERENZEN

Bernardo, JM, Smith, AFM: Bayesianische Theorie. Wiley, Chichester (1994)

Berry, DA: Statistik, eine bayesianische Perspektive. Wadsworth, Belmont, Kalifornien (1996)

Box, GEP, Tiao, GP: Bayesianische Inferenz in der statistischen Analyse. Addison-Wesley, Reading, MA (1973)

Carlin, B., Louis, TA: Bayes und empirische Bayes-Methoden. Chapman & Hall /CRC, London (2000)

Finetti, B. de: Wahrscheinlichkeitstheorie; Eine kritische Einführungsbehandlung. Trans. Ma chi, A., Smith, A. Wiley, New York (1975)

Gelman, A., Carlin, JB, Stern HS, Rubin, DB: Bayesianische Datenanalyse. Chapman & Hall /CRC, London (2003)

Lindley, DV: Die Philosophie der Statistik. Statistiker **49,** 293–337 (2000)

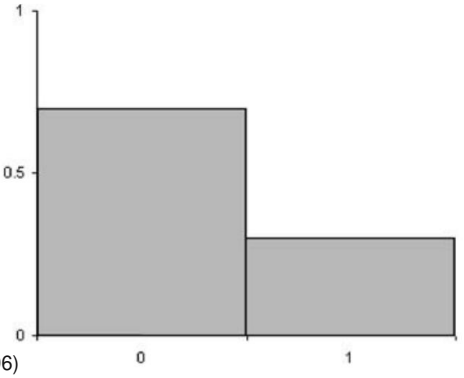
Robert, CP: Le choix Bayesien. Springer, Paris (2006)

Bernoulli-Verteilung

Eine **Zufallsvariable** *X* folgt einem Bernoulli Verteilung mit **Parameter** *p* , wenn ihre **Wahrscheinlichkeitsfunktion** die Form annimmt:

$$P(X = x) = \begin{cases} p & \text{für } x = 1 \\ q = 1 - p & \text{für } x = 0 \end{cases}.$$

wobei *p* und *q* jeweils die **Wahrscheinlichkeiten** von „Erfolg“ und „Misserfolg“ darstellen, symbolisiert durch die **Werte** 1 und 0.



Bernoulli-Gesetz, p = 0,3, q = 0,7

Die Bernoulli-Verteilung ist **diskret Wahrscheinlichkeitsverteilung.**

Machine Translated by Google

GESCHICHTE

Siehe **Binomialverteilung.**

MATHEMATISCHE ASPEKTE

Der **Erwartungswert** der Bernoulli-Verteilung ist per Definition:

$$E[X] = \sum_{x=0}^{1} x \cdot P(X = x) = 1 \cdot p + 0 \cdot q = p .$$

Die **Varianz** der Bernoulli-Verteilung beträgt per Definition:

$$\mathrm{Var}(X) = E[X2] - (E[X])^2 = 12 \cdot p + 02 \cdot q - p2 = p - p2 = pq .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Bernoulli-Verteilung wird verwendet, wenn

Ein Zufallsexperiment hat nur zwei mögliche Ergebnisse: „Erfolg“ oder „Misserfolg“. Diese Ergebnisse werden normalerweise durch 0 und 1 symbolisiert.

WEITERLESEN

Binomialverteilung

Diskrete Wahrscheinlichkeitsverteilung

Familie Bernoulli

Die aus Basel stammende Familie Bernoulli brachte mehrere Mathematiker in die Naturwissenschaften ein. Bernoulli, Jacques (1654–1705), studierte auf Wunsch seines Vaters Theologie an der Universität Basel und reiste anschließend mehrere Jahre, um zu lehren und sein eigenes Studium fortzusetzen. Nachdem er sich entschieden der Mathematik zugewandt hatte, wurde er 1687 Professor an der Universität Basel. Laut Stigler (1986)

Bernoulli, Jacques (in den meisten englischen Werken Bernoulli, James genannt) ist der Vater der Quantifizierung der Unsicherheit.

Erst 1713, sieben Jahre nach seinem Tod, wurde auf Betreiben seines Neffen Nicolas sein Hauptwerk *Ars Conjectandi* veröffentlicht. Dieses Werk ist in vier Teile gegliedert: Im ersten kommentiert der Autor die berühmte Abhandlung von Huygens; der zweite ist der Theorie der **Permutationen** und **Kombinationen gewidmet;** der dritte befasst sich mit der Lösung verschiedener Probleme im Zusammenhang mit Glücksspielen; und schließlich diskutiert der vierte die Anwendung der Wahrscheinlichkeitstheorie **auf** Fragen von moralischem Interesse und Wirtschaftswissenschaften.

Ein großer Teil der Werke von Bernoulli und Jacques wurde nie veröffentlicht.

Jean Bernoulli (1667–1748), der sich mehr für Mathematik als für Medizin interessierte

Nach der von seinem Vater vorgesehenen beruflichen Laufbahn wurde er ebenfalls Universitätsprofessor, zunächst im niederländischen Groningen und dann in Basel, wo er den durch den Tod seines Bruders frei gewordenen Lehrstuhl übernahm.

Die beiden Brüder hatten am Differential gearbeitet und Integralrechnung und Minimierungsprobleme und hatte auch Funktionen studiert.

VERWEISE

Stigler, S.: Die Geschichte der Statistik, die Messung der Unsicherheit vor 1900. Belknap, London (1986), S. 62–71

Bernoulli, Jakob

Bernoulli, Jakob (oder Jacques oder Jacob oder James) (1655–1705) und sein Bruder Jean waren Pioniere der Leibniz-Infinitesimalrechnung. Jakob formulierte das Problem der Berechnung einer Erwartung in eine Wahrscheinlichkeitsrechnung um. Er formulierte auch das schwache Gesetz der großen Zahl.

Machine Translated by Google

Bers, auf denen moderne Wahrscheinlichkeit und Statistiken basieren. Bernoulli, Jakob war 1671 zum *Maître ès lettres* ernannt . Während sein Theologiestudium, das er abbrach 1676 studierte er entgegen dem Willen seines Vaters (Nikolaus Bernoulli) Mathematik und Astronomie.

Im Jahr 1682 veröffentlichte Leibniz (in *Acta Eru ditorium)* eine Methode, die dazu verwendet werden konnte Bestimmen Sie die Integrale algebraischer Funktionen, eine kurze Abhandlung über die Differentialrechnung in einem algorithmischen Schema und einiges mehr Bemerkungen zum Grundgedanken des Integrals Infinitesimalrechnung. Dieses Papier erregte Aufmerksamkeit von Jakob und seinem Bruder, und sie halten es für deutlich, dass sie die bereits geleistete Arbeit deutlich verbessert haben von Leibniz. Leibniz selbst hat das erkannt Die Infinitesimalrechnung wurde größtenteils begründet von den Bernoulli-Brüdern und nicht von ihm selbst. Tatsächlich führte Jakob 1690 den Begriff ein "Integral."
1687 wurde er zum Professor ernannt Mathematik an der Universität Basel, wo er bis zu seinem Tod im Jahr 1705 blieb.
Ars conjectandi ist der Titel dessen, was allgemein als Bernoullis originellstes Werk gilt arbeiten. Es besteht aus vier Teilen: Der erste enthält einen Nachdruck von Huygens *De Ratiociniis in Ludo Aleae* (veröffentlicht 1657), das wird durch wichtige Änderungen vervollständigt. Im zweiten Teil seines Werkes stellt Bernoulli befasst sich mit der kombinatorischen Theorie. Der dritte Teil umfasst 24 Beispiele, die zur Veranschaulichung dienen das modifizierte Konzept des Erwartungswertes. Schließlich ist der vierte Teil der interessanteste und originell, auch wenn Bernoulli dies nicht tat Habe die Zeit, es zu Ende zu bringen. Es ist in diesem Teil das Bernoulli unterscheidet zwei Arten, das Klassische (exakt oder annähernd) zu definieren Maß der Wahrscheinlichkeit.
Um 1680 wurde auch Bernoulli, Jakob Interesse an Stochastik. Die Entwicklung seiner Ideen können in seinem wissenschaftlichen Tagebuch verfolgt werden *Meditationen.*

Einige der Hauptwerke und Artikel von Bernoulli, Jakob umfassen:

1677 Meditationes, Annotationes, Ani madversiones Theologicae & Philo sophicae, von mir JB. concinnatae & Collectae ab anno 1677. Universitäten Bibliothek, Basel, LI a 3.

1713 Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infini Dies ist das Scripta de Ludo der Epistola Gallice Pilae recticularis. Impensis Thurni siorum, Fratrum, Basel.

Bernoulli-Prozess

Die Bernoulli-Versuche sind wiederholte Tests eines **Experiment ,** das den folgenden Regeln gehorcht:

1. Jeder Versuch endet entweder mit Erfolg oder Misserfolg ure;
2. Die **Erfolgswahrscheinlichkeit** ist gleich für jeden Versuch; die Erfolgswahrscheinlichkeit ist mit p bezeichnet und die Ausfallwahrscheinlichkeit mit $q = 1 \ddot{y} p$;
3. Die Studien sind unabhängig.

GESCHICHTE

Die Bernoulli-Prozesse haben ihren Namen der Schweizer Mathematiker **Bernoulli, Jakob** (1713).
Bernoulli, Jakob (1654–1705) war der Älteste von vier Brüdern und es war der Wille seines Vaters dass er Theologie studieren sollte. Als er es getan hatte schloss sein Theologiestudium in Basel ab 1676 verließ er kurzzeitig die Stadt und kehrte dann zurück im Jahr 1680, um sich der Mathematik zu widmen matik. Er erhielt den Lehrstuhl für Mathematik Mathematik an der Universität Basel im Jahr 1687.

Machine Translated by Google

BEISPIELE

Das einfachste Beispiel eines Bernoulli-Prozesses ist das Werfen einer Münze. Wenn der Erhalt von „Kopf“ als Erfolg (S) gewertet wird, während der Erhalt von „Zahl“ als Misserfolg (F) gilt, gilt:

$$p = P(S) = \frac{1}{2}$$

Und

$$q = P(F) = 1 - p = \frac{1}{2}.$$

WEITERLESEN

- **Bernoulli-Verteilung**
- **Binomialverteilung**

REFERENZEN

Bernoulli, J.: ArsConjectandi, Opus Posthu Mama. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Basel (1713)

Ross, SM: Einführung in Wahrscheinlichkeitsmodelle, 8. Aufl. John Wiley, New York (2006)

Satz von Bernoulli

Der Satz von Bernoulli besagt, dass sich die relative **Erfolgshäufigkeit** in einer Folge von Bernoulli-Versuchen der **Erfolgswahrscheinlichkeit** annähert , wenn die Anzahl der Versuche zunimmt Unendlichkeit.

Es ist eine vereinfachte Form des **Gesetzes der großen Zahlen** und leitet sich von der Tschebyscheff-Ungleichung ab.

GESCHICHTE

Der Satz von Bernoulli, manchmal auch als „schwaches Gesetz der großen Zahlen“ bezeichnet, wurde erstmals **von Bernoulli, Jakob** (1713) in seinem Werk *Ars Conjectandi,* das (mit Hilfe seines Neffen Nikolaus) sieben Jahre nach seinem Tod veröffentlicht wurde.

MATHEMATISCHE ASPEKTE

Wenn *S* die Anzahl der während *n* Bernoulli-Versuchen erzielten Erfolge darstellt und *p* die Erfolgswahrscheinlichkeit ist , dann gilt:

$$\lim_{n\to\infty} P\left(\left|\frac{S}{N} - p\right| > \varepsilon\right) = 0 \quad ,$$

oder

$$\lim_{n\to\infty} P\left(\frac{S}{N} - p < \varepsilon\right) = 1 \quad ,$$

wobei $\varepsilon > 0$ und beliebig klein ist.
In äquivalenter Weise können wir schreiben:

$$\frac{S}{N} \to p \, ,$$

Das bedeutet, dass die relative **Erfolgshäufigkeit** tendenziell zur Erfolgswahrscheinlichkeit tendiert, wenn *n* gegen Unendlich geht.

WEITERLESEN

- **Bernoulli-Verteilung**
- **Bernoulli-Prozess**
- **Konvergenz**
- **Gesetz der großen Zahlen**

VERWEISE

Bernoulli, J.: ArsConjectandi, Opus Posthu mum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Basel (1713)

Rényi, A.: Wahrscheinlichkeitstheorie. Nordholland, Amsterdam (1970)

Machine Translated by Google

Beta-Verteilung

Eine **Zufallsvariable** *X* folgt einer Betaverteilung

Bution mit **den Parametern** ÿ und ÿ, wenn ihre **Dichtefunktion** die Form hat:

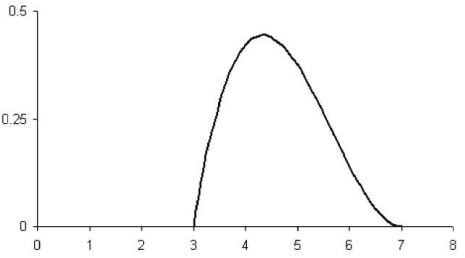
$$f(x) = \frac{1}{B(ÿ, ÿ)(x ÿ a)\, ÿÿ1(b ÿ x)\, ÿ(ÿ+ÿÿ1)} \cdot (b ÿ a) \quad ÿÿ1 \quad ,$$

$$a ÿ x ÿ b \quad , \quad ÿ > 0 \text{ und } ÿ > 0 \quad ,$$

Wo:

$$B(ÿ, ÿ) = \int_0^1 t\, ÿÿ1(1 ÿ t)\, ÿÿ1 dt = \frac{(ÿ)(ÿ)}{(ÿ + ÿ)} ,$$

ist die Gammafunktion (siehe **Gammaverteilung).**



Betaverteilung ÿ = 2, ÿ = 3, a = 3, b = 7

Die Betaverteilung ist eine **kontinuierliche Wahrscheinlichkeitsverteilung.**

MATHEMATISCHE ASPEKTE

Die **Dichtefunktion** des Standard-Beta

Die Verteilung wird durch Ausführen der erhalten **Variablenänderung** $Y = \frac{Xÿa}{_bÿa}$:

$$f(y) = \begin{cases} ÿ \frac{1}{B(ÿ,ÿ)yÿÿ1} \cdot (1 ÿ y)ÿÿ1 & \text{wenn } 0 < y < 1 \\ ÿÿÿ 0 & \text{wenn nicht} \end{cases} .$$

Betrachten Sie X, eine **Zufallsvariable ,** die folgt die Standard-Beta-Distribution. Das **erwartete Wert und Varianz von X** sind jeweils gegeben von:

$$E[X] = ÿ \frac{ÿ}{+ ÿ} ,$$

$$\text{Var}(X) = \frac{ÿ \cdot ÿ}{(ÿ + ÿ)2 \cdot (ÿ + ÿ + 1)} .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Beta-Distribution ist eine der häufigsten

Wird häufig zur Anpassung empirischer Verteilungen verwendet

wobei **der Bereich** (oder Variationsintervall) *[a,* b] ist bekannt.

Hier sind einige besondere Fälle, in denen die Beta Verteilung verwendet wird, bezogen auf andere **kontinuierliche Wahrscheinlichkeitsverteilungen:**

- *Wenn* X1 und X2 zwei **unabhängige Zufallszahlen** sind **Variablen** jeweils verteilt nach eine **Gammaverteilung** mit **Parametern** (ÿ1, 1) bzw. (ÿ2, 1), die Zufallsvariable

$$\frac{X1}{X1 + X2} ,$$

ist gemäß einer Betaverteilung mit Parametern (ÿ1, ÿ2) verteilt.

- Die Betaverteilung wird **einheitlich Verteilung** wann

$$ÿ = ÿ = 1 \quad .$$

- Wenn die **Parameter** ÿ und ÿ gegen Unendlich tendieren, tendiert die Beta-Verteilung in Richtung der Standard- **Normalverteilung tion.**

WEITERLESEN

Kontinuierliche Wahrscheinlichkeitsverteilung
Gammaverteilung
Normalverteilung
Gleichmäßige Verteilung

Machine Translated by Google

Voreingenommenheit

Aus statistischer Sicht ist der Bias definiert als die Differenz zwischen dem **erwarteten Wert** einer **Statistik** und dem wahren **Wert** des entsprechenden **Parameters.** Daher ist der Bias ein Maß für den systematischen Fehler eines **Schätzers.** Wenn wir den Mittelwert von berechnen

Mit einer großen Anzahl unvoreingenommener Schätzungen finden wir den richtigen Wert. Die Voreingenommenheit zeigt der Abstand des Schätzers vom Wahren Wert des Parameters.

GESCHICHTE Das Konzept eines unverzerrten **Schätzers** stammt von **Gauss, CF** (1821), als er an der Methode der **kleinsten Quadrate** arbeitete .

DOMÄNEN UND EINSCHRÄNKUNGEN

Wir sollten die Voreingenommenheit einer Schätzung nicht verwechseln

Parameter eines Parameters mit seinem Genauigkeitsgrad, der ein Maß für die Stichprobe ist Fehler.

Es gibt verschiedene Arten von Bias: Selektionsbias (aufgrund systematischer Unterschiede zwischen den verglichenen Gruppen), Ausschlussbias (aufgrund des systematischen Ausschlusses bestimmter Personen aus der Studie) oder analytischer Bias (aufgrund der Art und Weise, wie die Ergebnisse ausgewertet werden).

MATHEMATISCHE ASPEKTE

Betrachten Sie eine **Statistik** *T*, die zur Schätzung verwendet wird ein **Parameter** ÿ. Wenn *E[T]* = ÿ + b(ÿ) (wobei *E[T]* den **erwarteten Wert** von T darstellt), dann wird die Größe b(ÿ) als Bias der Statistik T bezeichnet.

Wenn b(ÿ) = 0, gilt *E[T]* = ÿ und *T* ist ein erwartungstreuer **Schätzer** von ÿ.

BEISPIELE

Betrachten Sie X1 , _ _ _ _ _ _ _ _ Wir kann den Bias des **Schätzers** berechnen

$$S2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2 ,$$

wird verwendet, um die Varianz ÿ2 der **Grundgesamtheit** auf folgende Weise zu schätzen:

$$E[S2] = E\left[\frac{1}{N}\sum_{i=1}^{N}(xi \ddot{y} \bar{x})^2\right]$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N}(xi \ddot{y} \ddot{y})2 \ddot{y} (x\bar{} \ddot{y} \ddot{y})2\right]$$

$$= \frac{1}{N} E\left[\sum_{i=1}^{N}(xi \ddot{y} \ddot{y})2 \ddot{y} E(x\bar{} \ddot{y} \ddot{y})2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E(xi \ddot{y} \ddot{y})2 \ddot{y} E(x\bar{} \ddot{y} \ddot{y})2$$

$$= \text{Var}(xi) \ddot{y} \text{Var}(x\bar{})$$

$$= \ddot{y}2 \ddot{y} \frac{\ddot{y}2}{N} .$$

Der Bias von S2 ist dann gleich $\frac{\ddot{y}\ddot{y}2}{N}$.

WEITERLESEN

- **Einschätzung**
- **Schätzer**
- **Erwarteter Wert**

VERWEISE

Gauss, CF: Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, Teile 1, 2 und suppl. Werke 4, 1–108 (1821, 1823, 1826)

Machine Translated by Google

Gauss, CF: Méthode des Moindres Car
rés. Erinnerungen an die Kombination des
Beobachtungen. Französische Übersetzung von J.
Bertrand. Mallet-Bachelier, Paris (1855)

Bienaymé, Irénée-Jules

Bienaymé, Irenée-Jules (1796–1878)
trat in das französische Finanzministerium ein und
wurde Generalinspektor für Finanzen
im Jahr 1836, obwohl er seine Anstellung verlor
im Jahr 1848 aus politischen Gründen. Kurz danach
Daraufhin begann er, an der Fakultät Unterricht zu geben
der Wissenschaften in Paris.

Ein Anhänger von **Laplace, Pierre Simon
de,** Bienaymé bewies das Bienaymé–
Tschebyscheff-Ungleichung einige Jahre zuvor
Tschebyschew, Pafnutii Lwowitsch. Die überarbeitete
Fassung von Bienaymés Artikel aus dem Jahr 1867
geht der französischen Version von Chebychev voraus
nachweisen. Diese Ungleichung wurde dann verwendet
Tschebyschews unvollständiger **Zentralbeweis
Grenzwertsatz,** der später von fertiggestellt wurde
Markow, Andrei Andrejewitsch.
Darüber hinaus hat Bienaymé richtig formuliert
der Satz für Verzweigungsprozesse im Jahr 1845.
Sein bekanntestes öffentliches Werk ist wohl das
Korrekturen, die er an der Verwendung von Duvillard vornahm
Sterbetafel.

Einige Hauptwerke und Artikel von Bien aymé, Irénée-Jules:

1853 Überlegungen zu Laplaces Entdeckungen ergaben, dass
es sich um eine Wahrscheinlichkeitsrechnung
handelte, die sich aus der Methode des kleinsten Autos ergab
rés. Comptes Rendus de l'Académie
des Sciences, Paris **37,** 5–13; überarbeitet
1867 im Journal de Liouville
vor dem Beweis des Bienaymé–
Tschebyscheff-Ungleichung in J. Math.
Rein. Appl., 12, 158–176.

Binärdaten

Binäre Daten treten auf, **wenn die interessierende
Variable** nur zwei Werte annehmen kann . Diese beiden
Werte werden im Allgemeinen gerade durch 0 und 1 dargestellt
wenn die Variable nicht quantitativ ist.
Geschlecht, das Vorhandensein oder Fehlen eines
Merkmals und der Erfolg oder Misserfolg eines
Experiments sind nur einige Beispiele für **Variablen**
die zu Binärdaten führen. Diese Variablen sind
sogenannte **dichotome Variablen.**

BEISPIELE

Ein Meteorologe möchte wissen, wie zuverlässig
seine Prognosen sind. Dazu untersucht er eine
Zufallsvariable, die die Vorhersage darstellt.
Diese **Variable** kann nur zwei **Werte annehmen:**

$$X = \begin{cases} 0, \text{ wenn die Vorhersage falsch war} \\ 1, \text{ wenn die Vorhersage richtig war} \end{cases}.$$

Der Meteorologe macht Vorhersagen für
ein Zeitraum von 50 aufeinanderfolgenden Tagen. Die
Vorhersage erwies sich 32 Mal als richtig und 18 Mal
als falsch.

Um herauszufinden, ob seine Vorhersagen besser sind
als diejenigen, die hätten erhalten werden können
indem man eine Münze wirft und das Wetter vorhersagt
basierend darauf, ob Kopf oder Zahl erzielt wird,
er beschließt, die **Nullhypothese** H0 zu testen ,
dass der Anteil *p* richtiger Vorhersagen ist
gleich 0,5, entgegen der **Alternativhypothese** H1,
dass sich dieses Verhältnis von 0,5 unterscheidet:

$$H0 : p = 0{,}5$$
$$H1 : p = 0{,}5 .$$

Berechnen wir den **Wert** von

$$ÿ2 = \frac{\text{beobachtete Häufigkeit} \; ÿ\text{theoretische Häufigkeit}2}{\text{theoretische Frequenz}},$$

Machine Translated by Google

wobei die beobachteten **Häufigkeiten** 18 betragen und 32, und die theoretischen Frequenzen sind: jeweils 25 und 25:

$$ÿ2 = \frac{(32 ÿ 25)2 + (18 ÿ 25)2}{25}$$

$$= \frac{49 + 49}{25} = \frac{98}{25} = 3{,}92 .$$

Unter der Annahme, dass der **zentrale Grenzwertsatz** gilt, vergleichen wir diesen **Wert** mit der Wert der **Chi-Quadrat-Verteilung** mit einem **Freiheitsgrad.** Für ein **Signifikanzniveau** von 5 % finden wir im **Chi-Quadrat Tisch:**

$$ÿ2_{1,0,95} = 3{,}84 .$$

Daher gilt, da 3,92 > 3,84, die **Null Hypothese** wird abgelehnt, was bedeutet, dass Der Meteorologe sagt das Wetter besser voraus als eine Münze.

WEITERLESEN

- **Bernoulli-Verteilung**
- **Kategoriale Daten**
- **Kontingenztabelle**
- **Daten**
- **Dichotome Variable**
- **Likelihood-Ratio-Test**
- **Logistische Regression**
- **Variable**

VERWEISE

Cox, DR, Snell, EJ: Die Analyse von Binärdaten. Chapman & Hall (1989)

Bishop, YMM, Fienberg, SE, Holland, PW: Diskrete multivariate Analyse: Theorie und Praxis. MIT Press, Cambridge, MA (1975)

Binomial

B

Algebraische Summen, die **Variablen** enthalten , sind sogenannte Polynome (vom griechischen „poly“) bedeutet „mehrere“). Ein Ausdruck, der zwei Terme enthält, wird Binomial genannt (aus dem lateinisch „bi“, was „doppelt“ bedeutet). Ein Monom (vom griechischen „mono“, was „einzigartig“ bedeutet) ist ein Ausdruck mit einem Begriff und einem Trinom (vom lateinischen „tri“, was „dreifach“ bedeutet) enthält drei Elemente.

GESCHICHTE

Siehe **Rechendreieck.**

MATHEMATISCHE ASPEKTE

Das Quadrat eines Binomials lässt sich leicht berechnen verwenden

$$(a + b)^2 = a2 + 2ab + b2 .$$

Binomialformeln für höhere Exponenten

Es gibt auch mehr als zwei, wie zum Beispiel:

$$(a + b)^3 = a3 + 3a2b + 3ab2 + b3 ,$$

$$(a + b)^4 = a4 + 4a3\,b + 6a2\,b2 + 4ab3 + b4 .$$

Wir können eine verallgemeinerte Binomialformel schreiben auf folgende Art:

$$(a + b)^N = an + nanÿ1b + \frac{n(n ÿ 1)}{2} anÿ2b2 + \cdots + 2^{\frac{n(n ÿ 1)}{}} a2bnÿ2 + nabnÿ1 + bn$$

$$= C0_N\, ein + C1_{n/A}^{nÿ1}\, b + C2\, nanÿ2b2 + \cdots + Cnÿ1\, abnÿ1 + Cn\, nbn$$

$$= \sum_{k=0}^{N} Ck\, nanÿk\, bk .$$

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Das mit *n* potenzierte Binomial *(p* + q) gibt der **Binomialverteilung** ihren Namen

weil es der Gesamtwahrscheinlichkeit entspricht,

die nach *n* **Bernoulli-Versuchen erhalten wurde.**

WEITERLESEN

Arithmetisches Dreieck
Binomialverteilung
Kombination

Binomialverteilung

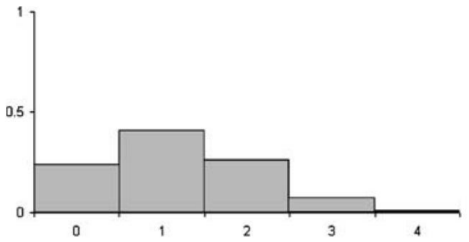
Eine **Zufallsvariable** *X* folgt einem Binomial

Verteilung mit **den Parametern** *n* und *p* , wenn ihre **Wahrscheinlichkeitsfunktion** die Form annimmt:

$$P(X = x) = Cx \cdot px_N \, qnÿx ,$$

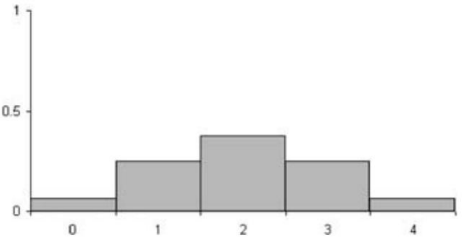
$$x = 0, 1, 2,...,n.$$

Wenn also ein Ereignis *x* „Erfolge“ und *(n* ÿ x) „Misserfolge“ umfasst, wobei *p* die **Wahrscheinlichkeit** des „Erfolgs“ und *q* = 1 ÿ *p* die Wahrscheinlichkeit des „Misserfolgs“ ist, ermöglicht uns die Binomialverteilung die Berechnung die Wahrscheinlichkeit, bei *n* unabhängigen Versuchen *x* Erfolge zu erzielen .



Binomialverteilung, p = 0,3, q = 0,7, n = 3

Die Binomialverteilung mit **den Parametern** *n* und p, bezeichnet als *B(n,* p), ist eine **diskrete Wahrscheinlichkeitsverteilung.**



Binomialverteilung, p = 0,5, q = 0,5, n = 4

GESCHICHTE

Die Binomialverteilung ist eine der alten bekannteste **Wahrscheinlichkeitsverteilungen.** Es wurde von **Bernoulli, J.** in seinem Werk mit dem Titel *Ars Conjectandi* (1713) entdeckt . Dieses Werk ist in vier Teile gegliedert: Im ersten kommentiert der Autor die Abhandlung von Huygens; der zweite Teil ist der Theorie der **Permutationen** und **Kombinationen gewidmet;** der dritte ist der Lösung verschiedener Probleme im Zusammenhang mit Glücksspielen gewidmet; Schließlich schlägt er im vierten Teil vor, die Wahrscheinlichkeitstheorie auf moralische Fragen und die Wirtschaftswissenschaft anzuwenden.

MATHEMATISCHE ASPEKTE

Wenn X1, *X2,...,Xn n* unabhängige **Zufallsvariablen** sind , die einer **Bernoulli-Verteilung** mit einem **Parameter** p folgen, dann ist die **Zufallsvariable**

$$X = X1 + X2 + ... + Xn$$

folgt einer Binomialverteilung *B(n,* p).
Um den **erwarteten** Wert von X zu berechnen, wird die folgende Eigenschaft verwendet, wobei *Y* und *Z* zwei **Zufallsvariablen sind:**

$$E[Y + Z] = E[Y] + E[Z] .$$

Wir haben daher:

$$E[X] = E[X1 + X2 + ... + Xn]$$
$$= E[X1] + E[X2] + ... + E[Xn]$$
$$= p + p + ... + p = np .$$

Machine Translated by Google

Um die **Varianz** *von X zu berechnen,* gehen Sie wie folgt vor Die Eigenschaft wird verwendet, wobei *Y* und *Z* zwei sind unabhängige **Variablen:**

$$\mathrm{Var}(Y + Z) = \mathrm{Var}(Y) + \mathrm{Var}(Z) .$$

Wir haben daher:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(X1 + X2 + \ldots + Xn) \\ &= \mathrm{Var}(X1) + \mathrm{Var}(X2) + \ldots \\ &\quad + \mathrm{Var}(Xn) \\ &= pq + pq + \ldots + pq = npq .\end{aligned}$$

Binomialverteilungstabellen (für die **Wahrscheinlichkeitsverteilung** und die **Verteilung Funktion)** wurden von Rao veröffentlicht et al. (1985) sowie vom National Bureau of Standards (1950). Erweitert Verteilungsfunktionstabellen finden Sie in die *Annalen des Computerlabors* (1955).

BEISPIELE

Eine Münze wird zehnmal umgeworfen. Betrachten Sie **die Zufallsvariable** X, die angibt, wie oft die Münze auf „Zahl“ landet.
Wir haben daher:

Anzahl von Versuchen:	$n = 10$
Wahrscheinlichkeit von ein Erfolg:	$p = \frac{1}{2}$ (Schwänze erhalten)
Wahrscheinlichkeit von eins Versagen:	$q = \frac{1}{2}$ (Köpfe erhalten)

Die **Wahrscheinlichkeit ,** *x* -mal Zahl zu erhalten unter den zehn Versuchen wird angegeben von

$$P(X = x) = C_{10}^{x} \cdot \left(\frac{1}{2}\right)^{x} \cdot \left(\frac{1}{2}\right)^{10-x} .$$

Die **Wahrscheinlichkeit ,** Zahl genau zu erhalten achtmal ist also gleich:

$$\begin{aligned}P(X = 8) &= C_{10}^{8} \cdot p^{8} \cdot q^{10-8} \\ &= \frac{10!}{8!(10-8)!} \cdot \left(\frac{1}{2}\right)^{8} \cdot \left(\frac{1}{2}\right)^{2} \\ &= 0{,}0439 .\end{aligned}$$

Die **Zufallsvariable** *X* folgt dem Bino mialverteilung $B(10, \frac{1}{2})$.

WEITERLESEN

Bernoulli-Verteilung
Binomialtabelle
Diskrete Wahrscheinlichkeitsverteilung
Negative Binomialverteilung

VERWEISE

Bernoulli, J.: ArsConjectandi, Opus Posthu mum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Basel (1713)

Harvard University: Tabellen der kumulativen binomialen Wahrscheinlichkeitsverteilung, Bd. 35. Annals of the Computation Lab oratory, Harvard University Press, Cam Bridge, MA (1955)

National Bureau of Standards.: Tabellen von die binomiale Wahrscheinlichkeitsverteilung. US-Handelsministerium. Angewandt Mathematikreihe 6 (1950)

Rao, CR, Mitra, SK, Matthai, A., Rama Murthy, KG: Formeln und Tabellen für Statistische Arbeit. Statistische Veröffentlichung Company, Kalkutta, S. 34–37 (1966)

Binomialtabelle

Die Binomialtabelle gibt die Werte für an die **Verteilungsfunktion** eines **Zufalls**

Machine Translated by Google

Variable , die einer **Binomialverteilung** folgt **Bution.**

GESCHICHTE

Siehe **Binomialverteilung.**

MATHEMATISCHE ASPEKTE

Die **Zufallsvariable** *X* folge dem **Bino mialverteilung** mit den Parametern *n* und p.

Seine **Wahrscheinlichkeitsfunktion** ist gegeben durch:

$$P(X = x) = Cx_N \cdot px \cdot qnyx ,$$

$$x = 0, 1, 2,...,n,$$

wo Cx_N ist der Binomialkoeffizient, gleich Zu $\frac{N!}{x!(nyx)!}$, Parameter *p* ist die **Wahrscheinlichkeit** vor Erfolg, und *q* = 1yp ist das Komplementär Wahrscheinlichkeit, die der Ausfallwahrscheinlichkeit entspricht (siehe **Normalverteilung).**

Die **Verteilungsfunktion** des Zufalls Variable *X* ist definiert durch:

$$P(X \ddot{y} x) = \sum_{ich=0}^{x} Ci_N \cdot pi \cdot qnyi ,$$

$$0 \ddot{y} x \ddot{y} n .$$

Die Binomialtabelle gibt den Wert von an P(X ÿ x) für verschiedene Kombinationen von x, *n* und p.

Zum Vergrößern wird diese Berechnung mühsam. Zum Glück können wir einiges sehr Gutes gebrauchen stattdessen Näherungen. Wenn min(np, n(1 ÿ p)) > 10, wir können es mit dem annähern Normalverteilung:

$$P(X \ddot{y} x) = \ddot{y} \frac{x + 2^{\frac{1}{-}} \ddot{y} np}{\ddot{y}npq} ,$$

wobei ÿ die Verteilungsfunktion für ist Standardnormalverteilung und die Kontinuitätskorrektur $\frac{1}{2}$ ist enthalten.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Binomialtabelle wird verwendet, um nichtparametrische Tests für **Statistiken** durchzuführen, die gemäß **der Binomialverteilung** verteilt sind , insbesondere den **Vorzeichentest** und den **Binomialtest.**

Das National Bureau of Standards (1950) veröffentlichte individuelle und kumulative Binomialverteilungswahrscheinlichkeiten für *n* ÿ 49, während kumulative Binomialverteilung Wahrscheinlichkeiten für *n* ÿ 1000 sind im angegeben *Annalen des Computerlabors* (1955).

BEISPIELE

Siehe Anhang B.

Wir können das für *n* = 2 und *p* = 0,5 überprüfen:

$$P(X \ddot{y} 1) = \sum_{ich=0}^{1} Ci\ 2(0,5)^{\ddot{}}(0,5)^{2\ddot{y}i}$$

$$= 0,75 .$$

oder das für *n* = 5 und *p* = 0,05:

$$P(X \ddot{y} 3) = \sum_{ich=0}^{3} Ci\ 5(0,05)^{\ddot{}}(0,95)^{5\ddot{y}i}$$

$$= 1,0000 .$$

Ein Beispiel für die Anwendung der Binomialtabelle finden Sie unter **Binomialtest.**

WEITERLESEN

Binomialverteilung

Binomialtest

Zeichentest

Statistische Tabelle

VERWEISE

National Bureau of Standards.: Tabellen von die binomiale Wahrscheinlichkeitsverteilung. US-Handelsministerium. Angewandt Mathematikreihe 6 (1950)

Machine Translated by Google

Harvard University: Tables of the Cumu lative Binomial Probability Distribution, vol. 35. Annals of the Computation Lab oratory, Harvard University Press, Cam Bridge, MA (1955)

Binomialtest

Der Binomialtest ist ein **parametrischer Hypothesentest ,** der angewendet wird, wenn die **Grundgesamtheit** in zwei Klassen unterteilt werden kann: jeder **Beobachter**

Die Bevölkerung dieser Bevölkerung wird der einen oder anderen dieser beiden Kategorien angehören.

MATHEMATISCHE ASPEKTE

Wir betrachten eine **Stichprobe** von *n* unabhängigen Versuchen. Jeder Versuch gehört entweder zur Klasse C1 oder zur *Klasse C2.* Wir notieren die Anzahl der Beobachtungen n1 , die in C1 fallen , und die Anzahl der Beobachtungen n2 , die in C2 fallen .

Jeder Versuch hat **eine Wahrscheinlichkeit** *p ,* zur Klasse C1 zu gehören, wobei *p* für alle *n* Versuche identisch ist , und eine Wahrscheinlichkeit *q* = 1 ÿ *p ,* zur Klasse C2 zu gehören .

Hypothesen

Der Binomialtest kann entweder **zweiseitig** sein **Test** oder ein **einseitiger Test.** Wenn p0 der angenommene Wert von p ist (0 ÿ p0 ÿ 1), werden die Hypothesen wie folgt ausgedrückt:

A: Zweiseitiger Fall

$$H0 : p = p0 , H1 : p = p0 .$$

B: Einseitiger Fall

$$H0 : p \ddot{y} p0 ,$$

$$H1 : p > p0 .$$

C: Einseitiger Fall

$$H0 : p \ddot{y} p0 ,$$

$$H1 : p < p0 .$$

Entscheidungsregeln

Fall A

Für $n < 25$ verwenden wir die **Binomialtabelle** mit *n* und *po* als Parametern.

Daher suchen wir zunächst nach dem **Wert, der** diesem am nächsten kommt $\frac{\ddot{y}}{2}$ inthistable (die wir mit ÿ1 bezeichnen), wobei ÿ das **Signifikanzniveau ist.** Wir bezeichnen den Wert, der ÿ1 entspricht, mit t1.

Als nächstes finden wir den Wert 1ÿÿ1 = ÿ2 in der Tabelle. Wir bezeichnen den Wert, der ÿ2 entspricht, mit t2.

Wir lehnen H0 auf der Ebene ÿ ab, wenn

$$n1 \ddot{y} t1 \text{ oder } n1 > t2 ,$$

Dabei ist n1 die Anzahl der Beobachtungen, die in die Klasse C1 fallen.

Wenn *n* größer als 25 ist, können wir **die Normalverteilung** als Näherung für die **Binomialverteilung verwenden.**

Die Parameter für die Binomialverteilung Sind:

$$\ddot{y} = n \cdot p0 ,$$

$$\ddot{y} = \ddot{y}n \cdot p0 \cdot q0 .$$

Die **Zufallsvariable** *Z ,* die dem folgt

Die Standardnormalverteilung ist dann gleich:

$$Z = \frac{X \ddot{y} \ddot{y}}{\ddot{y}} = \frac{X \ddot{y} n \cdot p0}{\ddot{y}n \cdot p0 \cdot q0} ,$$

wobei *X* eine zufällige Binomialvariable mit ist Parameter *n* und p0.

Die Näherungen für die Werte von t1 und t2 lauten dann:

$$t1 = n \cdot p0 + z\ddot{y}1 \cdot \ddot{y}n \cdot p0 \cdot q0 , t2 = n \cdot p0 + z\ddot{y}2 \cdot \ddot{y}n \cdot p0 \cdot q0 ,$$

wobei zÿ1 und zÿ2 die in der **Normaltabelle** gefundenen Werte sind, die den Niveaus ÿ1 und ÿ2 entsprechen .

Machine Translated by Google

Fall B

Für $n < 25$ nehmen wir t als den Wert in der Binomialtabelle, der 1 ÿ ÿ entspricht, wobei ÿ das **Signifikanzniveau** (oder den nächstgelegenen Wert) ist und n und p0 die zuvor beschriebenen Parameter sind. Wir lehnen H0 auf der Ebene el ÿ ab, wenn

$$n1 > t \quad ,$$

Dabei ist n1 die Anzahl der Beobachtungen, die in die Klasse C1 fallen.
Für n ÿ 25 können wir die Näherung machen:

$$t = n \cdot p0 + z\ddot{y} \sqrt{\ddot{y}n \cdot p0 \cdot q0} \ ,$$

wobei zÿ in der Normaltabelle für 1 ÿ ÿ zu finden ist.

Die Entscheidungsregel ist in diesem Fall dieselbe wie das für $n < 25$.

Fall C

Für n < 25 ist dies der Wert in der Binomialtabelle, der ÿ entspricht, wobei ÿ das Signifikanzniveau (oder der nächstliegende Wert) ist und mit n und p0 dieselben zuvor beschriebenen Parameter gelten.

Wir lehnen H0 auf der Ebene ÿ ab, wenn

$$n1 \ \ddot{y} \ t \quad ,$$

Dabei ist n1 die Anzahl der Beobachtungen, die in die Klasse C1 fallen.
Für n ÿ 25 machen wir die Näherung:

$$t = n \cdot p0 + z\ddot{y} \sqrt{\ddot{y}n \cdot p0 \cdot q0} \ ,$$

wobei zÿ in der Normaltabelle für das Signifikanzniveau ÿ zu finden ist. Dann ist die Entscheidungsregel dieselbe wie zuvor beschrieben.

DOMÄNEN UND EINSCHRÄNKUNGEN

Zwei Grundbedingungen, die bei der Durchführung des Binomialtests beachtet werden müssen, sind:

1. Die n Beobachtungen müssen gegenseitig sein unabhängig;
2. Jede Beobachtung hat eine **Wahrscheinlichkeit** p , in die erste Klasse zu fallen. Auch diese Wahrscheinlichkeit ist für alle Beobachtungen gleich.

BEISPIELE

Eine Maschine gilt als betriebsbereit, wenn maximal 5 % der von ihr produzierten Teile defekt sind. Die **Nullhypothese** mit der Bezeichnung H0 drückt diese Situation aus, während **die Alternativhypothese** mit der Bezeichnung H1 angibt, dass die Maschine ausfällt:

$$H0 : p \ \ddot{y} \ 0{,}05 \quad H1 : p > 0{,}05 \ .$$

Bei der Durchführung des Tests nehmen wir eine **Probe** von 10 Stück und stellen fest, dass es drei fehlerhafte Teile gibt ($n1 = 3$). Da die Hypothesen dem **einseitigen Test** *(Fall* B) entsprechen, wird Entscheidungsregel B verwendet. Wenn wir ein **Signifikanzniveau** von ÿ = 0,05 wählen, ist der **Wert** von t in der **Binomialtabelle** gleich 1 (für $n = 10$ und p0 = 0,05).
Wir lehnen H0 ab , weil $n1 = 3 > t = 1$, und kommen zu dem Schluss, dass die Maschine versagt.
Dann führen wir den Test erneut durch, diesmal jedoch an 100 Stück. Wir stellen fest, dass es 12 defekte Teile gibt. In diesem Fall kann der Wert oft wie folgt angenähert werden:

$$t = n \cdot p0 + z\ddot{y} \sqrt{\ddot{y}n \cdot p0 \cdot q0} \ ,$$

wobei zÿ aus der **Normaltabelle** für 1 ÿ ÿ = 0,95 ermittelt werden kann . Wir haben dann:

$$t = 100 \cdot 0{,}05 + 1{,}64\ddot{y} \sqrt{100 \cdot 0{,}05 \cdot 0{,}95}$$

$$= 8{,}57 \ .$$

Wir lehnen H0 erneut ab, da $n1 = 12 > t = 8{,}57$.

Machine Translated by Google

WEITERLESEN

- **Binomialtabelle**
- **Anpassungstest**
- **Hypothesentest**
- **Einseitiger Test**
- **Parametrischer Test**
- **Zweiseitiger Test**

VERWEISE

Abdi, H.: Binomialverteilung: Binomial- und Vorzeichentests. In: Salkind, NJ (Hrsg.) Enzyklopädie der Messung und Statistik. Salbei, tausend Eichen (2007)

Bio-Statistiken

Biostatistik ist das wissenschaftliche Gebiet, in dem Dazu werden statistische Methoden angewendet Beantworten Sie Fragen zur Humanbiologie und Medizin (das Präfix „bio“ kommt von Griechisch „bios“, was „Leben“ bedeutet. Die Domänen der Biostatistik sind hauptsächlich Epidemiologie, klinische und biologische Studien, und es wird auch bei der Untersuchung der verwendet Ethik dieser Prozesse.

GESCHICHTE

Die Biostatistik begann Mitte des 17. Jahrhunderts, als Petty, Sir William (1623–1687) und Graunt, John (1620–1674) entwickelte neue Methoden zur Analyse der *Londoner Sterbelisten.* Sie haben diese angewendet neue Methoden zur Analyse der Sterblichkeitsrate, Geburtenraten- und **Volkszählungsstudien ,** die Erstellung der Bereich der Biometrie. Dann, mittendrin das neunzehnte Jahrhundert, die Werke von Mendel, Gregor untersuchte die Vererbung bei Pflanzen. Sein Beobachtungen und Ergebnisse basierten auf systematisch erhobenen Daten, aber auch auf der Anwendung numerischer Methoden zur Beschreibung der Regelmäßigkeit der erblichen Übertragung.

Galton, Francis und **Pearson, Karl** waren zwei der wichtigsten Personen, die mit der Entwicklung dieser Wissenschaft verbunden waren. Sie nutzten die neuen Konzepte und Statistiken Methoden bei der Untersuchung der Ähnlichkeit in physischer, psychischer und verhaltensbezogener Hinsicht Daten zwischen Eltern und ihren Kindern. Pearson, Galton und Weldon, Walter Frank Raffael (1860–1906) war Mitbegründer der Zeitschrift *Biometrika.* **Fisher, Ronald Aylmer,** Während seines Agrarstudiums absolvierte er bei Rothamsted Experimental Station, vorgeschlagen eine Methode der Zufallsstichprobe , bei der Tiere in verschiedene Gruppen eingeteilt wurden und verschiedene Behandlungen zugeordnet, Kennzeichnung die ersten Studien in klinische Studien.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Domänen der Biostatistik sind hauptsächlich **Epidemiologie,** klinische Studien und Ethik Fragen dazu sowie biologische Versuche. Einer der Bereiche, in denen Biostatisti

Bei indirekten Messungen wurden kalische Konzepte zur Analyse einer spezifischen Fragestellung verwendet der Persistenz einer Substanz (z. B. Vitamine und Hormone), wenn sie einem Lebewesen verabreicht wird. Biostatistik bezieht sich daher hauptsächlich auf Statistiken, die zur Lösung verwendet werden Probleme, die in der biomedizinischen Wissenschaft auftreten zen.

Die am häufigsten verwendeten statistischen Methoden in der Biostatistik umfassen **Resampling-Methoden (Bootstrap,** Jackknife) und Multivariate Analyse, verschiedene Regressionsmethoden, Bewertungen der Unsicherheit im Zusammenhang mit Schätzungen, und die Behandlung fehlender Daten.

Die bekanntesten wissenschaftlichen Zeitschriften im Zusammenhang mit der Biostatistik sind: *Statistiken in Medizin, The Biometrical Journal, Controlled Clinical Trials, The Journal of Bio Pharma Statistics, Statistische Methoden in der medizinischen Forschung und Biometrie.*

Machine Translated by Google

WEITERLESEN

- **Demographie**
- **Epidemiologie**

VERWEISE

Armitage, P., Colton, T.: Encyclopedia of Biostatistics. Wiley, New York (1998)

Graunt, J.: Natürliche und politische Beobachtungen, die in einem folgenden Index erwähnt und auf den Sterbeurkunden gemacht werden: in Bezug auf die Regierung, Religion, Handel, Wachstum, Ayre, Krankheiten und die verschiedenen Veränderungen der besagten Stadt. Tho. Roycroft, für John Martin, James Allestry und Tho. Dicas (1662)

Greenwood, M.: Medizinische Statistik von Graunt bis Farr. Cambridge University Press, Cambridge (1948)

Petty, W.: Ein weiterer Aufsatz in politischer Arithmetik über das Wachstum der Stadt London. Kelley, New York (1682) (2. Aufl. 1963)

Petty, W.: Beobachtungen zu den Dublin Bills of Mortality, 1681, und dem Staat dieser Stadt. Kelley, New York (1683) (2. Aufl. 1963)

Block

Blöcke sind Sätze, in denen **experimentelle Einheiten** so gruppiert sind, dass die Einheiten innerhalb jedes Blocks möglichst ähnlich sind. Wir können davon ausgehen, dass der mit einem Block verbundene experimentelle **Fehler** kleiner ist als das man erhält, wenn man die gleiche Anzahl an Einheiten hat waren zufällig im gesamten Experimentierraum verteilt.

Die Blockaden werden im Allgemeinen unter Berücksichtigung sowohl kontrollierbarer Ursachen, die mit den untersuchten Faktoren in Zusammenhang stehen, als auch deren Ursachen bestimmt

Es kann schwierig oder unmöglich sein, über alle **Versuchseinheiten** hinweg konstant zu bleiben . Die Variationen zwischen den Blöcken sind dann eliminiert, wenn wir die Auswirkungen der **Faktoren vergleichen.**

Um die Auswirkungen einer oder mehrerer Quellen zu verringern, können verschiedene Arten von Umgruppierungen eingesetzt werden **Fehler.** Ein **randomisierter Blockentwurf** ergibt sich, wenn nur eine Fehlerquelle vorhanden ist.

GESCHICHTE

Das im Bereich **der Versuchsplanung** verwendete Blockkonzept geht auf Studien von **Fisher, RA** (1925) zurück, als er Leiter der Versuchsstation in Rothamsted war. Bei der Arbeit mit Agrarforschern stellte Fisher fest, dass der für das **Experiment** gewählte Boden (das Feld) offensichtlich heterogen war in dem Sinne, dass die Fruchtbarkeit von einem Punkt auf dem Boden zum anderen systematisch variiert.

WEITERLESEN

- **Versuchsplanung**
- **Experimentelle Einheit**
- **Faktor**
- **Griechisch-lateinisches quadratisches Design**
- **Lateinische quadratische Designs**
- **Randomisiertes Blockdesign**

VERWEISE

Fisher, RA: Statistische Methoden für Forscher. Oliver & Boyd, Edinburgh (1925)

Bonferroni, Carlo E.

Bonferroni, Carlo Emilio wurde 1892 in Bergamo, Italien, geboren. Er erwarb einen Abschluss in Mathematik in Turin und vervollständigte seine Ausbildung mit einem einjährigen Universitätsaufenthalt in Turin

Machine Translated by Google

Wien und an der Eidgenössischen Technischen HochschuleinZürich.Er war Militäroffizier während des Ersten Weltkriegs und danach Nach Kriegsende wurde er Assistenzprofessor am Polytechnikum Turin. 1923 erhielt er der Lehrstuhl für Finanzmathematik am Eco Institut für Wirtschaftswissenschaften in Bari, wo er tätig war Sieben Jahre lang Rektor, dann wechselte er schließlich 1933 ging er nach Florenz, wo er seinen Lehrstuhl innehatte bis zu seinem Tod.

Bonferroni befasste sich mit verschiedenen Themen, darunter versicherungsmathematische Mathematik, Wahrscheinlichkeitsrechnung und Statistische Mathematik, Analysis, Geometrie und Mechanik. Er gab dem seinen Namen zwei Bonferroni-Gleichheiten, die das erleichtern Behandlung statistischer Abhängigkeiten. Diese erschien erstmals 1936 im Artikel *Teoria statistica delle classi e calcolo*

Die Wahrscheinlichkeit. Die Entwicklung dieser Ungleichheiten führten zu einer Vielzahl neuer Dinge Literatur in diesem Bereich.

Bonferroni starb 1960 in Florenz, Italien.

Hauptartikel von Carlo Emilio Bonferroni:

1936 Theoria statistica delle classi e calcolo

Die Wahrscheinlichkeit. Publ. R. Istit. Super. Wissenschaft. Wirtschaft. Kommerziell. Florenz, 8, 1–62.

Bootstrap

Der Begriff Bootstrap beschreibt eine Familie von Techniken, die hauptsächlich zur Schätzung des Standardfehlers, des **Bias** und **des Konfidenzintervalls** eines Parameters (oder mehr) verwendet werden mehr als ein Parameter). Es basiert auf *n* unabhängigen Beobachtungen einer **Zufallsvariablen** mit unbekannter **Verteilungsfunktion** F, und ist besonders nützlich, wenn es sich bei den zu schätzenden Parametern um eine komplizierte Situation handelt Funktion von F.

Die Grundidee von Bootstrap (die der Idee hinter dem Klappmesser in gewisser Weise ähnelt). Methode) besteht darin, *F* mithilfe eines möglichen zu schätzen Verteilung Fˆ und dann eine Neuabtastung von Fˆ. Bootstrap-Verfahren erfordern in der Regel die Verwendung von Computern, da sie eine große Leistung erbringen können Anzahl der Simulationen in relativ kurzer Zeit Zeit. Bei Bootstrap-Methoden ersetzen automatische Simulationen die analytischen Berechnungen, die in den „traditionellen“ Methoden von verwendet werden Schätzung, und in bestimmten Fällen können sie mehr Freiheit bieten, beispielsweise wenn wir dies tun Wir wollen (oder können) keine Hypothese für die Struktur der Verteilung akzeptieren Daten. Bestimmte Bootstrap-Methoden sind enthalten in Statistiksoftware wie S-PLUS, SAS und MATLAB.

GESCHICHTE

Der Ursprung der Verwendung des Wortes „Bootstrap“ in Bezug auf die hier beschriebenen Methoden veranschaulicht anschaulich die reflexive Natur des von ihnen erzeugte Sekundärproben (die diejenigen, die aus Fˆ konstruiert wurden). Es stammt aus die literarische Figur Baron Münchhausen (aus „ Die *überraschenden Abenteuer des Barons“. Münchhausen* von Raspe, RE), der hineinfiel alake, zog sich aber an seinen eigenen Stiefelriemen (seinen Schnürsenkeln) heraus.

Die Geschichte von Bootstrap beginnt, wie viele selbststatistische Techniken, im Jahr 1979 die Veröffentlichung eines Artikels von Efron (1979), das auf viele einen großen Einfluss hatte Forscher. Dieser Artikel löste die Veröffentlichung einer großen Anzahl von Artikeln zum Thema aus Theorie und Anwendungen von Bootstrap.

MATHEMATISCHE ASPEKTE

Es wurden verschiedene Arten von Bootstrap vorgeschlagen; Das Ziel besteht hier nicht darin, alles zu überprüfen diese ausführlich, aber um eine Vorstellung davon zu geben der Art der verwendeten Methoden.

Machine Translated by Google

Seien x1, *x2,...,xn n* unabhängige Beobachtungen einer Zufallsvariablen mit einer Unbekannten Verteilungsfunktion F.

Uns interessiert die Schätzung eines unbekannter Parameter ÿ, der von *F* und der Zuverlässigkeit (also seinem Bias, seiner **Varianz** und seinem Konfidenzintervall) abhängt. Die Schätzung von ÿ erfolgt durch die **Statistik** *t = t(F),* die von F abhängt. Da *F* unbekannt ist, finden wir eine Schätzung Fˆ für *F* basierend auf der Stichprobe x1, *x2,...,xn* und we Schätzen Sie ÿ mit *t* Fˆ .

Klassische Beispiele für *t* sind der **mathematische Erwartungswert,** die Varianz und die Quantile von F. Wir bezeichnen die Zufallsvariable, die der Statistik *t* entspricht (die von x1, *x2,...,xn abhängt),* mit T.

Wir unterscheiden folgende Arten von Bootstraps: •

Parametrischer Bootstrap. Wenn *F* = Fÿ ein Mitglied der parametrischen Familie von Verteilungsfunktionen ist, können wir (den Vektor) ÿ mithilfe des Schätzers ÿˆ der **maximalen Wahrscheinlichkeit schätzen.** Wir schätzen *F* natürlich anhand von Fÿˆ. • *Nichtparametrischer Bootstrap.* Wenn *F* kein Mitglied der Parameterfamilie ist, wird es über eine empirische Verteilungsfunktion geschätzt, die auf der Grundlage einer Größenstichprobe berechnet wird. Die empirische Verteilungsfunktion in *x* ist definiert durch:

$$F\hat{}(x) = \frac{\text{Anzahl der Beobachtungen } xi \ddot{y} x}{N} .$$

Die Idee des Bootstrap ist die folgende.

1. Betrachten Sie *R* unabhängige Stichproben vom $x\ddot{y}_{1,}$ Typ xÿ *2,...,xÿ n* einer Zufallsvariablen der Verteilung Fˆ.
2. Berechnen Sie die Schätzwerte *t* tÿ *1, 2,...,* $T^{\ddot{y}}_{R}$ für den Parameter ÿ basierend auf dem Gen erstellte Bootstrap-Beispiele. Berechnen Sie zunächst die Schätzung Fˆ ÿ basierend auf der Stichprobe *i* *(i = 1,...,R)* und setze dann:

$$t^{\ddot{y}}_{i} = t F\hat{} \ddot{y} _ .$$

Nach der Bootstrap-Methode sind der Bias *b* und die Varianz *v* von *t* geschätzt gepaart mit:

$$\text{bboot} = \bar{t}^{\ddot{y}} \ddot{y} t = \frac{1}{R} \sum_{i=1}^{R} t^{\ddot{y}}_{i} \ddot{y} t ,$$

$$\text{vboot} = \frac{1}{R \ddot{y} 1} \sum_{i=1}^{R} (T^{\ddot{y}}_{_} \ddot{y} \bar{t} \ddot{y})^{2} .$$

Wir sollten anmerken, dass hier zwei Arten von Fehlern auftreten:

- Ein statistischer Fehler, der sich aus der Tatsache ergibt, dass wir den Bias und die Varianz für Fˆ berücksichtigen und nicht das tatsächliche F, das unbekannt ist.
- Ein Simulationsfehler, der aus der Tatsache resultiert, dass die Anzahl der Simulationen nicht stimmt hoch genug. In der Realität reicht es aus, zwischen *R* = 50 und 200 Mal erneut abzutasten .

Es gibt viele mögliche Methoden, Konfidenzintervalle mithilfe von Boot Strap zu ermitteln, und wir stellen hier nur einige vor: 1. *Normale Intervalle.* Angenommen, *T* folgt der Gaußschen (Normal-)Verteilung. In der Hoffnung, dass auch Tÿ der Normalverteilung folgt, finden wir dann das folgende Konfidenzintervall für den (bias)korrigierten Wert *tb = t* ÿ *b* von t:

$$tb \ddot{y} \ddot{y}\text{vboot} \cdot z1\ddot{y}\ddot{y}/2, \quad tb + \ddot{y}\text{vboot} \cdot z1\ddot{y}\ddot{y}/2 ,$$

wobei zÿ das ÿ-Quantil der Gausschen Verteilung N(0, 1) ist. Für eine Toleranzgrenze von ÿ verwenden wir den Wert aus der **Normaltabelle** $z1\ddot{y}\ddot{y}_{_}$, was ÿ = 5 % ergibt: $z1\ddot{y}0{,}05_{_} = 1{,}96$.

Machine Translated by Google

2. *Bootstrap-Intervalle.* Nehmen wir an, dass T ÿ ÿ hängt nicht von einer unbekannten Variablen ab. In diesem Fall berechnen wir das ÿ-Quantil aÿ von T ÿÿ. Dazu ordnen wir die Werte $\overset{ÿ}{T}_1$ *ÿt,* tÿ $_2$ *ÿt, . . . ,* tÿ RÿT in aufsteigender Reihenfolge:

$$\overset{ÿ}{T}_{(1)} \quad ÿ\ t, \text{ tÿ} \qquad ÿ\ t, \ldots, ÿ\ t\overset{ÿ}{T}_{(R)} \quad ,$$

und dann lesen wir die Quantile von inter ÿ *t* und est aÿ = *t* $\overset{ÿ}{T}_{((R+1)(1ÿÿ))}$ $\overset{ÿ}{((R+1)ÿ)}$ ÿ a1ÿÿ = ÿ t. Und soweit

$$1 ÿ 2ÿ = \Pr(aÿ ÿ\ T ÿ ÿ ÿ\ a1ÿÿ)$$
$$= \Pr\ 2t ÿ\ t \quad \overset{ÿ}{((R+1)(1ÿÿ))}\ ÿ\ ÿ$$
$$ÿ\ 2t ÿ\ t \quad \overset{ÿ}{((R+1)ÿ)}\ ,$$

Das Konfidenzintervall für *t* auf dem Signifikanzniveau ÿ beträgt:

$$2t ÿ\ t \quad \overset{ÿ}{((R+1)(1ÿÿ))},\ \overset{ÿ}{((R+1)ÿ)}\ 2t ÿ\ t \quad .$$

3. *Studentisierte Bootstrap-Intervalle.* Es besteht die Möglichkeit, die bisherige Schätzung zu verbessern unter Berücksichtigung von

$$Z = \frac{T ÿ\ ÿ}{ÿv},$$

wobei *v* die Varianz *oft ist,* die sein muss geschätzt (normalerweise unter Verwendung bekannter Methoden wie Delta-Methoden oder dem Klappmesser). Methode). Das Konfidenzintervall für *t* wird auf analoge Weise gefunden vorher beschrieben:

$$t ÿ\ ÿv \cdot \bar{z} \quad \overset{ÿ}{((R+1)(1ÿÿ))},$$
$$t + ÿv \cdot \bar{z} \quad \overset{ÿ}{((R+1)ÿ)}\ ,$$

wobei zÿ $_{((R+1)(1ÿÿ))}$ und zÿ $_{((R+1)ÿ)}$ Sind „empirisch“ die (1 ÿ ÿ und ÿ)-Quan-Kacheln von zÿ $= \frac{\overset{ÿ}{T} ÿt}{vÿ}$ *i=1,...,R.*

BEISPIELE

Wir betrachten ein typisches Beispiel für nichtparametrisches Bootstrap (aus der Arbeit von Davison, A. und Hinkley, DV (1997)). Der Die Daten beziehen sich auf die Bevölkerung von zehn amerikanischen Städten in den Jahren 1920 *(U)* und 1930 (X):

u	138	93	61	179	48
x	143	104	69	260	75

u	37	29	23	30	2
x	63	50	48	111	50

Uns interessiert der Wert ÿ der Statistik $T = \frac{EX)}{EU)}$ was es uns ermöglicht, das abzuleiten Populationen im Jahr 1930 im Vergleich zu denen im Jahr 1920. Wir Schätzen Sie ÿ mit $t = = 1{,}52$, aber was ist die Unsicherheit in dieser Schätzung? Mit dem Bootstrap können wir den abgetasteten ues t $\overset{ÿ}{_}$ Wert *(i = 1,...,R)* von ÿ simulieren eine bivariate Verteilung $Y = (U,$ X) mit an empirische Verteilungsfunktion Fˆ, die ergibt ein Gewicht $\frac{1}{10}$ zu jeder Beobachtung. In der Praxis ist dies eine Frage des Resamplings Geben Sie die zehn Datenwerte zurück und berechnen Sie die dazugehörigen

$$\overset{ÿ}{T}_{_} = \frac{\sum_{j=1}^{10} ij\ xÿ}{\sum_{j=1}^{10} ij\ uÿ}, \qquad i = 1,\ldots,R,$$

wobei xÿ und uÿ sind die Werte der Vari *ij ij* ablesX undU. Dies ermöglicht uns eine Annäherung der Bias und die Varianz von *t* unter Verwendung der zuvor angegebenen Formeln. Die folgende Tabelle fasst die mit $R = 5$ erhaltenen Simulationen zusammen :

	J						
	1	2	3	4	5	6	7
$uÿ_{1j}$	93	93	93	61	61	37	29
$xÿ_{1j}$	104	104	104	69	69	63	50
uÿ	138	61	48	48	37	37	29
$2j\ xÿ_{2j}$	143	69	75	75	63	63	50

Machine Translated by Google

	J						
	1	2	3	4	5	6	7
$u\ddot{y}_{3j}$	138	93	179	37	30	30	30
$x\ddot{y}_{3j}$	143	104	260	63	111	111	111
$u\ddot{y}_{4j}$	93	61	61	48	37	29	29
$x\ddot{y}_{4j}$	104	69	69	75	63	50	50
$u\ddot{y}_{5j}$	138	138	138	179	48	48	48
x5j ÿ	143	143	143	260	75	75	75

	J			Summe
	8	9	10	
$u\ddot{y}_{1j}$	29	30	2	$u\ddot{y}_{1j} = 528$
$x\ddot{y}_{1j}$	50	111	50	$x\ddot{y}_{1j} = 774$
$u\ddot{y}_{2j}$	30	30	2	$u\ddot{y}_{2j} = 460$
$x\ddot{y}_{2j}$	111	111	50	$x\ddot{y}_{2j} = 810$
$u\ddot{y}_{3j}$	30	2	2	$u\ddot{y}_{3j} = 571$
$x\ddot{y}_{3j}$	111	50	50	$x\ddot{y}_{3j} = 1114$
$u\ddot{y}_{4j}$	23	23	2	$u\ddot{y}_{4j} = 406$
$x\ddot{y}_{4j}$	48	48	50	$x\ddot{y}_{4j} = 626$
$u\ddot{y}_{5j}$	29	23	30	$u\ddot{y}_{5j} = 819$
x5j ÿ	50	48	111	$x\ddot{y}_{5j} = 1123$

Also, die $\ddot{t}$ Sind:

$\ddot{t}_1 = 1{,}466$, $t\ddot{y} = 1{,}761\,2$, $t\ddot{y} = 1{,}951\,3$,

$\ddot{t}_4 = 1{,}542$, $t\ddot{y}_5 = 1{,}371$.

Daraus lässt sich der Bias leicht berechnen
die Varianz von *t* unter Verwendung der angegebenen Formeln
bisher:

$b = 1{,}62$ ÿ $1{,}52 = 0{,}10$ und $v = 0{,}0553$.

Es ist möglich, diese Ergebnisse zur Berechnung der normalen Konfidenzintervalle zu verwenden, vorausgesetzt, dass *T* ÿ ÿ normalverteilt ist. *N (0,1,* 0,0553). Wir erhalten für eine Signifikanz Bei einem Niveau von 5 % ergibt sich folgendes Konfidenzintervall für t: [1,16, 2,08]. Die geringe Anzahl der berücksichtigten Städte und Simulationen bedeuten, dass wir nicht viel haben können Vertrauen in die erhaltenen Werte.

WEITERLESEN

- Monte-Carlo-Methode
- Resampling
- Simulation

VERWEISE

Davison, AC, Hinkley, DV: Bootstrap Methoden und ihre Anwendung. Cambridge University Press, Cambridge (1997)

Efron, B.: Bootstrap-Methoden: eine andere Schau dir das Klappmesser an. Ann. Stat. 7, 1–26 (1979)

Efron, B., Tibshirani, RJ: Eine Einführung zum Bootstrap. Chapman & Hall, Neu York (1993)

Hall, P.: The Bootstrap und Edgeworth Erweiterung. Springer, Berlin Heidelberg New York (1992)

Shao, J., Tu, D.: Das Klappmesser und der Stiefelriemen. Springer, Berlin Heidelberg Neu York (1995)

Boscovich, Roger J.

Boscovich, Roger Joseph wurde in Ragusa geboren (heute Dubrovnik, Kroatien) im Jahr 1711. Er besuchte damals das *Collegium Ragusinum* Er ging nach Rom, um am *Collegio* zu studieren *Romano;* Beide Colleges waren jesuitisch. Er starb 1787 in Mailand.

Im Jahr 1760 entwickelte Boscovich eine geometrische Methode zur Bestimmung einer einfachen Regression Linie L1 (um Fehler bei Sternbeobachtungen zu korrigieren). **Laplace, Pierre Simon de** missbilligte die Tatsache, dass Boscovich verwendete geometrische Terminologie und übersetzte diese Methode in eine algebraische Version 1793.

Machine Translated by Google

Das Hauptwerk von Boscovich, Roger Joseph:

1757 De Litteraria Expeditione per Pontifi ciam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex exemplaria etiam sensorum impres sa. Bononiensi Scientiarium et Ar tium Instituto Atque Academia Commentarii, Tomus IV, S. 353–396.

VERWEISE

Whyte, Lancelot Law: Roger Joseph Boscovich, Studien zu seinem Leben und Werk zum 250. Jahrestag seiner Geburt. Allen und Unwin, London (1961)

Box, EP George

Box, George EP wurde 1919 in England geboren. Er diente als Chemiker in den Briten Armeeingenieure im Zweiten Weltkrieg. Nach Nach dem Krieg erlangte er einen Abschluss in Mathematik und Statistiken vom University College London. In den 1950er Jahren war er als Gastprofessor am Institute of Statistics in den USA tätig an der University of North Carolina. Im Jahr 1960 Er zog nach Wisconsin, wo er als diente der erste Vorsitzende der Statistikabteilung. Er erhielt die British Empire Medal im Jahr 1946 und die Shewhart-Medaille im Jahr 1968. Sein Hauptinteresse galt der experimentellen Statistik und der Versuchsplanung. Sein Buch *Statistik für Experimentatoren,* Mitautor mit Hunter, J. Stuart und Hunter, William G. ist einer der am meisten empfohlenen Texte auf diesem Gebiet. Box hat auch pünktlich geschrieben Serienanalyse.

Empfohlene Veröffentlichung von Box, George:

2005 (mit Hunter, William G. und Hunter, J. Stuart) Statistiken für Exper Themen: Design, Innovation und Discovery, 2. Aufl. Wiley, New York

WEITERLESEN

Versuchsplanung

Box-Plot

Das Boxplot ist eine Möglichkeit, die folgenden fünf Größen für einen Datensatz darzustellen: den **Median;** das erste **Quartil** und das dritte Quartil; die Maximal- und Minimalwerte .

Der Boxplot ist ein Diagramm (eine Box), das veranschaulicht Preise:

- Das **Maß der zentralen Tendenz** (in Prinzipal der **Median);**
- Die Variabilität und;
- Die Symmetrie.

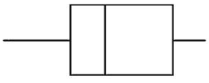
Es wird oft verwendet, um mehrere Sätze zu vergleichen **Beobachtungen.**

GESCHICHTE

Der „Box-and-Whisker-Plot“ oder Boxplot war 1972 von Tukey eingeführt, zusammen mit anderen Methoden zur semigraphischen Darstellung von Daten, von denen eine der bekanntesten die ist **Stamm- und Blattdiagramm.**

MATHEMATISCHE ASPEKTE

Es gibt verschiedene Möglichkeiten, einen Boxplot darzustellen. Wir werden es wie folgt präsentieren:



Das zentrale Rechteck repräsentiert 50 % des Beobachtungen. Es wird als **Interquar-** Kachelbereich bezeichnet. Die untere Grenze dieses Rechtecks ist auf das erste **Quartil** und die Obergrenze festgelegt

Machine Translated by Google

im dritten Quartil. Die Position des Medi **ans** wird durch eine Linie durch das Rechteck angezeigt.

Ein Liniensegment (ein „Whisker“) verbindet beide **Quartil** zum entsprechenden Extremwert (Minimum oder Maximum, es sei denn, es handelt sich um Ausreißer). gegenwärtig; siehe unten) Wert auf jeder Seite von das Rechteck.
In dieser Darstellung werden **Ausreißer** behandelt auf besondere Weise. Wenn die Beobachtungen sind Sehr ausgebreitet definieren wir zwei Werte namens die „internen Grenzen“ oder „Whisker“ durch:

int.lim.1 = 1. **Quartil**

ÿ *(1,5* · **Interquartilbereich)**

int.lim.2 = 3. **Quartil**

\+ *(1,5* · **Interquartilbereich)**

Für jedes interne Limit wählen wir dann das aus Datenwert, der dem Grenzwert am nächsten kommt, aber immer noch innerhalb des Intervalls zwischen der inneren Grenze es ist. Diese beiden Datenwerte werden als benachbarte Punkte bezeichnet. Nun, wenn wir die Box konstruieren Diagramm, die Liniensegmente verbinden die **Quartile** zu den angrenzenden Punkten. Beobachtungen draußen des Intervalls zwischen den internen Grenzen (Ausreißer) werden im Diagramm durch Sterne dargestellt.

BEISPIEL

Das folgende Beispiel stellt die Einnahmen dar Indizes pro Einwohner für jeden Schweizer Kanton (Schweizer Umsatz = 100) im Jahr 1993:

Kanton	Index Kanton	Index
Zürich	125,7 Schaffhouse	99,2
Bern	86.2 Appenzell Rh.-Ext.	84,2
Luzern	87,9 Appenzell Rh.-Int.	72,6
Uri	88,2 St. Gallen	89,3
Schwyz	94,5 Graubünden	92,4
Obwald	80,3 Argovie	98,0
Nidwald	108,9 Thurgau	87,4
Glarus	101,4 Tessin	87,4
Zug	170,2 Waadt	97,4
Freiburg	90,9 Wallis	80,5
Soleure	88,3 Neuenburg	87,3
Basel-Stadt	124.2 Genf	116,0
Ballen Kampagne	105,1 Jura	75.1

Quelle: Bundesamt für Statistik (1993)

Wir berechnen nun die verschiedenen Quartile:
Der Boxplot gibt uns Auskunft über die zentrale Tendenz, über die Streuung der Verteilung und den Grad der Symmetrie:

- Zentrale Tendenz:
  Der **Median** beträgt 90,10.
- Streuung:
  Das **Interquartilintervall** gibt an **Intervall, das** 50 % der Beobachtungen enthält tionen, und diese Beobachtungen sind der Schluss est zum Verteilungszentrum.
  In unserem Beispiel haben wir:
  – 50 % der Kantone der Schweiz haben einen Index, der in das Intervall fällt [102,33ÿ87,03] (mit einer Breite von 15,3).

Die Grenzen des Boxplots sind die folgenden:

lim.int.1 = 1. **Quartil**

ÿ *(1,5* · **Interquartilintervall)**

= *87,03* ÿ *1,5* · *15,3* = *64,08* .

lim.int.2 = 3. **Quartil**

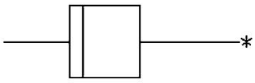
\+ *(1,5* · **Interquartilintervall)**

= *102,33* + *1,5* · *15,3* = *125,28* .

Die Werte größer als 125,28 sind daher Ausreißer. Es gibt zwei davon: 125,7

Machine Translated by Google

(Zürich) und 170,2 (Zoug). Der Boxplot ist dann wie folgt:



WEITERLESEN

Grafische Darstellung

Explorative Datenanalyse

Standortmaß

Maß für die zentrale Tendenz

B

VERWEISE

Tukey, JW: Einige grafische und semigrafische Darstellungen. In: Bancroft, TA (Hrsg.) Sta tistical Papers in Honor of George W. Snedecor. Iowa State University Press, Ames, IA, S. 293–316 (1972)

Machine Translated by Google

Kategoriale Daten

Kategoriale Daten bestehen aus der Anzahl der Beobachtungen, die in bestimmte Klassen fallen. Wir können zwischen verschiedenen Arten von kategorialen Daten unterscheiden: • Binär, charakterisierend das Vorhandensein oder Fehlen einer Immobilie;
• Ungeordnetes Multikategorial (auch „nominal“ genannt); • Geordnete multikategoriale (auch „ordinal“ genannt); • Ganze Zahlen.

Wir stellen die kategorialen Daten in Form einer Kontingenztabelle dar.

DOMÄNEN UND EINSCHRÄNKUNGEN

Im Wesentlichen kontinuierliche Variablen können auch als kategoriale Variablen dargestellt werden. Ein Beispiel ist „Alter“, eine kontinuierliche Variable, das Alter kann jedoch weiterhin in Klassen gruppiert werden, sodass es weiterhin als kategoriale Daten dargestellt werden kann.

BEISPIELE

Bei einer Meinungsumfrage zur Zustimmung oder Ablehnung eines neuen Gesetzes können die abgegebenen Stimmen entweder „Ja“ oder „Nein“ lauten. Die Ergebnisse können wir in Form einer Kontingenztabelle darstellen:

	ja	Nein
Stimmen	8546	5455

Unterteilt man die Mitarbeiter eines Unternehmens nach Berufen (und es werden mindestens drei Berufe dargestellt), erhält man als Daten ungeordnete multikategoriale Daten (es gibt keine natürliche Reihenfolge der Berufe).
Wenn wir uns dagegen für die Anzahl der Personen interessieren, die unterschiedliche Bildungsniveaus erreicht haben, wird es wahrscheinlich eine natürliche Reihenfolge der Kategorien geben: „Grundschule, Sekundarstufe“ und dann Universität. Solche Daten wären daher ein Beispiel für geordnete multikategoriale Daten.
Wenn wir Mitarbeiter schließlich auf der Grundlage der Größe der Familie jedes Mitarbeiters (d. h. der Anzahl der Familienmitglieder) in Kategorien einteilen, erhalten wir kategoriale Daten, bei denen es sich bei den Kategorien um ganze Zahlen handelt.

WEITERLESEN

- **Analyse kategorialer Daten**
- **Binärdaten**
- **Kategorie**
- **Daten**
- **Dichotome Variable**
- **Qualitative kategoriale Variable**
- **Zufällige Variable**

VERWEISE

Siehe **Analyse kategorialer Daten.**

Machine Translated by Google

Kategorie

Eine Kategorie stellt eine Menge von Personen oder Objekten dar, die ein gemeinsames Merkmal haben.

Wenn wir die Menschen in einer **Bevölkerung** untersuchen wollen , können wir sie in „natürliche“ Kategorien einordnen, zum Beispiel nach Geschlecht (Männer und Frauen), oder in Kategorien, die durch andere Kriterien definiert sind, wie z. B. Beruf (Manager, Sekretärinnen, Bauern . . .).

WEITERLESEN

- **Binärdaten**
- **Kategoriale Daten**
- **Dichotome Variable**
- **Bevölkerung**
- **Zufällige Variable**
- **Variable**

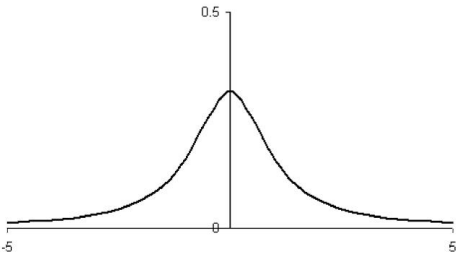
Cauchy-Verteilung

Eine **Zufallsvariable** *X* folgt einer Cauchy-Verteilung, wenn ihre **Dichtefunktion** die Form hat:

$$f(x) = \frac{1}{\cdot 1 + \frac{x\ddot{y}\ddot{y}}{\ddot{y}}^{2}}, \quad \ddot{y}1 \qquad \ddot{y} > 0 .$$

Die **Parameter ÿ** und ÿ sind die Orts- bzw. Ausbreitungsparameter.
Die Cauchy-Verteilung ist symmetrisch um *x* = ÿ, das den Median darstellt . Das erste **Quartil** und das dritte Quartil sind durch ÿ ± ÿ gegeben.

Die Cauchy-Verteilung ist eine **kontinuierliche Wahrscheinlichkeitsverteilung.**



Cauchy-Verteilung, ÿ = 1, ÿ = 0

MATHEMATISCHE ASPEKTE

Der **Erwartungswert** *E[X]* und die **Varianz** Var(X) existieren nicht.

Wenn ÿ = 0 und ÿ = 1, ist die Cauchy-Verteilung identisch mit der **Student-Verteilung** mit ein **Freiheitsgrad.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Seine Bedeutung in der Physik beruht hauptsächlich auf der Tatsache, dass es die Lösung des Differentials ist Gleichung, die Kraftresonanz beschreibt.

WEITERLESEN

- **Kontinuierliche Wahrscheinlichkeitsverteilung**
- **Studentenverteilung**

VERWEISE

Cauchy, AL: Über die Ergebnisse meiner Beobachtungen selbst in der Natur und über die Ergebnisse mit größerer Wahrscheinlichkeit. CR Acad. Wissenschaft. **37,** 198–206 (1853)

Kausale Epidemiologie

Das Ziel der kausalen Epidemiologie besteht darin, herauszufinden, wie Ursache und Wirkung im Hinblick auf die menschliche Gesundheit zusammenhängen.

Mit anderen Worten handelt es sich um die Untersuchung von Krankheitsursachen und den Versuch, statistische Daten zu finden.

Machine Translated by Google

Technischer Beweis für einen Kausalzusammenhang oder einen Zusammenhang zwischen der Krankheit und dem Faktor vorgeschlagen, die Krankheit zu verursachen.

GESCHICHTE

Siehe **Epidemiologie.**

MATHEMATISCHE ASPEKTE

Siehe **Ursache und Wirkung in der Epidemiologie, Chancen und Chancenverhältnis, relatives Risiko, zuschreibbares Risiko, vermeidbares Risiko, Inzidenzrate, Prävalenzrate.**

BEREICHE UND GRENZEN Studien

zum Zusammenhang zwischen Tabakrauchen und der Entstehung von Lungenkrebs sowie zum Zusammenhang zwischen HIV und AIDS sind Beispiele für die kausale Epidemiologie. Die Erforschung eines Kausalzusammenhangs ist oft sehr komplex und erfordert viele Studien sowie die Einbeziehung und Kombination verschiedener Datensätze aus biologischen und Tierversuchen bis hin zu klinischen Studien.

Während Ursachen nicht immer genau identifiziert werden können, ist die Kenntnis der mit einer Krankheit verbundenen Risikofaktoren und damit der

Durch die Erfassung gefährdeter Personengruppen können wir präventiv und gesundheitserhaltend eingreifen.

BEISPIELE

Als Beispiel können wir den Zusammenhang zwischen Rauchen und der Entstehung von Lungenkrebs untersuchen. Betrachten Sie eine Studie mit 2.000 Probanden: 1.000 Rauchern und 1.000 Nichtrauchern. Die Altersverteilung und das Männer/Frauen-Verhältnis sind in beiden Gruppen identisch. Lassen Sie uns eine Zusammenfassung der über viele Jahre gesammelten Daten analysieren. Die Ergebnisse sind in der folgenden Tabelle dargestellt:

	Nichtraucher	Raucher
Anzahl Lungenkrebsfälle (/1000)	50	10
Anteil der Personen in dieser Gruppe, die an einer Erkrankung erkrankt sind Lungenkrebs	5 %	1 %

Wenn wir den Anteil der Raucher, die an Lungenkrebs erkranken, mit dem Anteil der Nichtraucher vergleichen, die daran erkranken, erhalten wir 5 %/1 % = 5. Daraus können wir schließen, dass das Risiko, an Lungenkrebs zu erkranken, bei Rauchern fünfmal höher ist als bei Nichtrauchern. Im Allgemeinen bewerten wir auch die Signifikanz dieses Ergebnisses durch Berechnung ***des p-* Werts.** Nehmen wir an, dass **der Chi-Quadrat-Test** $p < 0{,}001$ ergab . Normalerweise wird angenommen, dass die erhaltenen Ergebnisse statistisch signifikant sind, wenn der **p-Wert** kleiner als 0,05 ist.

Angenommen, wir führen dieselbe Studie durch, aber die Dimension jeder Gruppe (Raucher und Nichtraucher) beträgt 100 statt 1000, und wir beobachten die gleichen Anteile von Menschen, die an Lungenkrebs erkranken:

	Nichtraucher	Raucher
Anzahl der Fälle von Lungenkrebs (/100)	5	1
Anteil der Personen daran Gruppe, die an der Lunge erkrankte Krebs	5 %	1 %

Wenn wir denselben statistischen Test durchführen, beträgt der **p-Wert** 0,212. Da dieser Wert größer als 0,05 ist, können wir keine eindeutige Schlussfolgerung ziehen, dass es sich um eine Existenz handelt

eines signifikanten statistischen Zusammenhangs zwischen Rauchen und Lungenkrebs. Dies verdeutlicht, dass, um eine statistisch signifikante

Machine Translated by Google

Der Unterschied zwischen den Ergebnissen kann nicht groß sein

Für verschiedene Populationen, die aus epidemiologischen Studien gewonnen wurden, ist es normalerweise notwendig, große Stichproben zu untersuchen.

WEITERLESEN

Siehe **Epidemiologie.**

VERWEISE

Siehe **Epidemiologie.**

Ursache und Wirkung in der Epidemiologie

In der Epidemiologie ist die „Ursache“ ein Erreger (mikrobielle Keime, verschmutztes Wasser, Rauchen usw.), der die Gesundheit verändert, und die „Wirkung“ beschreibt die Art und Weise, wie der Erreger die Gesundheit verändert. Der Erreger ist häufig potenziell pathogen (in diesem Fall wird er als „Risikofaktor“ bezeichnet).

Der Effekt ist somit faktisch ein Risikovergleich. Wir können in diesem Zusammenhang zwei verschiedene Arten von Risiken definieren:

- Die absolute Wirkung einer Ursache drückt die Erhöhung des Risikos bzw. des Zusätzlichen aus
  Anzahl der daraus resultierenden Krankheitsfälle bzw
  könnte aus der Exposition gegenüber dieser Ursache resultieren. Es wird am **zurechenbaren Risiko** und seinen Derivaten gemessen.
- Die relative Wirkung einer Ursache drückt die Stärke des Zusammenhangs zwischen dem Erreger und der Krankheit aus.

Eine Ursache, die selbst eine Wirkung hervorruft, wird als *ausreichend bezeichnet.*

GESCHICHTE

Die Begriffe „Ursache“ und „Wirkung“ wurden definiert

bei der Geburt der **Epidemiologie,** die im 17. Jahrhundert stattfand.

MATHEMATISCHE ASPEKTE

Formal haben wir:

$$\text{Absoluter Effekt} = \text{Risiko für Exponierte} \; \ddot{y} \; \text{Risiko für unbelichtete}.$$

Der absolute Effekt drückt das zusätzliche Risiko oder die Krankheitsfälle aus, die sich aus der Exposition gegenüber der Ursache ergeben (oder ergeben könnten).

$$\text{Relativer Effekt} = \text{Risiko} \frac{\text{Risiko für ausgesetzt}}{\text{für Nichtbelichtete}}.$$

Der relative Effekt drückt die Stärke des Zusammenhangs zwischen der Krankheit und dem aus

Ursache. Es wird anhand des **relativen Risikos** und des **Odds Ratio gemessen.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Streng genommen reicht die Stärke eines Zusammenhangs zwischen einem bestimmten Faktor und einer Krankheit nicht aus, um einen kausalen Zusammenhang zwischen ihnen herzustellen. Wir müssen auch Folgendes berücksichtigen:

- Das „Zeitlichkeitskriterium“ (wir müssen sicher sein, dass die vermeintliche Ursache der Wirkung vorausgeht) und;
- Grundlegende und experimentelle Forschungselemente, die es uns ermöglichen, dies sicherzustellen
  Der vermeintliche Kausalfaktor ist eigentlich kein „Verwirrungsfaktor“ (ein Faktor, der nicht kausal ist, sondern statistisch mit dem nicht identifizierten tatsächlichen Kausalfaktor zusammenhängt).

Diesen relativen und absoluten Wirkungen entsprechen zwei Arten von Kausalität. Die relative Kausalität ist unabhängig

von den klinischen oder öffentlichen Auswirkungen der Wirkung; Im Allgemeinen erlaubt es uns nicht, die Auswirkungen der damit verbundenen Wirkung auf die klinische oder öffentliche Gesundheit vorab zu beurteilen. Es kann stark sein, wenn das Risiko einer Exposition sehr hoch ist oder wenn das Risiko, nicht exponiert zu sein, sehr gering ist.

Die absolute Kausalität drückt die klinische oder öffentliche Auswirkung der damit verbundenen Wirkung aus.

Machine Translated by Google

und ermöglicht uns daher die Beantwortung der Frage: Wenn wir die Ursache unterdrückt hätten, was?
Ausmaß der Auswirkungen auf die Bevölkerung (Intervall von Krankheitsfälle) hätten vermieden werden können? Wenn Der Patient hatte mit dem Rauchen aufgehört, was auch immer die Verringerung seines Risikos, an einer Lunge zu erkranken Krebs oder ein Herzinfarkt sein?
Ein Risikofaktor, der mit einem hohen Relativwert verbunden ist Wirkung, die jedoch nur eine kleine Anzahl von Personen betrifft, wird weniger Krankheiten verursachen und Todesfälle sind dann ein damit verbundener Risikofaktor mit einem kleineren relativen Effekt, aber wo viele mehr Personen sind exponiert. Es ist daher klar, dass die Bedeutung eines Kausalzusammenhangs wichtig ist Die Beziehung variiert je nachdem, ob wir berücksichtigen relative oder absolute Kausalität.

Wir sollten hier auch einen wichtigen Punkt ansprechen über kausale Wechselwirkungen. Es kann viele geben Ursachen für die gleiche Krankheit. Während alle diese Wenn Ursachen zum gleichen Ergebnis beitragen, können sie es auch interagieren. Die wichtigste Konsequenz davon Der kausale Zusammenhang besteht darin, dass wir nicht vorab urteilen können die Auswirkung einer gleichzeitigen Exposition gegenüber den Ursachen *A* und *B* (bezeichnet als *A+B)*, basierend auf dem, was wir tun Wissen Sie, welche Auswirkungen es hat, wenn Sie nur Aor ausgesetzt sind nurB. Im Gegensatz zum Fall von Independent Ursachen, wir müssen die gemeinsame Wirkung abschätzen, nicht beschränken wir uns auf die isolierten Analysen der zusammenwirkenden Ursachen.

BEISPIELE

Der relative Effekt und der absolute Effekt sind unterliegen unterschiedlichen Interpretationen, wie das folgende Beispiel zeigt.
Angenommen, wir haben zwei Populationen P1 *und P2,* jeweils 100.000 Personen. In der Bevölkerung P1 beträgt das Risiko, an einer bestimmten Krankheit zu erkranken, 0,2 % für die Betroffenen und 0,1 % für die Betroffenen unbelichtet. In PopulationP2 ist das Risiko für belichtet beträgt 20 % und zwar für unbelichtet beträgt 10 %, wie in der folgenden Tabelle dargestellt:

Bevölkerung	Risiko für die ausgesetzt (%)	Risiko für die unbelichtet (%)
	A	B
P1	0,2	0,1
P2	20	10

Bevölkerung	Relativ Wirkung $\frac{A}{B}$	Absolut Wirkung (%) C = A − B	Vermeidbar Fälle C ×100000
P1	2,0	0,1	100
P2	2,0	10	10000

Der relative Effekt ist für die Populationen P1 und P2 derselbe (das Verhältnis des Risikos für die Populationen P1 und P2). dem Risiko für das Unbelichtete ausgesetzt ist 2),
Die Wirkung derselben Präventionsmaßnahmen wäre jedoch in beiden Bevölkerungsgruppen sehr unterschiedlich, da die absolute Wirkung bei zehn liegt mal wichtiger in P2: die Anzahl der potenziell vermeidbare Fälle beträgt daher 100 in Bevölkerung P1 und 10000 in Bevölkerung P2.
Betrachten Sie nun die **Inzidenzrate** der Lunge Krebs in einer Population von Personen, die 35 oder mehr Zigaretten pro Tag rauchen: 3,15/1000/Jahr. Während diese Rate scheinen mag klein, es maskiert die Tatsache, dass es ein Starkes gibt relativer Effekt (das Risiko ist 45-mal größer). für Raucher dann für Nichtraucher) aufgrund die Tatsache, dass Lungenkrebs bei Nichtrauchern sehr selten ist (die Inzidenzrate für Nichtraucher). beträgt 0,07/1000/Jahr).

WEITERLESEN

- Zurechenbares Risiko
- Vermeidbares Risiko
- Inzidenzrate
- Quoten und Quotenverhältnis
- Prävalenzrate
- Relatives Risiko

Machine Translated by Google

VERWEISE

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Kleiner Brauner, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Rothmann, JK: Epidemiologie. Eine Einleitung. Oxford University Press (2002)

Volkszählung

Eine Volkszählung ist eine Operation, die aus Folgendem besteht:
Beobachtung aller Individuen einer **Population.** Das Wort
„Zensus“ kann sich auf eine Volkszählung beziehen, also
auf eine Bevölkerung
zählen, es kann sich aber auch auf Anfragen beziehen (genannt
„erschöpfende“ Anfragen), die wir abrufen
Informationen über eine Population durch Beobachtung
aller Individuen in der Population.
Es ist klar, dass solche Untersuchungen für sehr große
Bevölkerungsgruppen sehr kostspielig sein werden. Darum
umfassende Untersuchungen werden selten durchgeführt;
Probenahme, die aus der Beobachtung von besteht
nur ein Teil der Bevölkerung (genannt
Stattdessen wird in der Regel **eine Probe** verwendet.

GESCHICHTE

Volkszählungen haben ihren Ursprung in den großen
Zivilisationen der Antike, als große Gebiete erfasst wurden
von Imperien und der damit verbundenen Komplexität
Um sie zu regieren, waren Kenntnisse über die erforderlich
Bevölkerungsgruppen beteiligt.
Es gehört zu den ältesten Zivilisationen
Es ist bekannt, dass Volkszählungen durchgeführt wurden
in Sumer (zwischen 5000 und 2000 v. Chr.),
wo die beteiligten Personen Listen von gemeldet haben
Männer und Waren auf Tontischen in Keilschrift
Figuren.

Volkszählungen wurden auch in Mesopotamien (ca. 3000
v. Chr.) sowie in durchgeführt
altes Ägypten ab der ersten Dynastie; Diese
Volkszählungen wurden fällig durchgeführt
zumilitärischen und fiskalischen Zielen. Unter Ama sis II
musste jeder (unter Lebensgefahr)
Geben Sie ihren Beruf und die Quelle(n) ihres Rev. an
enue.

Die Situation in Israel war komplexer:
Volkszählungen waren manchmal obligatorisch und
manchmal aufgrund des alten Testa verboten
ment. Dieser Einfluss auf die christliche Zivilisation
tion dauerte eine ganze Weile; mitten drin
Alter, St. Augustin und St. Ambroise waren
Ich verurteile immer noch Volkszählungen.

In China wurden Volkszählungen durchgeführt
seit mindestens 200 v. Chr. in verschiedenen Formen
und für verschiedene Zwecke. Hecht, J. (1987)
berichtete über die wichtigsten Volkszählungen:

1. *Han-Dynastie (200 Jahre v. Chr. bis 200 Jahre).*
   AD): Volkszählungen bezogen sich auf
   das System der Wehrpflicht.
2. *Verbindungen zwischen drei Königreichen und fünf*
   Dynastien (221–959 n. Chr.): Bezogen auf das System
   der territorialen Verteilung.
3. *Song- und Yuan-Dynastie (960–1368).*
   AD): Volkszählungen wurden für das Finanzwesen durchgeführt
   Zwecke.
4. *Ming-Dynastie (1368–1644 n. Chr.):* „Gelb
   Register“ wurden für die Dauer von zehn Jahren eingerichtet
   Volkszählungen. Sie listeten Namen, Beruf, Geschlecht
   und Alter jeder Person auf.
5. *Qing-Dynastie (seit 1644 n. Chr.):* Volkszählungen
   wurden durchgeführt, um die Bevölkerungsmigration
   zu erfassen.

Im Japan des Mittelalters war das anders
Es wurden verschiedene Volkszählungsarten verwendet. Die erste
Volkszählung wurde wahrscheinlich unter dieser Herrschaft durchgeführt
von Kaiser Sujin (86 v. Chr.).
Schließlich gibt es in Indien eine politische und wirtschaftliche
Wissenschaftliche Abhandlung mit dem Titel „Arthasastra“ (Profit
Vertrag) gibt einige Hinweise zur Nutzung

Machine Translated by Google

einer äußerst detaillierten Aufzeichnung. Diese Abhandlung wurde von Kautilya, Premierminister unter Chandragupta Maurya (313–289 v. Chr.), verfasst.

Eine andere sehr wichtige Zivilisation, die Inkas, verwendeten ebenfalls Volkszählungen. Sie nutzten ein Statistiksystem namens „quipos“. Jeder Quipo war sowohl ein Instrument als auch ein Informationsregister. Die Farben, Kombinationen und Knoten der Schnüre bestanden aus einer Reihe von Schnüren und hatten eine genaue Bedeutung. Die Quipos wurden an speziell initiierte Wachen weitergegeben, die alle diese Statistiken zusammenstellten.

In Europa praktizierten auch die alten Griechen und Römer Volkszählungen. Aristoteles berichtete, dass die Griechen eine Portion Weizen spendeten pro Geburt und ein Maß Gerste pro Tod an die Göttin Athéna. In Rom wurde die erste Volkszählung auf Geheiß von König Servius Tullius (578–534 v. Chr.) durchgeführt, um die Einnahmen zu überwachen und folglich die Steuern zu erhöhen.

Später wurden Volkszählungen je nach Land mit unterschiedlicher Häufigkeit und in unterschiedlichem Umfang durchgeführt.

Im Jahr 786 ordnete Karl der Große an, alle seine über zwölf Jahre alten Untertanen zu zählen; Auch in Italien wurden im 12. Jahrhundert Bevölkerungszählungen eingeführt; Viele Städte führten im 15. Jahrhundert Volkszählungen ihrer Einwohner durch Jahrhundert, darunter Nürnberg (1449) und Straßburg (1470). Im 16. Jahrhundert führte Frankreich Familienstandsregister ein.

Im 17. Jahrhundert entwickelten sich drei verschiedene Denkschulen: eine deutsche Schule, die sich mit **deskriptiver Statistik beschäftigte,** eine französische Schule, die sich mit der Ideologie und Methodik der Volkszählung beschäftigte, und eine englische Schule, die zur modernen Statistik führte.

In der Geschichte der Volkszählungen in Europa gibt es ein Land, das einen besonderen Platz einnimmt. Im Jahr 1665 führte Schweden Kirchenbücher ein ioner, die von Pfarrern unterhalten wurden; Im Jahr 1668 wurde die Eintragung in diesen Registern durch ein Dekret verpflichtend, und die Register waren nicht mehr religiös, sondern administrativ. 1736 erschien ein weiteres Dekret, das besagte, dass der Gouverneur jeder Provinz dem Parlament alle Veränderungen in der Bevölkerung der Provinz melden musste. Die schwedische Bevölkerungsstatistik wurde am 3. Februar 1748 durch die Schaffung des „Tabellverket“ (Verwaltungstabellen) offiziell anerkannt.

Die erste Zusammenfassung für alle schwedischen Provinzen, die 1749 erstellt wurde, kann als die erste richtige Volkszählung in Europa angesehen werden, und der 11. November 1756 markierte die Gründung der „Tabellkommissionen“, der ersten offiziellen Abteilung für Statistik. Seit 1762 werden diese Zahlentabellen von der Akademie der Wissenschaften gepflegt.

Anfangs wurden die schwedischen Volkszählungen jährlich (1749–1751) und dann alle drei Jahre (1754–1772) durchgeführt, seit 1775 werden sie jedoch alle fünf Jahre durchgeführt.

Ende des 18. Jahrhunderts wurde dort ein offizielles Institut für Volkszählungen gegründet Frankreich (1791). Im Jahr 1787 wurde das Prinzip der Volkszählung in der Verfassungscharta der USA (CC USA) verankert.

Siehe auch **amtliche Statistiken.**

WEITERLESEN

- **Datensammlung**
- **Demographie**
- **Offizielle Statistiken**
- **Bevölkerung**
- **Probe**
- **Umfrage**

VERWEISE

Hecht, J.: L'idée du dénombrement jusqu'à la Révolution. Für eine Geschichte von la

Machine Translated by Google

statistique, Band 1, S. 21–82. Econo
Glimmer/INSEE (1978)

Zentraler Grenzwertsatz

Der zentrale Grenzwertsatz ist von grundlegender Bedeutung
Satz der Statistik. In seiner einfachsten Form ist es
schreibt vor, dass die Summe einer ausreichend großen
Die Anzahl der unabhängigen, identisch verteilten
Zufallsvariablen ergibt sich näherungsweise
eine **Normalverteilung.**

GESCHICHTE

Der zentrale Grenzwertsatz wurde erstmals aufgestellt
im Rahmen von **Binomi** durchgeführt
Alvertrieb durch **Moivre, Abraham de**
(1733). **Laplace, Pierre Simon de** (1810)
formulierte den Beweis des Satzes.
Poisson, Siméon Denis(1824) war ebenfalls tätig
Zu diesem Theorem und **Chebyshev gab Pafnu tii**
Lvovich (1890–1891) eine strenge Stellungnahme
Demonstration davon in der Mitte der Neun
10. Jahrhundert.
Zu Beginn des 20. Jahrhunderts entdeckte der
russische Mathematiker Liapounov,
Aleksandr Mikhailovich (1901) schuf das
allgemein anerkannte Form des zentralen Limes-
Theorems durch Einführung seiner charakteristischen
Funktionen. Markov, **Andrei Andreevich**
(1908) arbeitete ebenfalls daran und war der Erste
Verallgemeinern Sie den Satz auf den Fall
unabhängiger Variablen.
Laut Le Cam, L. (1986) wurde ihm die Bezeichnung
„zentral“ von George Polyà verliehen
(1920) aufgrund der wesentlichen Rolle, die es spielt
in der Wahrscheinlichkeitstheorie.

MATHEMATISCHE ASPEKTE

Seien X1, *X2,...,Xn n* unabhängige Zufallsvariablen
mit identischer Verteilung
ed (mit beliebiger Verteilung) mit einem Mittelwertÿand
eine endliche **Varianz** ÿ2.

Wir definieren die Summe $Sn = X1 + X2 + ...+ Xn$
und wir ermitteln das Verhältnis:

$$\frac{Sn \ddot{y}\ n \cdot \ddot{y}}{\ddot{y} \cdot \ddot{y}\overline{n}},$$

wobei n·ÿ und ÿ · $\ddot{y}\overline{n}$ den Mittelwert darstellen und
die **Standardabweichung** von *Sn* .
Der zentrale Grenzwertsatz legt fest, dass die
Die Verteilung dieses Verhältnisses tendiert zum Standard
Normalverteilung, wenn *n* gegen Unendlich geht.
Das bedeutet, dass:

$$P\left(\frac{Sn \ddot{y}\ n \cdot \ddot{y}}{\ddot{y}\ \ddot{y}\overline{n}} \ddot{y}\ x\right) \ddot{y}\ddot{y}n\ddot{y}+\ddot{y}\ (x)$$

wobei *(x)* die **Verteilungsfunktion** der
Standardnormalverteilung **ist,**
ausgedrückt durch:

$$(x) = \int_{\ddot{y}\ddot{y}}^{x} \frac{1}{\ddot{y}2\ddot{y}} \exp\left(\frac{\ddot{y}x2}{2}\right) dx ,$$

$$\ddot{y}\ \ddot{y} < x < \ddot{y} \quad .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Der zentrale Grenzwertsatz bietet eine einfache Lösung
Methode zur näherungsweisen Berechnung von
Wahrscheinlichkeiten, die sich auf die Summe der Zufallsvariationen beziehen
fähig.
Abgesehen von seinem Interesse an der **Sam**
Pling- Theorem, wobei die Summen und die
Mittelwerte spielen eine wichtige Rolle, der zentrale Lim-
It-Satz wird zur Annäherung an **den Normalwert** verwendet
Verteilungen , die aus der Summierung identischer
Verteilungen abgeleitet werden. Wir können zum Beispiel mit
die Hilfe des zentralen Grenzwertsatzes nutzen
die **Normalverteilung** zur Annäherung
Binomialverteilung, die **Poisson-Verteilung,** die
Gamma-Verteilung, die **Chi-Quadrat-Verteilung,**
die **Student-Verteilung,** die **hypergeometrische**
Verteilung,
die **Fisher-Verteilung** und die **Lognormalverteilung**
Verteilung.

Machine Translated by Google

BEISPIELE

Bei einer großen Menge elektrischer Artikel ist die **Wahrscheinlichkeit ,** einen defekten Artikel auszuwählen, gleich bei $\frac{1}{8}$. Wie hoch ist die Wahrscheinlichkeit, dass 25 defekten Artikeln 4 defekte *Artikel* ausgewählt werden? ausgewählt?

Sei X die **dichotome** Variable

Ergebnis eines Versuchs:

$$X = \begin{cases} 1, \text{ wenn der ausgewählte Artikel defekt ist;} \\ 0, \text{ wenn nicht.} \end{cases}$$

Die **Zufallsvariable** X folgt **aBernoul li-Verteilung** mit **Parameter** p. Folglich folgt die Summe Sn = *X1+X2+...+Xn* einer **Binomialverteilung** mit einem **Mittelwert,** *np* und eine **Varianz** np(1ўp), die nach dem zentralen Grenzwerttheorem durch **eine Normalverteilung** mit einem Mittelwert angenähert werden kann ў = *np* und eine Varianz ў2 = np(1 ў p).
Wir bewerten diese Werte:

$$ў = n \cdot p = 25 \cdot \frac{1}{8} = 3{,}125$$

$$ў2 = n \cdot p(1 ў p) = 25 \cdot \frac{1}{8} \cdot 1 ў \frac{1}{8}$$

$$= 2{,}734 .$$

Wir berechnen dann P(Sn > 4) in zwei verschiedenen Wege:

1. Mit der Binomialverteilung:

   Aus der **Binomialtabelle** die Wahrscheinlichkeit von $P(Sn ў 4) = 0{,}8047$. Die Wahrscheinlichkeit von $P(Sn > 4)$ ist dann:

$$P(Sn > 4) = 1 ў P(Sn ў 4) = 0{,}1953 ,$$

2. Mit der Normalnäherung (erhalten aus dem zentralen Grenzwertsatz):
   Um das diskrete Zeichen zu berücksichtigen Akteur der Zufallsvariablen *Sn* müssen wir eine Kontinuitätskorrektur vornehmen; das heißt, wir Berechnen Sie die Wahrscheinlichkeit, dass *Sn* größer ist dann $4 + \frac{1}{2} = 4{,}5$

Wir haben:

$$z = \frac{Sn ў n \cdot p}{ўn \cdot p (1 ў p)} = \frac{4{,}5 ў 3{,}125}{1.654}$$

$$= 0{,}832 .$$

Aus der **normalen Tabelle** erhalten wir die Wahrscheinlichkeit:

$$P(Z > z) = P(Z > 0{,}832)$$

$$= 1 ў P(Z ў 0{,}832)$$

$$= 1 ў 0{,}7967 = 0{,}2033 .$$

WEITERLESEN

Binomialverteilung

Chi-Quadrat-Verteilung

Konvergenz

Konvergenzsatz

Fisher-Verteilung

Gammaverteilung

Hypergeometrische Verteilung

Gesetz der großen Zahlen

Lognormalverteilung

Normalverteilung

Poisson-Verteilung

Wahrscheinlichkeit

Wahrscheinlichkeitsverteilung

Studentenverteilung

REFERENZ

Laplace, PS de: Erinnerung an Approximationen von Formeln, die ihre Funktion haben Es gibt sehr viele und Ihre Anwendung mit Wahrscheinlichkeiten. Erinnerungen de l'Académie Royale des Sciences de Paris, 10. Abgebildet in: Œuvres de Laplace **12,** 301–347 (1810)

Le Cam, L.: Der zentrale Grenzwertsatz um 1935. Stat. Wissenschaft. 1, 78–96 (1986)

Machine Translated by Google

Liapounov, AM: Über einen Vorschlag zur Wahrscheinlichkeitstheorie. Bulletin de l'Academie Imperiale des Sciences de St.-Petersbourg 8, 1–24 (1900)

Markov, AA:Erweiterung der theoretischen Berechnung von Wahrscheinlichkeitsrechnungen für Kettengrößen. Mem. Acad. Wissenschaft. St. Petersburg 8, 365–397 (1908)

Moivre, A. de: Approximatio ad summam terminorum binomii *(a + b)n,* in seriem expansi. Supplementum II zu Miscellanae Analytica, S. 1–7 (1733). Fotografisch nachgedruckt in einer seltenen Broschüre über Moivre und einige seiner Entdeckungen. Kneipe

gelistet von Archibald, RC Isis 8, 671–683 (1926)

Poisson, SD: Surla probabilité des résultats moyens des beobachtungen. Connaissance des temps pour l'an 1827, S. 273–302 (1824)

Polyà, G.: Über den Grenzwert zentralen Satz der Wahrscheinlichkeitsrechnung und das Momentproblem. Mathematische Zeitschrift 8, 171–181 (1920)

Tschebytschew, PL (1890–1891). Über zwei relative Theorien und Wahrscheinlichkeiten. Acta Math. **14,** 305–315

Tschebyschew, Pafnutii Lwowitsch

Chhebyshev, Pafnutii Lvovich (1821–1894) begann 1837 sein Studium an der Moskauer Universität, wo er von Zernov, Nikolai Efimovich (dem ersten Russen, der in Mathematik promovierte) und Brashman, Nikolai Dmitrievich, beeinflusst wurde. Nach seinem Abschluss konnte er in Moskau keine Lehrtätigkeit mehr finden und ging nach St. Petersburg, wo er Konferenzen über Algebra und Wahrscheinlichkeitstheorie organisierte. Im Jahr 1859 absolvierte er den Wahrscheinlichkeitskurs von Buni Akovsky, Viktor Jakowlewitsch an der Universität St. Petersburg.

Sein Name lebt durch die von ihm bewiesene **Tschebyscheff-Ungleichung** (auch Bienaymé-Tschebyscheff-Ungleichung genannt) weiter. Dies wurde auf Französisch veröffentlicht, kurz nachdem Bien **aymé, Irénée-Jules** einen Artikel zum gleichen Thema im *Journal de Math ématiques Pures et Appliquées* (auch Journal of Liouville genannt) veröffentlicht hatte.

Er leitete rigorose Arbeiten zur Erstellung einer allgemeinen Version des **zentralen Grenzwerttheos** ein **rem** und gilt als Gründer von

die mathematische Schule von St. Petersburg.

Einige Hauptwerke und Artikel von Tschebyschew, Pafnutii Lwowitsch:

1845 Ein Essay über die Elementaranalyse der Wahrscheinlichkeitstheorie (Dissertation) Crelles Tagebuch.

1867 Preuve de l'inégalité de Tchebychev. J. Mathe. Rein. Appl., 12, 177–184.

WEITERLESEN

Zentraler Grenzwertsatz

VERWEISE

Heyde, CE, Seneta, E.: IJ Bienaymé. Statistische Theorie erwartet. Springer, Berlin Heidelberg New York (1977)

Tschebyscheffs Ungleichung

Siehe **Gesetz der großen Zahlen.**

Chi-Quadrat-Distanz

Betrachten Sie eine **Häufigkeitstabelle** mit *n* Zeilen und *p* Spalten. Es ist möglich, Zeilenprofile und Spaltenprofile zu berechnen. Lassen Sie uns dann planen

Machine Translated by Google

die *n* oder *p* Punkte aus jedem Profil. Wir können Definieren Sie die **Abstände** zwischen diesen Punkten. Der euklidische Abstand zwischen den Komponenten der Profile, auf dem eine Gewichtung erfolgt definiert ist (jeder Begriff hat ein Gewicht, das ist Kehrwert seiner **Frequenz)** wird Chi-Quadrat-Abstand genannt. Der Name der Entfernung ist abgeleitet von der Tatsache, dass das Mathematische Der Ausdruck, der den Abstand definiert, ist identisch zu dem, was bei der Ausarbeitung des **Chi-Quadrat-Anpassungstest.**

MATHEMATISCHE ASPEKTE

Sei (fij) die **Häufigkeit** der i-ten Zeile und jte Spaltein einer Häufigkeitstabellemitnrowsan *p* Spalten. Der Chi-Quadrat-Abstand zwischen zwei Zeilen *i* und *i* ergibt sich aus der Formel:

$$d(i,i) = \sqrt{\sum_{j=1}^{P} \left(\frac{fij}{fi.} - \frac{fij}{fi.}\right)^2 \cdot \frac{1}{fj}},$$

Wo

fi. ist die Summe der Komponenten des ith Reihe;

fj ist die Summe der Komponenten des jth Spalte;

$\left[\frac{fij}{fi.}\right]$ ist das i-te Zeilenprofil fürj = *1, 2,...,p.*

Ebenso ist der Abstand zwischen zwei Säulen *j* und *j* gegeben durch:

$$d(j,j) = \sqrt{\sum_{i=1}^{N} \left(\frac{fij}{fj} - \frac{fij}{fj}\right)^2 \cdot \frac{1}{fi.}},$$

wo $[fij \quad \frac{}{fj}]$ ist das j-te Spaltenprofil für *j* = *1,...,n.*

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Chi-Quadrat-Abstand beinhaltet ein Gewicht, das umgekehrt proportional zu ist die Summe jeder Zeile (oder Spalte), die erhöht die Bedeutung kleiner Abweichungen in den Zeilen (oder Spalten), die vorhanden sind eine kleine Summe im Vergleich zu denen mit mehr wichtiges Summenpaket.

Der Chi-Quadrat-Abstand hat die Eigenschaft Verteilungsäquivalenz, was bedeutet, dass es sorgt dafür, dass die **Abstände** zwischen Reihen und Spalten sind invariant, wenn zwei Spalten (bzw zweireihig) mit identischen Profilen werden aggregiert.

BEISPIELE

Betrachten Sie eine **Notfalltabelle ,** die die Vorgehensweise aufzeigt Zufriedene Mitarbeiter, die für drei verschiedene Unternehmen arbeiten, sind zufrieden. Stellen Sie **eine Distanz** her **Tabelle** unter Verwendung des Chi-Quadrat-Abstands.

Werte für die untersuchte **Variable X** können fallen eine von drei **Kategorien:**

- X1: hohe Zufriedenheit;
- X2: mittlere Zufriedenheit;
- X3: geringe Zufriedenheit.

Die aus Proben von gesammelten Beobachtungen Nachfolgend sind Personen aus den drei Unternehmen aufgeführt:

	Busi ness 1	Busi Ness 2	Busi Ness 3	Gesamt
X1	20	55	30	105
X2	18	40	15	73
X3	12	5	5	22
Insgesamt	50	100	50	200

Die relative **Häufigkeitstabelle** wird erhalten durch Alle Elemente der Tabelle durch dividieren 200, die Gesamtzahl der Beobachtungen:

	Busi ness 1	Busi Ness 2	Busi Ness 3	Gesamt
X1	0,1	0,275	0,15	0,525
X2	0,09	0,2	0,075	0,365
X3	0,06	0,025	0,025	0,11
Insgesamt	0,25	0,5	0,25	1

Machine Translated by Google

Wir können den Unterschied in der Mitarbeiterzufriedenheit zwischen den drei Unternehmen berechnen. Die Säulenprofilmatrix ist unten angegeben:

	Busi ness 1	Busi Ness 2	Busi Ness 3	Gesamt
X1	0,4	0,55	0,6	1,55
X2	0,36	0,4	0,3	1.06
X3	0,24	0,05	0,1	0,39
Gesamt 1		1	1	3

Dadurch können wir die **Entfernungen** berechnen zwischen den verschiedenen Spalten:

$$d2(1,2) = \frac{1}{0,525} \cdot (0,4 \text{ ÿ } 0,55)^2 + \frac{1}{0,365} \cdot (0,36 \text{ ÿ } 0,4)^2 + \frac{1}{0,11} \cdot (0,24 \text{ ÿ } 0,05)^2 = 0,375423$$

$$d(1,2) = 0,613$$

Wir können d(1, 3) und d(2, 3) auf ähnliche Weise berechnen. Die ermittelten Distanzen sind in der folgenden **Distanztabelle** zusammengefasst :

	Busi ness 1	Busi Ness 2	Busi Ness 3
Geschäft 1	0	0,613	0,514
Geschäft 2	0,613	0	0,234
Geschäft 3	0,514	0,234	0

Wir können auch die **Abstände** zwischen berechnen die Zeilen, also der Unterschied in Mitarbeiterzufriedenheit; dazu brauchen wir die Linienprofiltabelle:

	Busi ness 1	Busi Ness 2	Busi Ness 3	Gesamt
X1	0,19	0,524	0,286	1
X2	0,246	0,548	0,206	1
X3	0,546	0,227	0,227	1
Insgesamt	0,982	1,299	0,719	3

Dadurch können wir die **Entfernungen** berechnen zwischen den verschiedenen Zeilen:

$$d2(1,2) = \frac{1}{0,25} \cdot (0,19 \text{ ÿ } 0,246)^2 + \frac{1}{0,5} \cdot (0,524 \text{ ÿ } 0,548)^2 + \frac{1}{0,25} \cdot (0,286 \text{ ÿ } 0,206)^2 = 0,039296$$

$$d(1,2) = 0,198$$

Wir können (1, 3) und d(2, 3) auf ähnliche Weise berechnen. Die Unterschiede zwischen den Graden der Mitarbeiterzufriedenheit sind abschließend in der folgenden **Distanztabelle** zusammengefasst :

	X 1	X 2	X 3
X1	0	0,198	0,835
X2	0,198	0	0,754
X3	0,835	0,754	0

WEITERLESEN

Kontingenztabelle
Distanz
Entfernungstabelle
Frequenz

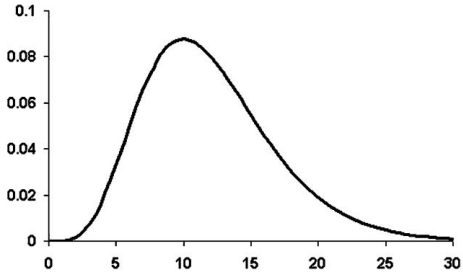
Chi-Quadrat-Verteilung

Eine **Zufallsvariable** *X* folgt einem Chi-Quadrat Verteilung mit *n* **Freiheitsgradenwenn** es **Dichtefunktion** ist:

$$f(x) = \frac{2^{-\frac{N}{2}} \text{ ÿ } 1 \quad \text{Erfahrung} \quad \bar{y}x^{\frac{N}{2}}}{2^{\frac{N}{2}}}, \quad x \text{ ÿ } 0 \quad ,$$

Wo ist die Gammafunktion **(siehe Gamma Verteilung).**

Machine Translated by Google



ÿ2- Verteilung, ÿ = 12

Die Chi-Quadrat-Verteilung ist **kontinuierlich Wahrscheinlichkeitsverteilung.**

GESCHICHTE

Nach Sheynin (1977) ist das Chi-Quadrat Die Verbreitung wurde von Ernst Karl entdeckt Abbe im Jahr 1863. Maxwell erhielt es für drei Freiheitsgrade vor ein paar Jahren (1860) und Boltzman entdeckten den General Fall im Jahr 1881.

Laut Lancaster (1966) Bienaymé erhielt 1838 die Chi-Quadrat-Verteilung als Grenze des diskreten **Laufs dom-Variable**

$$\sum_{i=1}^{k} \frac{(ni \ddot{y} npi)2}{npi},$$

wenn (N1, *N2,...,Nk)* einer gemeinsamen Multinomialverteilung der **Parameter** *n,* p1 folgen, *p2,...,pk.*
Ellis zeigte 1844, dass die Summe von *k* **Zufallsvariablen** entsprechend verteilt zu einer Chi-Quadrat-Verteilung mit zwei **Grad der** Freiheit folgt einer Chi-Quadrat-Verteilung mit *2k* Freiheitsgraden. Die allgemeine Das Ergebnis wurde 1852 von Bien Aymé demonstriert.
Die Arbeiten **von Pearson, Karl** sind auf diesem Gebiet sehr wichtig. Im Jahr 1900 nutzte er das Chi Quadratverteilung zur Annäherung an die Chi-Quadrat- **Statistik ,** die in verschiedenen Tests verwendet wird auf **Kontingenztabellen.**

MATHEMATISCHE ASPEKTE

Die Chi-Quadrat-Verteilung erscheint im Theorie verteilter **Zufallsvariablen** gemäß einer **Normalverteilung.** In diesem, es ist die Verteilung der Summe der Quadrate von normale, zentrierte und reduzierte Zufallsvariablen (mit einem **Mittelwert** gleich 0 und einer **Varianz) .** gleich 1).
Betrachten Sie Z1, *Z2,...,Zn, n* unabhängig, Standardnormale **Zufallsvariablen.** Ihre Quadratsumme:

$$X = Z2_1 + Z2_2 + ... + Z2_N = \sum_{i=1}^{N} Z2_i$$

ist eine Zufallsvariable, die gemäß verteilt ist eine Chi-Quadrat-Verteilung mit *n* **Graden von Freiheit.**

Der **Erwartungswert** der Chi-Quadrat-Verteilung ergibt sich aus:

$$E[X] = n \quad .$$

Die **Varianz** ist gleich:

$$\mathrm{Var}(X) = 2n \quad .$$

Die Chi-Quadrat-Verteilung hängt mit anderen zusammen **kontinuierliche Wahrscheinlichkeitsverteilungen:**

- Die Chi-Quadrat-Verteilung ist eine Besonderheit Fall der **Gammaverteilung.**
- Wenn zwei **Zufallsvariablen** X1 und X2 einer Chi-Quadrat-Verteilung mit n1 bzw. n2 **Freiheitsgraden** folgen , dann die Zufallsvariable

$$Y = \frac{X1/n1}{X2/n2}$$

folgt **einer Fisher-Verteilung mit** n1 und n2 Freiheitsgrade.
- Wenn die Anzahl **der Freiheitsgrade** *n* gegen Unendlich geht, tendiert die Chi-Quadrat-Verteilung (relativ langsam) in Richtung einer **Normalverteilung.**

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Chi-Quadrat-Verteilung wird in vielen Ansätzen zum Testen von Hypothesen verwendet . Der wichtigste ist der **Test auf Güte der Anpassung ,** bei dem die beobachteten **Häufigkeiten** mit den hypothetischen Häufigkeiten bestimmter Klassen verglichen werden.
Es wird auch für Vergleiche zwischen der **beobachteten Varianz** und der hypothetischen Varianz normalverteilter **Stichproben** und zum Testen der **Unabhängigkeit** zweier **Variablen verwendet.**

WEITERLESEN

- **Chi-Quadrat-Anpassungstest**
- **Chi-Quadrat-Tisch**
- **Chi-Quadrat-Test**
- **Kontinuierliche Wahrscheinlichkeitsverteilung**
- **Fisher-Verteilung**
- **Gammaverteilung**
- **Normalverteilung**

VERWEISE

Lancaster, HO: Vorläufer des Pear-Son-Chi-Quadrats. Aust. J. Stat. 8, 117–126 (1966)

Pearson, K.: Über das Kriterium, dass ein gegebenes System von Abweichungen vom Wahrscheinlichen im Fall eines korrelierten Variablensystems so beschaffen ist, dass vernünftigerweise angenommen werden kann, dass es durch Zufallsstichproben entstanden ist. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, S. 339–357. Erstveröffentlichung 1900 in Philos.Mag. (5. Ser) **50,** 157–175 (1948)

Sheynin, OB: Zur Geschichte einiger statistischer Verteilungsgesetze. In: Kendall, M., Plackett, RL (Hrsg.) Studies in the History of Statistics and Probability, vol. II. Grif fin, London (1977)

Chi-Quadrat Gütetest

Der Chi-Quadrat-Anpassungstest ist neben dem **Kolmogorov-Smirnov-Test** einer der am häufigsten verwendeten Anpassungsgütetests
Tests.
Mit diesem Test soll festgestellt werden, ob dies der Fall ist
Es ist möglich, eine beobachtete Verteilung durch eine bestimmte **Wahrscheinlichkeitsverteilung** (**Normalverteilung, Poisson-Verteilung** usw.) anzunähern.

GESCHICHTE

Der Chi-Quadrat-Anpassungstest ist der älteste und bekannteste Test auf die Anpassungsgüte. Es wurde erstmals 1900 von **Pearson, Karl, vorgestellt.**

MATHEMATISCHE ASPEKTE

Sei X1,...,Xn eine Stichprobe von Beobachtungen. Die Schritte zur Durchführung des Chi-Quadrat-Anpassungstests lauten dann wie folgt: 1.
Stellen Sie die Hypothese auf. **Die Nullhypothese** wird die folgende Form annehmen:

$$H0 : F = F0 ,$$

wobei F0 die angenommene **Verteilungsfunktion** der Verteilung ist.

2. Verteilen Sie die Beobachtungen in *k* disjunkte Klassen:

$$[aiÿ1, ai] .$$

Wir bezeichnen die Anzahl der Beobachtungen enthalten in der i-ten Klasse, $i = 1,...,k,$ von $ni.$

3. Berechnen Sie die theoretischen Wahrscheinlichkeiten für jede Klasse auf Basis der angenommenen Verteilungsfunktion F0:

$$pi = F0(ai) ÿ F0(aiÿ1) , \qquad i = 1,...,k.$$

Machine Translated by Google

4. Ermitteln Sie die erwarteten Häufigkeiten für alle Klasse

$$ei = n \cdot pi\,, \qquad i = 1,\ldots,k,$$

wobei *n* die Größe der **Stichprobe ist.**

5. Berechnen Sie die ÿ2- Statistik (Chi-Quadrat):

$$ÿ2 = \sum_{i=1}^{k} \frac{(ni ÿ ei)^2}{ei}\,.$$

Wenn H0 wahr ist, folgt die ÿ2- Statistik einer **Chi-Quadrat-Verteilung** mit *v* Graden von Freiheit, wo:

$$v = k ÿ 1 ÿ \text{Anzahl der geschätzten Parameter}\,.$$

Zum Beispiel beim Testen der Güte Aufgrund der Anpassung an eine **Normalverteilung** beträgt die Anzahl der Freiheitsgrade:

- *k* ÿ 1, wenn der **Mittelwert** ÿ und die **Standardabweichung** ÿ der **Grundgesamtheit** sind sind bekannt;
- kÿ2wenn einer vonÿoderÿ unbekannt ist und wird geschätzt, um fortfahren zu können mit dem Test;
- *k* ÿ 3, wenn beide Parameter ÿ und ÿ sind unbekannt und beide werden geschätzt aus die entsprechenden Werte der Probe.

6. Verwerfen Sie H0 , wenn die Abweichung zwischen dem beobachteten und geschätzten Häufigkeiten ist groß; das ist:

$$\text{wenn } ÿ2 > ÿ2_{v,ÿ}\,,$$

wobei $ÿ2_{v,ÿ}$ ist der im **Chi** angegebene **Wert quadratischer Tisch** für eine **besondere Bedeutung Ebene** ÿ.

DOMÄNEN UND EINSCHRÄNKUNGEN

Um die Chi-Quadrat-Güte von fittest,it anzuwenden Wichtig ist, dass *n* groß genug ist und dass die Die geschätzten Häufigkeiten *ei* sind nicht zu klein. Normalerweise geben wir an, dass die geschätzten Häufigkeiten größer als 5 sein müssen, mit Ausnahme von extreme Klassen, in denen sie kleiner sein können dann 5, aber größer als 1. Wenn diese Einschränkung gilt Wenn wir nicht zufrieden sind, müssen wir die Klassen neu gruppieren um dieser Regel zu genügen.

BEISPIELE

Güte der Binomialanpassung Verteilung

Wir werfen viermal eine Münze und zählen wie oft „Köpfe“ vorkommt.
Dieses **Experiment** wird 160 Mal durchgeführt.
Die beobachteten Häufigkeiten sind wie folgt:

Anzahl der „Köpfe“ xi	Anzahl der Experimente (ni)
0	17
1	52
2	54
3	31
4	6
Gesamt	160

1. Wenn **das Experiment** korrekt durchgeführt wurde und die Münze nicht gefälscht ist, erhält man die Verteilung der Anzahl „Kopf“. sollte **der Binomialverteilung folgen.**
   Wir stellen dann eine **Nullhypothese** auf , dass die Die beobachtete Verteilung kann durch die Binomialverteilung angenähert werden, und wir Ich werde mit einem **Eignungstest** fortfahren um festzustellen, ob diese **Hypothese** akzeptiert werden kann oder nicht.
2. In diesem Beispiel ist die unterschiedliche Anzahl von Berücksichtigt werden jeweils die pro Versuch erreichbaren „Köpfe“ (0, 1, 2, 3 und 4). eine Klasse sein.

Machine Translated by Google

3. Die **Zufallsvariable** *X* (Anzahl der „Kopf“ nach vier Würfen von erzielt eine Münze) folgt einer Binomialverteilung, wenn

$$P(X = x) = Cx_{N} \cdot px \cdot qnyx ,$$

Wo:

n ist die Anzahl der unabhängigen Versuche = 4;

p ist die **Erfolgswahrscheinlichkeit** („Köpfe“) = 0,5;

q ist die Wahrscheinlichkeit eines Ausfalls („Tails“) = 0,5;

$C\chi_N$ ist die Anzahl der Kombinationen von *x* Objekte aus n.

Wir haben dann die folgende Theorie Wahrscheinlichkeiten für vier Würfe:

$$P(X = 0) = \frac{1}{16}$$
$$P(X = 1) = \frac{4}{16}$$
$$P(X = 2) = \frac{6}{16}$$
$$P(X = 3) = \frac{4}{16}$$
$$P(X = 4) = \frac{1}{16}$$

4. Nachdem das Experiment durchgeführt wurde Das 160-fache der erwarteten Anzahl an Köpfen für jeden möglichen **Wert** von *X* ist gegeben durch:

$$ei = 160 \cdot P(X = xi) .$$

Wir erhalten die folgende Tabelle:

Anzahl „Köpfe“ xi	Beobachtet Frequenz (hi	Erwartet Frequenz (ei)
0	17	10
1	52	40
2	54	60
3	31	40
4 610		
Gesamt	160	160

5. Die ÿ2- Statistik (Chi-Quadrat) lautet dann:

$$ÿ2 = \sum_{i=1}^{k} \frac{(ni \ ÿ \ ei)2}{ei} ,$$

wobei *k* die Anzahl der möglichen Werte ist von X.

$$ÿ2 = \frac{(17 ÿ 10)2}{10} + ... + \frac{(6 ÿ 10)2}{10}$$

$$= 12,725 .$$

6. Wir wählen **ein Signifikanzniveau von** 5 % Finden Sie den Wert von $ÿ2_{v,ÿ}$ für *k* ÿ 1 = 4 Freiheitsgrade sind:

$$ÿ2_{4,0,05} = 9,488 .$$

Da der berechnete Wert von ÿ2 größer ist als der aus der Tabelle ermittelte Wert, Wir lehnen die **Nullhypothese** ab und kommen zu dem Schluss, dass dies bei der Binomialverteilung der Fall ist geben keine gute Annäherung an unsere beobachtete Verteilung. Wir können dann con schlussfolgern, dass die Münzen wahrscheinlich gefälscht waren, oder dass sie nicht richtig geworfen wurden.

Güte der Anpassung an die Normalverteilung

Untersucht wurden die Durchmesser der von einer Fabrik hergestellten Kabel.
Nachfolgend finden Sie eine **Häufigkeitstabelle** der beobachteten Durchmesserverteilung:

Kabeldurchmesser (in mm)	Beobachtete Häufigkeit ni
19.70–19.80 Uhr	5
19.80–19.90 Uhr	12
19.90–20.00 Uhr	35
20.00–20.10	42
20.10–20.20	28
20.20–20.30 Uhr	14
20.30–20.40 Uhr	4
Gesamt	140

Machine Translated by Google

1. Wir führen einen **Anpassungstest** für durch eine **Normalverteilung.** Die Nullhypothese lautet daher, dass die beobachtete Verteilung

   Bution kann durch eine Normale angenähert werden Verteilung.

2. Die vorherige Tabelle zeigt die Beobachtungen Durchmesser in Klassen eingeteilt.

3. Wenn die Zufallsvariable *X* (Kabeldurchmesser) folgt der Normalverteilung, der Zufallsvariablen

$$Z = \frac{X - \mu}{\sigma}$$

folgt einer Standardnormalverteilung.

Der **Mittelwert** ü und die **Standardabweichung ÿ der Grundgesamtheit** sind unbekannt und werden anhand des Mittelwerts $\bar{x}$ und des geschätzt Standardabweichung *S* der **Stichprobe:**

$$\bar{x} = \frac{\sum_{i=1}^{7} yi \cdot ni}{N} = \frac{2806.40}{140} = 20.05 \quad ,$$

$$S = \sqrt{\frac{\sum_{i=1}^{7} ni \cdot (yi - \bar{x})^2}{n - 1}} = \sqrt{\frac{2.477}{139}}$$

$$= 0{,}134$$

wobei die *yi* die Mittelpunkte der Klasse es sind (in diesem Beispiel der mittlere Durchmesser von eine Klasse; für $i = 1$: ÿ1 = 19,75) und *n* ist die Gesamtzahl der Beobachtungen.
Dann können wir das Theoretische berechnen

Wahrscheinlichkeiten, die jeder Klasse zugeordnet sind. Die detaillierten Berechnungen für die ersten beiden Klassen sind:

$$p1 = P(X < 19{,}8) = P\left(Z < \frac{19{,}8 - \bar{x}}{S}\right) = P(Z < -1{,}835) = 1 - P(Z < 1{,}835)$$

$$p2 = P(19{,}8 < X < 19{,}9) = P\left(\frac{19{,}8 - \bar{x}}{S} < Z < \frac{19{,}9 - \bar{x}}{S}\right) = P(-1{,}835 < Z < -1{,}089) = P(Z < 1{,}835) - P(Z < 1{,}089)$$

Diese Wahrscheinlichkeiten können anhand der **Normaltabelle ermittelt werden.** Wir bekommen:

$$p1 = P(X < 19{,}8) = 0{,}03325$$

$$p2 = P(19{,}8 < X < 19{,}9) = 0{,}10476$$

$$p3 = P(19{,}9 < X < 20{,}0) = 0{,}22767$$

$$p4 = P(20{,}0 < X < 20{,}1) = 0{,}29083$$

$$p5 = P(20{,}1 < X < 20{,}2) = 0{,}21825$$

$$p6 = P(20{,}2 < X < 20{,}3) = 0{,}09622$$

$$p7 = P(X > 20{,}3) = 0{,}02902$$

4. Die erwarteten Häufigkeiten für die Klassen sind dann gegeben durch:

$$ei = n \cdot pi \,,$$

was die folgende Tabelle ergibt:

Kabel Durchmesser (in mm)	Beobachtet Frequenz ni	Erwartet Frequenz ei
19.70–19.80 Uhr	5	4.655
19.80–19.90 Uhr	12	14.666
19.90–20.00 Uhr	35	31.874
20.00–20.10	42	40.716
20.10–20.20	28	30.555
20.20–20.30 Uhr	14	13.471
20.30–20.40 Uhr	4	4.063
Gesamt	140	140

5. Die ÿ2- Statistik (Chi-Quadrat) lautet dann:

$$\chi^2 = \sum_{i=1}^{k} \frac{(ni - ei)^2}{ei} \,,$$

Machine Translated by Google

wobei $k = 7$ die Anzahl der Klassen ist.

$$\ddot{y}2 = \frac{(5 \ddot{y} 4{,}655)2}{4.655} + \frac{(12 \ddot{y} 14{,}666)2}{14.666} + \ldots + \frac{(4 \ddot{y} 4.063)2}{4.063} = 1{,}0927 .$$

6. Wahl eines Signifikanzniveaus ÿ = 5 %,
Wir finden, dass der Wert von $\ddot{y}2_{\ddot{y},\ddot{y}}$ mit kÿ3 = 7 ÿ 3 = 4 Freiheitsgrade in der Chi-Quadrat-Tabelle ist:

$$\ddot{y}2_{4,0,05} = 9{,}49 .$$

Da der aus ÿ2 berechnete Wert ist kleiner als der aus dem erhaltene Wert

In der Chi-Quadrat-Tabelle lehnen wir die Null nicht ab Hypothese und wir kommen zu dem Schluss, dass der Unterschied zwischen der beobachteten Verteilung und die Normalverteilung ist bei einem **Signifikanzniveau** von 5 % nicht signifikant.

WEITERLESEN

Chi-Quadrat-Verteilung
Chi-Quadrat-Tisch
Anpassungstest
Hypothesentest
Kolmogorov-Smirnov-Test

REFERENZ

Pearson, K.: Über das Kriterium, dass eine Gegebenheit System der Abweichungen vom Wahrscheinlichen in Bei einem korrelierten Variablensystem kann vernünftigerweise davon ausgegangen werden, dass es durch Zufallsstichproben entstanden ist. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, S. 339–357. Erstveröffentlichung 1900 in Philos.Mag. (5. Ser) **50,** 157–175 (1948)

Chi-Quadrat-Tisch

Die Chi-Quadrat-Tabelle gibt die Werte an aus der **Verteilungsfunktion** erhalten einer **Zufallsvariablen ,** die einem **Chi** folgt **quadratische Verteilung.**

GESCHICHTE

Einer der ersten Chi-Quadrat-Tische wurde 1902 von Elderton veröffentlicht. Es beinhaltet **Verteilungsfunktionswerte ,** die angegeben werden auf sechs Dezimalstellen genau.

Im Jahr 1922 berichtete **Pearson, Karl** über eine Tabelle mit Werte für die unvollständige Gammafunktion, bis auf sieben Dezimalstellen.

MATHEMATISCHE ASPEKTE

Sei X eine folgende **Zufallsvariable** eine **Chi-Quadrat-Verteilung** mit v Grad von Freiheit. Die **Dichtefunktion** der Zufallsvariablen X ist gegeben durch:

$$f(t) = \frac{\tau^{\frac{v}{2}\ddot{y}1} \exp \ddot{y} 2^{\tau}}{2^{\frac{v}{2}} \quad \frac{v}{2}} , \qquad t \ddot{y} 0 \quad ,$$

wo stellt die Gammafunktion dar (siehe **Gammaverteilung).**
Die **Verteilungsfunktion** des Zufalls Variable X ist definiert durch:

$$F(x) = \mathrm{P}(X \ddot{y} x) = \int_0^x f(t)\, dt .$$

Die Chi-Quadrat-Tabelle gibt die Werte der an Verteilungsfunktion $F(x)$ für verschiedene Werte von v.

Wir verwenden die Chi-Quadrat-Tabelle oft in umgekehrter Reihenfolge, um den Wert von x zu ermitteln , der einer gegebenen Wahrscheinlichkeit entspricht .
Wir bezeichnen im Allgemeinen als $\ddot{y}2_{\ddot{y},\ddot{y}}$ der Wert der Zufallsvariable X für welche

$$\mathrm{P}(X \ddot{y} \ddot{y}2\ v,\ddot{y}) = 1 \ddot{y} \ddot{y} .$$

Hinweis: Die Schreibweise ÿ2 lautet „Chi-Quadrat“.

Machine Translated by Google

BEISPIELE

Siehe Anhang F.

Die Chi-Quadrat-Tabelle ermöglicht es uns für eine gegebene Zeit
Anzahl der Freiheitsgrade v, um meine zu bestimmen:

1. Der **Wert** der **Verteilungsfunktion**
   F(x), gegebenes x.
2. Der Wert von ÿ2 $_{v,ÿ,}$ angesichts der **Wahrscheinlichkeit**
   P(X ÿ ÿ2 v,ÿ).

WEITERLESEN

- **Chi-Quadrat-Verteilung**
- **Chi-Quadrat-Anpassungstest**
- **Chi-Quadrat-Test**
- **Chi-Quadrat-Test der Unabhängigkeit**
- **Statistische Tabelle**

VERWEISE

Elderton, WP: Tabellen zum Testen der guten
Anpassung der Theorie an die Beobachtung.
Biometrie 1, 155–163 (1902)

Pearson, K.: Tabellen des Unvollständigen -
Funktion. HM Stationery Office (Cambridge University
Press, Cambridge seit
1934), London (1922)

Chi-Quadrat-Test

Es gibt eine Reihe von Chi-Quadrat-Tests
Dabei werden die Testergebnisse mit verglichen
die Werte **aus der Chi-Quadrat-Verteilung.**
Die bekanntesten dieser Tests sind Einführungstests
unten aufgeführt:

- Der **Chi-Quadrat-Test der Unabhängigkeit** ist
  Wird verwendet, um zu bestimmen, ob zwei
  qualitative kategoriale Variablen miteinander verbunden sind
  eine **Stichprobe** sind unabhängig.
- Der **Chi-Quadrat-Anpassungstest** ist
  Wird verwendet, um zu bestimmen, ob die Verteilung
  Der für eine Probe beobachtete Wert kann durch
  eine theoretische Verteilung angenähert werden. Wir
  Vielleicht möchten Sie zum Beispiel wissen, ob
  die für das Sam beobachtete Verteilung
  ple entspricht einer bestimmten
  Wahrscheinlichkeitsverteilung (Normalverteilung,
  Poisson-Verteilung usw.).
- Der Chi-Quadrat-Test für eine unbekannte **Vari**
  ance wird verwendet, wenn wir testen wollen, ob
  Diese Varianz nimmt eine bestimmte Konstante an
  Wert.
- Zur Prüfung wird der Chi-Quadrat-Test verwendet
  Homogenität der berechneten Varianzen
  für viele Proben, die aus einem Normal entnommen wurden
  verteilte **Bevölkerung.**

GESCHICHTE

Im Jahr 1937 schlug MS Bartlett eine Methode vor
Testen der Homogenität der **Varianz** für
viele Stichproben aus einer normalverteilten **Population.**

Siehe auch **Chi-Quadrat-Test der Unabhängigkeit**
und **Chi-Quadrat-Anpassungstest.**

MATHEMATISCHE ASPEKTE

Die mathematischen Aspekte des **Chi-Quadrats**
Der Unabhängigkeitstest und der **Chi-Quadrat-**
Anpassungstest werden in behandelt
die entsprechenden Einträge.
Der Chi-Quadrat-Test dient zur Überprüfung, ob
Eine unbekannte **Varianz** nimmt einen bestimmten
konstanten **Wert** an :
Sei *(x1,...,xn)* eine Zufallsstichprobe **aus** einer
normalverteilten **Grundgesamtheit**
mit unbekanntem **Mittelwert** ÿ und unbekannter **Varianz**
ÿ2.

Wir haben guten Grund zu der Annahme, dass die Varianz
der Grundgesamtheit einem angenommenen Wert entspricht
sagt ÿ2 $_0$. Die Hypothesen für jeden Fall sind
nachstehend beschrieben.

A: Zweiseitiger Fall:

$$H0: \quad ÿ2 = ÿ2_0$$

$$H1: \quad ÿ2 = ÿ2_0$$

Machine Translated by Google

B: Einseitiger Test:

$$H0: \quad \ddot{y}2 \ddot{y} \ddot{y}2_0$$

$$H1: \quad \ddot{y}2 > \ddot{y}2_0$$

C: Einseitiger Test:

$$H0: \quad \ddot{y}2 \ddot{y} \ddot{y}2_0$$

$$H1: \quad \ddot{y}2 < \ddot{y}2_0$$

Anschließend ermitteln wir die **Statistik** des Gegebenen Chi-Quadrat-Test mit:

$$\ddot{y}2 = \frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}{\ddot{y}2_0}.$$

Diese Statistik ist unter H0 Chi-Quadrat-verteilt mit *n* ÿ 1 Freiheitsgraden. In Mit anderen Worten, wir suchen nach dem Wert von $\ddot{y}2_{n\ddot{y}1,\ddot{y}}$ in der **Chi-Quadrat-Tabelle,** und dann vergleichen wir diesen Wert mit dem berechneten Wert ÿ2. Die Entscheidungsregeln hängen vom Einzelfall ab und sind wie folgt.

Fall A

Wenn $\ddot{y}2 \ddot{y} \ddot{y}2_{n\ddot{y}1,\ddot{y}1}$ oder wenn $\ddot{y}2 \ddot{y} \ddot{y}2_{n\ddot{y}1,1\ddot{y}\ddot{y}2}$ Wir lehnen Sie die **Nullhypothese** H0 für die **Alternativhypothese** H1 ab, bei der wir gespalten haben das **Signifikanzniveau** ÿ in ÿ1 und ÿ2 wie z dass ÿ1+ÿ2 = ÿ. Ansonsten lehnen wir nicht ab die Nullhypothese H0.

Fall B

Wenn $\ddot{y}2 < \ddot{y}2_{n\ddot{y}1,\ddot{y}}$ Wir lehnen die Null nicht ab Hypothese H0. Wenn $\ddot{y}2 \ddot{y} \ddot{y}2_{n\ddot{y}1,\ddot{y}}$ wir lehnen das ab Nullhypothese H0 für die Alternativhypothese H1.

Fall C

Wenn $\ddot{y}2 > \ddot{y}2_{n\ddot{y}1,1\ddot{y}\ddot{y}}$ Wir lehnen das nicht ab Nullhypothese H0. Wenn $\ddot{y}2 \ddot{y} \ddot{y}2_{n\ddot{y}1,1\ddot{y}\ddot{y}}$ Wir Lehnen Sie die Nullhypothese H0 für die Alternativhypothese H1 ab.
Weitere Chi-Quadrat-Tests werden im vorgeschlagen Arbeit von Ostle, B. (1963).

DOMÄNEN UND EINSCHRÄNKUNGEN

Die ÿ2 (Chi-Quadrat)-Statistik muss berechnet werden unter Verwendung absoluter und nicht relativer Häufigkeiten Einsen.

Beachten Sie, dass der Chi-Quadrat-Test unzuverlässig sein kann für kleine Proben, insbesondere wenn einige davon die geschätzten Häufigkeiten sind gering (< 5). Dieses Problem kann oft durch die Gruppierung von Kategorien gelöst werden, sofern eine solche Gruppierung vorhanden ist vernünftig.

BEISPIELE

Betrachten Sie eine Charge von Artikeln, die von hergestellt werden ein Gerät. Sie können unterteilt werden in Klassen abhängig von ihren Durchmessern (in mm), wie in der folgenden Tabelle:

Durchmesser (mm) x_i	Anzahl der Teile n_i
59,5	2
59,6	6
59,7	7
59,8	15
59.9	17
60,0	25
60.1	15
60.2	7
60.3	3
60.4	1
60,5	2
Gesamt	$N = 100$

Wir lassen eine **Zufallsstichprobe** ziehen eine normalverteilte **Population, wobei die Mittelwert** und **Varianz** sind nicht bekannt. Der Anbieter dieser Artikel möchte die Varianz ÿ2 kleiner oder gleich 0,05 sein. Wir prüfen

Machine Translated by Google

die folgenden Hypothesen:

Nullhypothese H0 :	ÿ2 ÿ 0,05
Alternativhypothese H1 :	ÿ2 > 0,05 .

In diesem Fall verwenden wir die einseitige Hypothese prüfen.

Wir beginnen mit der Berechnung des Mittelwerts für die Stichprobe:

$$x^- = \frac{\sum_{i=1}^{11} ni \cdot xi}{N} = \frac{5995}{100} = 59{,}95 .$$

Wir können dann die ÿ2 -Statistik berechnen :

$$ÿ2 = \frac{\sum_{i=1}^{11} ni \cdot (xi ÿ \bar{x})2}{ÿ2_0}$$

$$= \frac{2 \cdot (ÿ0{,}45)2 + ... + 2 \cdot (0{,}55)2}{0{,}05}$$

$$= \frac{3{,}97}{0{,}05} = 79{,}4 .$$

Unter Verwendung eines **Signifikanzniveaus** von ÿ = 5 % haben wir Finden Sie dann den **Wert** von ÿ2 99,0,05 (= *123,2)* Zoll der **Chi-Quadrat-Tisch.**

Da ÿ2 = *79,4* < ÿ2 *99,0,05,* wir lehnen nicht ab die Nullhypothese, was bedeutet, dass die Der Verkäufer sollte diese Artikel gerne verkaufen da sie sich nicht wesentlich unterscheiden Durchmesser.

WEITERLESEN

- **Chi-Quadrat-Verteilung**
- **Chi-Quadrat-Anpassungstest**
- **Chi-Quadrat-Tisch**
- **Chi-Quadrat-Test der Unabhängigkeit**

VERWEISE

Bartlett, MS: Einige Beispiele für statistische Forschungsmethoden in der Landwirtschaft und Angewandte Biologie. J. Roy. Stat. Soc. (Ergänzung) 4, 137–183 (1937)

Ostle, B.: Statistik in der Forschung: Grundlegende Konzepte und Techniken für Forscher. Iowa State College Press, Ames, IA (1954)

C

Chi-Quadrat-Test der Unabhängigkeit

Der Chi-Quadrat-Test der Unabhängigkeit zielt darauf ab Bestimmen Sie, ob zwei Variablen verknüpft sind mit einer **Stichprobe** unabhängig sind oder nicht. Der Die untersuchten Variablen sind kategorisch qualitativ Variablen.

Der Chi-Quadrat-Unabhängigkeitstest wird mithilfe einer **Kontingenztabelle durchgeführt.**

GESCHICHTE

Die ersten Kontingenztabellen wurden nur verwendet zur Aufzählung. Allerdings ermutigt von das Werk von **Quetelet, Adolphe** (1849), Statistiker begannen, sich dafür zu interessieren Assoziationen zwischen den verwendeten Variablen die Tische. Zum Beispiel **Pearson, Karl** (1900) führte grundlegende Arbeiten zu Kontingenztabellen durch.

Yule, George Udny (1900) vorgeschlagen ein etwas anderer Ansatz zur Untersuchung von Kontingenztabellen nach Pearson, die führen zu einer Meinungsverschiedenheit zwischen ihnen. Pearson argumentierte auch mit **Fisher, Ronald Aylmer** über die Anzahl der Freiheitsgrade Verwendung im Chi-Quadrat-Unabhängigkeitstest. Jeder benutzte unterschiedliche Zahlen, bis schließlich bewiesen wurde, dass es sich um Fisher, RA (1922) handelte richtig.

MATHEMATISCHE ASPEKTE

Betrachten Sie zwei qualitative kategoriale Variablen *X* und Y. Wir haben eine **Stichprobe ,** die *n* Beobachtungen dieser Variablen enthält .

Machine Translated by Google

Diese Beobachtungen können in dargestellt werden eine **Kontingenztabelle.**
Wir bezeichnen die beobachtete **Häufigkeit** des **Kategorie** i der Variablen X und der Kategorie j der Variablen *Y* als *nij.*

		Kategorien der Variablen *Y*		
		Y1 ...	*Yc* Gesamt	
Kategorien von Variable *X*	X1	n11 ...	n1c	n1.
	...	...	... ...	...
	Xr	*nr1* ...	*nrc* Gesamt	*Nr.*
		n.1 ...	*nc*	*N..*

Die zu prüfenden Hypothesen sind:

Null-Hyp. H0 : die beiden Variablen sind unabhängig,

Alternative Hyp. H1 : die beiden Variablen sind nicht unabhängig.

Schritte des Tests

1. Berechnen Sie die erwarteten Häufigkeiten, bezeichnet mit *eij,* für jeden Fall in der **Kontingenztabelle** unter der **Unabhängigkeit Hypothese:**

$$eij = \frac{ni. \cdot nj}{N..},$$

$$ni. = \sum_{k=1}^{C} nik \text{ und } nj = \sum_{k=1}^{R} nkj,$$

wobei *c* die Anzahl der Spalten darstellt (oder Anzahl der Kategorien der Variablen *X* in die Kontingenztabelle) *und die* Anzahl der Zeilen (oder die Anzahl der Kategorien der Variablen Y).

2. Berechnen Sie den **Wert** von ÿ2 (Chi-Quadrat) Statistik, die eigentlich ein Maß dafür ist Abweichung der beobachteten Frequenzen *nij* aus den erwarteten Häufigkeiten *eij:*

$$ÿ2 = \sum_{i=1}^{C} \sum_{j=1}^{R} \frac{(nij - eij)2}{eij}.$$

3. Wählen Sie das zu verwendende **Signifikanzniveau ÿ** im Test und vergleichen Sie den berechneten **Wert** von ÿ2 mit dem aus die **Chi-Quadrat-Tabelle,** $ÿ2_{v,ÿ.}$ Die Nummer der Freiheitsgrade entsprechen dem Anzahl der Fälle in der Tabelle, die dauern können beliebige Werte; die von der angenommenen Werte andere Fälle werden ihnen von der auferlegt Zeilen- und Spaltensummen. Also die Anzahl der Freiheitsgrade sind gegeben durch:

$$v = (r - 1)(c - 1).$$

4. Wenn der berechnete ÿ2 kleiner ist, dann ist der $ÿ2_{v,ÿ}$ Aus der Tabelle lehnen wir die Null nicht ab Hypothese. Die beiden Variablen können als unabhängig betrachtet werden.
Allerdings ist das berechnete ÿ2 größer dann das $ÿ2_{v,ÿ}$ Aus der Tabelle lehnen wir das ab Nullhypothese für die Alternativhypothese. Daraus können wir schließen, dass die beiden Variablen sind nicht unabhängig.

DOMÄNEN UND EINSCHRÄNKUNGEN

Dazu müssen bestimmte Bedingungen erfüllt sein um den Chi-Quadrat-Test unabhängig anwenden zu können:

1. Die **Stichprobe,** die *keine* Beobachtungen enthält , muss eine Zufallsstichprobe sein;
2. Jede einzelne Beobachtung kann nur erscheinen für jede **Variable** in einer **Kategorie .** Mit anderen Worten: Jede einzelne Beobachtung kann nur in einer vorkommen Zeile und eine Spalte der **Kontingenz Tisch.**

Beachten Sie, dass der Chi-Quadrat-Unabhängigkeitstest für kleine Stichproben nicht sehr zuverlässig ist vor allem, wenn die geschätzten Frequenzen sind klein (< 5). Wir können dieses Problem vermeiden Gruppieren Sie Kategorien zusammen, aber nur, wenn dies der Fall ist Die erhaltenen Gruppen sind sinnvoll.

Machine Translated by Google

BEISPIELE

Wir wollen feststellen, ob der Anteil der Raucher geschlechtsunabhängig ist.
Die beiden zu untersuchenden Variablen sind kategorisch und qualitativ und enthalten jeweils zwei Kategorien:

- Variable „Geschlecht:“ M oder F;
- Variable „Raucherstatus“: „raucht“ oder „raucht nicht.“

Die Hypothesen lauten dann:

H0 : Die Chance, Raucher zu sein, ist unabhängig vom Geschlecht

H1 : Die Wahrscheinlichkeit, Raucher zu sein, besteht nicht unabhängig vom Geschlecht.

Die **Kontingenztabelle** erhalten von eine **Stichprobe** von 100 Personen *(n* = 100) ist unten dargestellt:

Rauchstatus

		„raucht“	„tut nicht Rauch"	Gesamt
Geschlecht	M	21	44	65
	F	10	25	35
	Insgesamt	31	69	100

Wir bezeichnen nun die beobachteten Frequenzen als *nij (i* = *1,* 2, *j* = *1,* 2).
Anschließend schätzen wir alle Häufigkeiten im Tabelle basierend auf der **Hypothese ,** dass die beiden Variablen sind unabhängig voneinander. Wir bezeichnen diese geschätzten Häufigkeiten mit *eij:*

$$eij = \frac{ni. \cdot nj}{N..} .$$

Wir erhalten daher:

$$e11 = \frac{65 \cdot 31}{100} = 20{,}15$$

$$e12 = \frac{65 \cdot 69}{100} = 44{,}85$$

$$e21 = \frac{35 \cdot 31}{100} = 10{,}85$$

$$e12 = \frac{35 \cdot 69}{100} = 24{,}15 .$$

Die geschätzte **Häufigkeitstabelle** ist angegeben unter:

Rauchstatus

		„raucht“	„tut nicht Rauch"	Gesamt
Geschlecht	M	20.15	44,85	65
	F	10,85	24.15	35
	Insgesamt	31	69	100

Wenn die **Nullhypothese** H0 wahr ist, ist die Statistik

$$ÿ2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(nij - eij)2}{eij}$$

ist Chi-Quadrat-verteilt mit *(r* ÿ 1) *(c* ÿ 1) = (2 ÿ 1) (2 ÿ 1) = 1 Grad der Freiheit und

$$ÿ2 = 0{,}036 + 0{,}016 + 0{,}066 + 0{,}030$$
$$= 0{,}148 .$$

Wenn ein **Signifikanzniveau** von 5 % gewählt wird, beträgt das **Wert** von $ÿ2_{1,0,05}$ beträgt 3,84, aus dem **Chi-Quadrat Tisch.**

Da der berechnete Wert von ÿ2 kleiner ist dann der in der Chi-Quadrat-Tabelle gefundene Wert, Wir lehnen die Nullhypothese nicht ab und wir schlussfolgern, dass die beiden untersuchten Variablen sind unabhängig.

WEITERLESEN

- **Chi-Quadrat-Verteilung**
- **Chi-Quadrat-Tisch**
- **Kontingenztabelle**
- **Test der Unabhängigkeit**

Machine Translated by Google

REFERENZ

Fisher, RA: Zur Interpretation von ÿ2
aus Kontingenztabellen und der Berechnung von
PJRoy. Stat. Soc. Ser. **A85,** 87–94
(1922)

Pearson, K.: Über das Kriterium, dass eine Gegebenheit
System der Abweichungen vom Wahrscheinlichen in
Bei einem korrelierten Variablensystem kann
vernünftigerweise davon ausgegangen werden, dass
es durch Zufallsstichproben entstanden ist. In: Karl
Pearson's Early Statistical Papers. Cambridge
University Press,
S. 339–357. Erstveröffentlichung 1900 in
Philos.Mag. (5. Ser) **50,** 157–175 (1948)

Quetelet, A.: Briefe an HRHthe
Großherzog von Sachsen-Coburg und Gotha, am
die Wahrscheinlichkeitstheorie in ihrer Anwendung
die Moral- und Politikwissenschaften. (Französisch
Übersetzung von Downs, Olinthus Gregory).
Charles und Edwin Layton, London (1849)

Yule, GU: Über die Zuordnung von Attributen
Statistik: mit Illustration vom Kumpel
Rial der Kindergesellschaft.Philos.Trans.
Roy. Soc. London. Ser. A **194,** 257–319
(1900)

Einstufung

Unter Klassifizierung versteht man die Gruppierung von ähnliche Objekte. Wenn jedes Objekt durch *p* Variablen charakterisiert ist , kann eine Klassifizierung erfolgen nach rationalen Kriterien durchgeführt werden. Abhängig von den verwendeten Kriterien ein Objekt könnte möglicherweise mehreren Klassen angehören.

GESCHICHTE

Klassifizierung der Einwohner einer Ortschaft bzw Ein Land nach seinem Geschlecht und anderen körperlichen Merkmalen ist eine Aktivität, die stammt aus der Antike. Die Hindus, die Die alten Griechen und die Römer entwickelten sich alle hat mehrere Typologien für den Menschen entwickelt. Der älteste stammt aus Galen (129–199). ANZEIGE).

Später verbreitete sich das Konzept der Klassifikation zu den Bereichen Biologie und Zoologie; Die In diesem Zusammenhang sind Werke von Linné (1707–1778) zu nennen.

Die ersten Ideen zu tatsächlichen Methoden von **Clusteranalysen** werden Adanson zugeschrieben (achtzehntes Jahrhundert). Zubin (1938), Tryon (1939) und Thorndike (1953) versuchten es ebenfalls Einige Methoden müssen entwickelt werden, aber die tatsächliche Entwicklung der Klassifizierungsmethoden fällt zusammen mit dem Aufkommen des Computers.

MATHEMATISCHE ASPEKTE

Klassifizierungsmethoden können unterteilt werden in zwei große **Kategorien,** eine basierend auf **Wahrscheinlichkeiten** und die andere nicht.

Die erste Kategorie umfasst beispielsweise diskriminierende Analysen. Die zweite Kategorie lassen sich weiter in zwei Gruppen unterteilen. Die erste Gruppe enthält sogenannte optimale Klassifizierungsmethoden. In dieser Sekunde Gruppe können wir zwischen mehreren unterscheiden Unterarten der Klassifizierungsmethode:

- Partitionierungsmethoden, die aus „Verteilen“ bestehen
  ing *n* Objekte unter *g* Gruppen in solchen
  weg, dass jedes Objekt ausschließlich dazugehört
  auf nur eine Gruppe. Die Anzahl der Gruppen *g*
  wird vorher festgelegt, und die Partition
  Die angewandte Methode erfüllt die
  Klassifizierungskriterien am ehesten.
- Partitionierungsmethoden, die verletzende Klassen
  einbeziehen, wo ein Objekt erlaubt ist
  gehören mehreren Gruppen gleichzeitig an.
- Hierarchische Klassifizierungsmethoden,
  wobei die Struktur der **Daten** bei Diff
  Es werden verschiedene Klassifizierungsebenen berücksichtigt
  Konto. Wir können dann berücksichtigen
  Beziehungen, die zwischen den verschiedenen

Machine Translated by Google

verschiedene Gruppen, die während der Partitionierung erstellt wurden.
Zwei verschiedene Arten hierarchischer Technologie
Es gibt Techniken: Agglomerationstechniken und Trenntechniken.
Agglomerierende Methoden beginnen mit getrennten Objekten, das heißt, die *n* Objekte werden zunächst in *n* Gruppen verteilt. In jedem weiteren Schritt werden zwei Gruppen zusammengeballt, bis nur noch eine Gruppe übrig bleibt.

Im Gegensatz dazu beginnen Teilungsmethoden mit der Gruppierung aller Objekte, was bedeutet, dass sich zunächst alle *n* Objekte in einer einzigen Gruppe befinden. Bei jedem weiteren Schritt werden neue Gruppen erstellt, bis es welche gibt
n Gruppen.

- Geometrische Klassifizierung, bei der die Objekte in einem **Streudiagramm** dargestellt und dann nach ihrer Position im Diagramm gruppiert werden. Bei einer **grafischen Darstellung** entsprechen die Nähen der Objekte zueinander in der Grafik den Ähnlichkeiten zwischen den Objekten.

Die ersten drei Klassifizierungsarten werden allgemein unter dem Begriff **Clusteranalyse zusammengefasst.** Gemeinsam ist ihnen, dass die zu klassifizierenden Objekte eine gewisse Struktur aufweisen müssen, die es uns ermöglicht, den Grad der Ähnlichkeit zwischen den Objekten zu messen.

Jeder Klassifizierungstyp enthält eine Vielzahl von Methoden, mit denen wir Klassen erstellen können von ähnlichen Objekten.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Klassifizierung kann in zwei Fällen verwendet werden:

- Beschreibungsfälle; •
Vorhersagefälle.

Im ersten Fall erfolgt die Klassifizierung anhand einiger allgemein anerkannter Standardmerkmale. Berufe lassen sich beispielsweise in freiberufliche Berufe einteilen. Manager, Arbeiter usw., und man kann für jede Klasse die Durchschnittsgehälter, die durchschnittliche Häufigkeit von Gesundheitsproblemen usw. berechnen.

Im zweiten Fall führt die Klassifizierung zu einer Vorhersage und dann zu einer Aktion. Wenn beispielsweise die Füchse in einer bestimmten Region apathisches Verhalten und übermäßigen Speichelfluss zeigen, können wir auf eine neue Tollwut schließen
Epidemie. Dies sollte dann eine Impfkampagne auslösen.

WEITERLESEN

Clusteranalyse
Vollständige Verknüpfungsmethode
Datenanalyse

REFERENZEN

Everitt, BS: Clusteranalyse. Halstead, London (1974)

Gordon, AD: Klassifizierung. Methoden zur explorativen Analyse multivariater Daten. Chapman & Hall, London (1981)

Kaufman, L., Rousseeuw, PJ: Gruppen in Daten finden: Eine Einführung in die Clusteranalyse. Wiley, New York (1990)

Thorndike, RL: Wer gehört in eine Familie? Psychometrika **18,** 267–276 (1953)

Tryon, RC: Clusteranalyse. McGraw Hill, New York (1939)

Zubin, J.: Eine Technik zur Messung von Gleichgesinntheit. J. Abnorm. Sozialpsychol. **33,** 508–516 (1938)

Clusteranalyse

Beim Clustering handelt es sich um die Aufteilung eines Datensatzes in Teilmengen oder Cluster, sodass der Grad der Assoziation zwischen Mitgliedern desselben Clusters stark und zwischen Mitgliedern desselben schwach ist

Machine Translated by Google

verschiedene Cluster entsprechend einem definierten Abstandsmaß.

Es gibt mehrere Methoden zur Durchführung **einer Clusteranalyse :**

• Partitionelles Clustering • Hierarchisches Clustering.

GESCHICHTE

Siehe **Klassifizierung** und **Datenanalyse.**

MATHEMATISCHE ASPEKTE

Um eine **Clusteranalyse** für eine Menge von *n* Objekten durchzuführen , müssen wir einen **Abstand** zwischen den Objekten (oder allgemeiner ein Maß für die Ähnlichkeit zwischen den Objekten) definieren, die klassifiziert werden müssen. Die Existenz irgendeiner Struktur innerhalb der

Es wird eine Menge von Objekten angenommen.

Um eine hierarchische Klassifizierung einer Menge *E* von Objekten {x1, *x2,...,xn} durchzuführen,* ist es notwendig, einen mit *E* verknüpften **Abstand** zu definieren, der zum Erhalten einer **Abstandstabelle** verwendet werden kann zwischen den Objekten von E. Ebenso muss auch für alle Teilmengen von E ein Abstand definiert werden.

Ein Ansatz zur hierarchischen Clusterbildung ist die Verwendung der Agglomerationsmethode. Es kann im folgenden Algorithmus zusammengefasst werden: 1. Suchen Sie das Paar von Objekten *(xi, xj) ,* die jeweils den kleinsten **Abstand** haben andere.

2. Aggregieren Sie das Objektpaar *(xi, xj)* zu einem einzelnen Element ÿ und erstellen Sie eine neue **Distanztabelle neu.** Dies wird erreicht, indem die mit *xi* und *xj* verknüpften Zeilen und Spalten unterdrückt und durch eine Zeile und eine mit ÿ verknüpfte Spalte ersetzt werden. Der Die neue Distanztabelle wird eine Linie und haben eine Spalte weniger als die vorherige Tabelle.
3. Wiederholen Sie diese beiden Vorgänge, bis die gewünschte Anzahl an Klassen erreicht ist oder bis alle Objekte in derselben Klasse zusammengefasst sind.

Beachten Sie, dass der Abstand zwischen der aus aggregierten Elementen gebildeten Gruppe und den anderen Elementen unterschiedlich definiert werden kann Wege, die zu unterschiedlichen Methoden führen.

Beispiele hierfür sind die Single-Link-Methode und die **Complete-Linkage-Methode.**

Die Single-Link-Methode ist eine hierarchische Klassifizierungsmethode, die das Euklidische verwendet **Abstand ,** um eine **Abstandstabelle zu erstellen,** und der Abstand zwischen zwei Klassen wird durch den euklidischen Abstand zwischen den beiden Clos angegeben est-Elemente (der Mindestabstand).

Bei der **vollständigen Verknüpfungsmethode** ist der **Abstand** zwischen zwei Klassen durch den euklidischen Abstand zwischen den beiden Elementen gegeben am weitesten entfernt (maximaler Abstand).

Da der einzige Unterschied zwischen diesen beiden Methoden der Abstand zwischen zwei ist Klassen ist entweder das Minimum und das Maximum Aufgrund der maximalen Entfernung wird hier nur die Single-Link-Methode berücksichtigt.

Für eine Menge *E* = {X1, *X2,...,Xn}* ergibt sich dann die **Distanztabelle** für die Elemente von *E* gern.

Da diese Tabelle symmetrisch und entlang ihrer Diagonale null ist, wird nur eine Hälfte der Tabelle berücksichtigt:

d(X1, X2) *d(X1,* X3) ... *d(X1, Xn) d(X2,* X3) ... *d(X2, Xn)*

... ...

d(Xnÿ1, Xn)

wobei *d(Xi, Xj)* der euklidische Abstand zwischen *Xi* und *Xj* für i < j ist, wobei die **Werte** von *i* und *j* zwischen 1 und n liegen.

Der **Algorithmus** für die Single-Link-Methode lautet wie folgt:

• Suchen Sie nach dem Minimum *d(Xi, Xj)* für *i* < j;

Machine Translated by Google

- Die Elemente Xi und Xj werden zu aggregiert
  eine neue Gruppe *Ck = Xi ÿ Xj;*
- Die Menge *E* wird dann partitioniert

  {X1},...,{Xiÿ1},{Xi, Xj},{Xi+1},...,

  {Xjÿ1},{Xj+1},...,{Xn} ;

- Anschließend wird die **Distanztabelle** neu erstellt
  ohne die Zeilen- und Spaltenzuordnungen

  ated mit *Xi* und *Xj,* und mit einer Linie
  und eine Spalte, die die **Abstände** darstellt
  zwischen *Xm* und *Ck, m = 1, 2,...,nm =*
  i und *m* = j, gegeben durch:

  d(Ck, Xm) = min{d(Xi, Xm); *d(Xj,* Xm)}.

Der Algorithmus wird wiederholt, bis der gewünschte Wert erreicht ist
Anzahl der Gruppen erreicht ist bzw. bis dahin
ist nur eine Gruppe, die alle Elemente enthält
gen.

Im **allgemeinen Fall der Abstand zwischen zwei**
Gruppen ist gegeben durch:

d(Ck, Cm) = min{d(Xi, Xj) mit *Xi*
zu *Ck* und *Xj* zu *Cm gehören}* ,

Es gilt die zuvor angegebene Formel
Besonderer Fall, wenn die Gruppen jeweils aus zwei
Elementen und einem bestehen
einzelnes Element.
Diese Reihe von Ansammlungen kann durch ein
Dendrogramm dargestellt werden , wobei die Abszisse
sa den Abstand zwischen den Objekten angibt.
Beachten Sie, dass wir mehr als ein Paar finden könnten
wenn wir nach dem Paar der nächsten Elemente
suchen. In diesem Fall gilt dies nicht für das Paar, das
im ersten Schritt zur Aggregation ausgewählt wurde
beeinflussen spätere Schritte (vorausgesetzt der Algorithmus
endet in diesem Schritt nicht), da das andere Paar der
nächstgelegenen Elemente im folgenden Schritt
aggregiert wird. Die Aggregationsreihenfolge wird im
Dendrogramm nicht angezeigt
weil es die Entfernung anzeigt, die trennt
zwei gruppierte Objekte.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Wahl des **Abstandes** zwischen den
Gruppe aus aggregierten Elementen und
Die anderen Elemente können in mehreren betrieben werden
Art und Weise, je nach der verwendeten Methode,
wie zum Beispiel bei der Single-Link-Methode und
in der **vollständigen Verknüpfungsmethode.**

BEISPIELE

Lassen Sie uns veranschaulichen, wie die **Single-Link-Methode** funktioniert
der Clusteranalyse können auf die angewendet werden
Prüfungsnoten von fünf Studierenden
jeder studiert vier Kurse: Englisch, Französisch,
Mathematik und Physik.
Wir wollen diese fünf Schüler in zwei teilen
Gruppen mithilfe der Single-Link-Methode.
Die in den Prüfungen erzielten Noten,
die von 1 bis 6 reichen, sind in zusammengefasst
die folgende Tabelle:

	Englisch	Französisch	Mathematik	Physi
Alain	5.0	3.5	4,0	4.5
Jean	5.5	4,0	5,0	4.5
März	4.5	4.5	4,0	3.5
Paul	4.0	5.5	3.5	4,0
Pierre	4.0	4.5	3,0	3.5

Anschließend berechnen wir die euklidischen **Abstände**
zwischen den Schülern und nutzen sie zum Gestalten
eine **Distanztabelle:**

	Alain	Jean	Marc	Paul	Pierre
Alain	0	1,22	1,5	2,35	2
Jean	1,22	0	1.8	2,65	2,74
März	1.5	1,8	0	1,32	1,12
Paulus	2,35	2,65	1,32	0	1.22
Pierre	2	2,74	1,12	1,22	0

Indem ich nur den oberen Teil davon betrachte
symmetrische Tabelle, wir erhalten:

Machine Translated by Google

	Jean Marc Paul			Pierre
Alain	1.22	1.5	2,35	2
Jean		1.8	2,65	2,74
Marc			1,32	1.12
Paul				1.22

Der Mindestabstand **beträgt** 1,12
Marc und Pierre; wir bilden also den ersten
Gruppe aus diesen beiden Studenten. Anschließend
berechnen wir die neuen Abstände.

Beispielsweise berechnen wir auf der einen Seite
den neuen Abstand zwischen Marc und Pierre
und Alain andererseits, indem der Mindestabstand
zwischen Marc und Alain eingehalten wird und
der Mindestabstand zwischen Pierre und
Alain:

d({Marc,Pierre},Alain)
= min{d(Marc,Alain);d(Pierre,Alain)}
= min{1,5; 2} = *1,5* ,

Auch

d({Marc,Pierre},Jean)
= min{d(Marc,Jean);d(Pierre,Jean)}
= min{1,8; 2,74} = *1,8* ,

Und

d({Marc,Pierre},Paul)
= min{d(Marc,Paul);d(Pierre,Paul)}
= min{1,32; 1,22} = 1,22 .

Die neue **Distanztabelle** sieht wie folgt aus
form:

	Jean Marc und Paul		
	Pierre		
Alain	1,22 1,5		2,35
Jean		1.8	2,65
Marc und Pierre			1.22

Der Mindestabstand beträgt jetzt 1,22
Alain und Jean und auch zwischen der Gruppe
von Marc und Pierre einerseits und Paul
auf der anderen Seite (also zwei Paare).
den Mindestabstand aufweisen); lasst uns wählen
um zunächst Alain und Jean neu zu gruppieren. Das andere
Das Paar wird im nächsten Schritt aggregiert. Wir
Erstellen Sie die **Distanztabelle** neu und erhalten Sie:

d({Alain,Jean}, {Marc,Pierre})
= min{d(Alain,{Marc,Pierre}) ,
d(Jean,{Marc,Pierre})}
= min{1,5; 1,8} = *1,5*

ebenso gut wie:

d({Alain,Jean},Paul)
= min{d(Alain,Paul);d(Jean,Paul)}
= min{2,35; 2,65} = 2,35 .

Daraus ergibt sich die folgende **Distanztabelle:**

	Marc und Pierre Paul
Alain und Jean 1.5	2,35
Marc und Pierre	1.22

Beachten Sie, dass Paulus jetzt integriert werden muss
Die Gruppe bestand aus Marc und Pierre und
Der neue **Abstand** beträgt:

d({(Marc,Pierre),Paul},{Alain,Jean})
= min{d({Marc,Pierre},{Alain,Jean}),
d(Paul,{Alain,Jean})}
= min{1,5; 2,35} = *1,5*

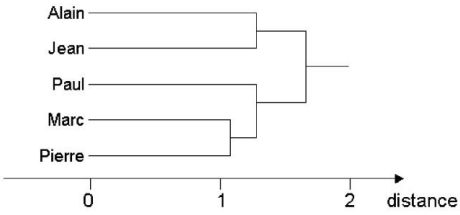
was die folgende **Distanztabelle ergibt:**

	Alain und Jean
Marc, Pierre und Paul	1.5

Wir erhalten schließlich zwei Gruppen:
{Alain und Jean} und {Marc, Pierre und
Paul}, die einen **Abstand** von haben
1.5.

Machine Translated by Google

Das folgende **Dendrogramm** veranschaulicht dies aufeinanderfolgende Aggregationen:



WEITERLESEN

Einstufung

Vollständige Verknüpfungsmethode

Dendrogramm

Distanz

Entfernungstabelle

VERWEISE

Celeux, G., Diday, E., Govaert, G., Lecheval lier, Y., Ralambondrainy, H.: Classifi cation automatique des données – Aspekte Statistik und Informatik. Dunod, Paris (1989)

Everitt, BS: Clusteranalyse. Halstead, London (1974)

Gordon, AD: Klassifizierung. Methoden für die explorative Analyse multivariater Methoden Daten. Chapman & Hall, London (1981)

Jambu, M., Lebeaux, MO: Klassifizierung Automatische Analyse der Données. Dunod, Paris (1978)

Kaufman, L., Rousseeuw, PJ: Finden Gruppen in Daten: Eine Einführung in die Clusteranalyse. Wiley, New York (1990)

Lerman, LC: Klassifizierung und ordnungsmäßige Analyse der Données. Dunod, Paris (1981)

Tomasson, R., Daudin, JJ, Danzart, M., Masson, JP: Diskriminierung und Klassifizierung. Masson, Paris (1988)

Cluster-Sampling

Beim **Clustersampling** besteht der erste Schritt in der Division Die **Bevölkerung** wird in Teilmengen, sogenannte Cluster, unterteilt. Jeder Cluster besteht aus Individuen, die es sind soll repräsentativ für die Bevölkerung sein.

Beim Cluster- **Sampling** geht es dann um die Auswahl eine **Zufallsstichprobe** von Clustern und dann die Beobachtung aller zu jedem Cluster gehörenden Individuen von ihnen.

GESCHICHTE

Siehe **Probenahme.**

MATHEMATISCHE ASPEKTE

Beim Cluster-Sampling handelt es sich um den Prozess der zufälligen Extraktion repräsentativer Mengen (bekannt als Cluster) aus einer größeren Population von Einheiten und dann einen Fragebogen auf alle anwenden die Einheiten in den Clustern. Die Cluster oft bestehen aus geografischen Einheiten, etwa Stadtbezirken. In diesem Fall besteht die Methode darin, eine Stadt in Bezirke zu unterteilen und diese dann auszuwählen Bezirke, die in die Stichprobe einbezogen werden sollen. Endlich, alle Personen oder Haushalte im Auserwählten Bezirk werden befragt.

Es gibt zwei Hauptgründe für eine Leistung Cluster-Sampling. In vielen Anfragen gibt es keine vollständige und verlässliche Liste der Bevölkerungseinheiten, die als Grundlage für die Stichprobe dienen könnten Oder es kann sein, dass die Erstellung einer solchen Liste zu teuer ist. Beispielsweise ist es in vielen Ländern, auch in den Industrieländern, selten vollständige und aktuelle Listen aller zu haben die Mitglieder der Bevölkerung, Haushalte oder ländliche Anwesen. In dieser Situation Probenahme kann auf geografische Weise erreicht werden: Jede Stadtregion ist in Stadtteile unterteilt und jede ländliche Region in Landgüter. Der Bezirke und die landwirtschaftlich genutzten Flächen werden als Cluster betrachtet und wir nutzen die Kom-

Machine Translated by Google

vollständige Liste der Cluster, da wir keine haben
eine vollständige und aktuelle Liste aller
Bevölkerungseinheiten. Daher stichproben wir eine
erforderliche Anzahl von Clustern aus der Liste und dann
Befragen Sie alle Einheiten im ausgewählten Cluster
ter.

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Vorteil des Cluster- **Samplings** besteht darin, dass es
Es ist nicht erforderlich, über eine vollständige und aktuelle
Liste aller Bevölkerungseinheiten zu **verfügen**
Analyse durchzuführen.
In vielen Ländern gibt es beispielsweise keine
aktualisierte Personen- oder Wohnungslisten. Die Kosten
Die Erstellung solcher Listen ist oft unerschwinglich.
Es ist daher einfacher, Teilmengen davon zu analysieren
Bevölkerung (bekannt als Cluster).
Im Allgemeinen liefert Cluster- **Sampling Schätzungen ,**
die nicht so präzise sind wie **einfache**
Zufallsstichprobe, aber dieser Rückgang in der Genauigkeit
wird leicht durch die weitaus geringeren Kosten des Clusters ausgeglichen
Probenahme.
Um **Clustersampling** möglichst effizient durchzuführen:

- Die Cluster sollten nicht zu groß sein
  Es sollte eine ausreichend große Anzahl vorhanden sein
  Cluster,
- Clustergrößen sollten so einheitlich wie möglich sein
  möglich;
- Die zu jedem Cluster gehörenden Individuen müssen
  möglichst heterogen sein
  in Bezug auf den beobachteten Parameter.

Ein weiterer Grund für die Verwendung von Cluster-Sampling ist
kosten. Auch wenn eine vollständige und aktuelle
Möglicherweise existiert keine Liste aller Bevölkerungseinheiten
Verwenden Sie vorzugsweise Cluster-Sampling aus einem
Aus wirtschaftlicher Sicht ist dies schneller, erfordert
weniger Arbeitskräfte und minimiert die Transportkosten.
Es ist also mehr
angemessen, Cluster-Sampling zu verwenden, wenn die
Das dadurch eingesparte Geld ist weitaus bedeutsamer
icant als die Zunahme der Stichprobenvarianz
das wird sich ergeben.

BEISPIELE

Betrachten Sie N, die Größe der **Bevölkerung** von
Stadt X. Wir wollen die Verteilung von untersuchen
„Alter“ für Stadt *X* ohne Durchführung einer Volkszählung.
Die Bevölkerung ist in *G*- Teile unterteilt .
Es wird **eine einfache Zufallsstichprobe** durchgeführt
Unter diesen *G*- Teilen erhalten wir *g*- Teile.
Die **endgültige Probe** wird aus allen zusammengesetzt sein
Einzelpersonen in den *g* ausgewählten Teilen.

WEITERLESEN

- **Probenahme**
- **Einfache Zufallsstichprobe**

VERWEISE

Hansen, MH, Hurwitz, WN, Madow,
MG: Beispielbefragungsmethoden und die Theorie. Bd.
I. Methoden und Anwendungen.
Bd. II. Theorie. Chapman & Hall, London (1953)

Koeffizient der Entschlossenheit

Das Bestimmtheitsmaß wird bezeichnet
R2 ist der Quotient der erklärten Variation (Summe der
Quadrate aufgrund der Regression) zur
Gesamtvariation (Gesamtsumme der Quadrate Gesamt-SS
(TSS)) in einem **Modell** der einfachen oder **multiplen**
linearen Regression:

$$R2 = \frac{\text{Erklärte Variation}}{\text{Totale Variation}}.$$

Er entspricht dem Quadrat des **Korrelationskoeffizienten**
und kann Werte zwischen 0 und 1 annehmen.
Sie wird oft in **Prozent ausgedrückt.**

GESCHICHTE

Siehe **Korrelationskoeffizient.**

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Betrachten Sie das folgende **Modell** für die multiple Regression:

$$Yi = ÿ0 + ÿ1Xi1 + \cdots + ÿpXip + ÿi$$

für $i = 1,...,n,$ wo

Yi sind *die* abhängigen *Variablen ,*

_ _ _ _ _ _ _ _ _ _ *.,p)* sind die zu schätzenden Parameter.

Die Schätzung erfolgt durch Schätzung der Parameter ÿ0, *ÿ1,..., ÿp*

$$Yˆi = ÿˆ 0 + ÿˆ 1Xi1 + \cdots + ÿˆ pXip .$$

Das Bestimmtheitsmaß ermöglicht es uns

Messen Sie die Qualität der Anpassung der Regressionsgleichung an die Messwerte.

Um die Qualität der Anpassung der **Regressionsgleichung** zu bestimmen , berücksichtigen Sie die Lücke zwischen dem beobachteten **Wert** und dem Schätzwert

Vergleichswert für jede **Beobachtung** des **Probe.** Diese Lücke (oder **Residuum)** kann auch folgendermaßen ausgedrückt werden:

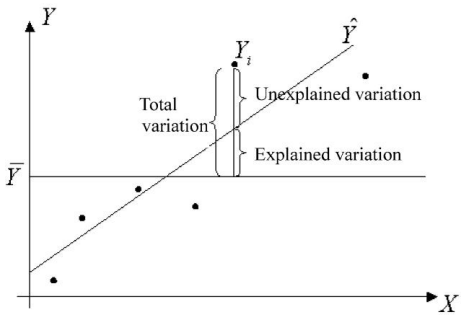
$$\sum_{i=1}^{N} (Yi - Y^-)^2 = \sum_{i=1}^{N} (Yi - Yˆi)^2 + n \sum_{i=1} (Yˆi - Y^-)^2$$

$$TSS = RSS + REGSS$$

Wo

TSS ist die Gesamtsumme der Quadrate, **RSS** die Restquadratsumme und **REGSS** die Summe der Quadrate der Regression.

Diese Konzepte und die Beziehungen zwischen ihnen werden in der folgenden Grafik dargestellt:



Mit diesen Konzepten können wir R2 definieren, Das ist der Bestimmungskoeffizient. Es misst den Variationsanteil der **Variablen** Y, der durch die **Regressionsgleichung** wie folgt beschrieben wird :

$$R2 = \frac{REGSS}{TSS} = \frac{\sum_{i=1}^{N} (Yˆi - Y^-)^2}{\sum_{i=1}^{N} (Yi - Y^-)^2}$$

Wenn die **Regressionsfunktion** verwendet werden soll, um Vorhersagen über nachfolgende Beobachtungen zu treffen, ist ein hoher **Wert** von R2 vorzuziehen , denn je höher der Wert von R2, desto kleiner ist die unerklärte Variation.

BEISPIELE

Die folgende Tabelle enthält Werte für das Bruttosozialprodukt (BSP) und die Nachfrage nach inländischen Produkten für den Zeitraum 1969–1980 für ein bestimmtes Land.

Jahr	BSP X	Nachfrage nach Dom Estische Produkte Y
1969	50	6
1970	52	8
1971	55	9
1972	59	10

Machine Translated by Google

Jahr BSP	X	Nachfrage nach Dom Estische Produkte Y
1973	57	8
1974	58	10
1975	62	12
1976	65	9
1977	68	11
1978	69	10
1979	70	11
1980	72	14

Wir werden versuchen, die Nachfrage im Kleinen abzuschätzen
Güter als Funktion des BSP gemäß
Modell

$$Yi = a + b \cdot Xi + \ddot{y}i\,, \qquad ich = 1,..., 12 \quad .$$

Schätzen der **Parameter** *a* und *b* durch
Die Methode **der kleinsten Quadrate** ergibt Folgendes
Schätzer:

$$b\hat{} = \frac{\sum_{i=1}^{12} (Xi \ddot{y} X^{-})(Yi \ddot{y} Y^{-})}{\sum_{i=1}^{12} (Xi \ddot{y} X^{-})^{2}} = 0{,}226 \quad ,$$

$$a\hat{} = Y^{-} \ddot{y} b\hat{} \cdot X^{-} = \ddot{y}4.047 \; .$$

Die geschätzte Zeile wird geschrieben

$$Y\hat{} = \ddot{y}4{,}047 + 0{,}226 \cdot X \qquad .$$

Die Qualität der Passung der gemessenen Punkte
zur Regressionsgeraden ergibt sich aus dem
Determinationskoeffizienten:

$$R2 = \frac{REGSS}{TSS} = \frac{\sum_{i=1}^{12} (Y\hat{}i \ddot{y} Y^{-})^{2}}{\sum_{i=1}^{12} (Yi \ddot{y} Y^{-})^{2}} \; .$$

Wir können den **Mittelwert** berechnen mit:

$$Y^{-} = \frac{\sum_{i=1}^{12} Yi}{N} = 9{,}833 \; .$$

Yi	Y_i	$(Yi \; Y_2)$	$(Y_i \; Y)^2$
6	7.260	14.694	6.622
8	7.712	3.361	4.500
9	8.390	0,694	2.083
10	9.294	0,028	0,291
8	8.842	3.361	0,983
10	9.068	0,028	0,586
12	9.972	4.694	0,019
9	10.650	0,694	0,667
11	11.328	1.361	2.234
10	11.554	0,028	2.961
11	11.780	1.361	3.789
14	12.232	17.361	4.754
Gesamt		47.667	30.489

Wir erhalten daher:

$$R2 = = \frac{30.489}{47.667} = 0{,}6396$$

oder, in Prozent:

$$R2 = 63{,}96\ \% \; .$$

Daraus können wir schließen, dass gem
Das gewählte **Modell** beträgt 63,96 % der Variation
Die Nachfrage nach Kleingütern wird erklärt durch
die Variation des BSP.

Offensichtlich darf der **Wert** von R2 nicht überschritten werden
100%. Obwohl 63,96 % relativ hoch sind, ist dies der Fall
nicht nah genug an 100 %, um einen Versuch auszuschließen
um das **Modell** weiter zu modifizieren.
Diese Analyse zeigt auch, dass andere Variablen
außer dem Bruttosozialprodukt berücksichtigt werden
Berücksichtigung bei der Bestimmung der Funktion
entsprechend der Nachfrage nach Kleingütern,
da das BSP das nur teilweise erklärt
Variation.

Machine Translated by Google

WEITERLESEN

Korrelationskoeffizient

Multiple lineare Regression

Regressionsanalyse

Einfache lineare Regression

Kurtosis-Koeffizient

Zur Messung wird der Kurtosis-Koeffizient verwendet

Stellen Sie sicher, dass eine Kurve spitze oder flach ist. Es basiert auf den **Momenten** der Verteilung.

Dieser Koeffizient ist eines der **Maße von Kurtosis.**

GESCHICHTE

Siehe **Schiefekoeffizient.**

MATHEMATISCHE ASPEKTE

Der Kurtosis-Koeffizient (ÿ2) basiert auf dem zentrierten **Moment** vierter Ordnung von eine **Verteilung ,** die gleich ist:

$$ÿ4 = E\,(X\ ÿ\ ÿ)4\ .$$

Um einen Kurtosis-Koeffizienten zu erhalten das von den Maßeinheiten unabhängige ist, wird das **Moment** vierter Ordnung geteilt durch die **Standardabweichung** der **Grundgesamtheit** ÿ, erhöht auf die vierte Potenz. Der Kurtosis-Koeffizient beträgt dann:

$$ÿ2 = \frac{ÿ4}{ÿ4}\ .$$

Für eine **Stichprobe** (x1, *x2,...,xn),* der **Schätzer** dieses Koeffizienten wird mit b2 bezeichnet. Es ist gleich Zu:

$$b2 = \frac{m4}{S4}\ ,$$

wobei m4 die zentrierte vierte Ordnung ist **Zeitpunkt** der Probe, gegeben durch:

$$m4 = \frac{1}{N}\cdot \sum_{i=1}^{N} (xi\ ÿ\ \bar{x})^4\ ,$$

Dabei ist $\bar{x}$ das **arithmetische Mittel und** *n* die Summe Anzahl der **Beobachtungen** und S4 ist die **Standardabweichung** der Stichprobe vierte Potenz.

Für den Fall, dass eine **zufällige Vari fähiges** *X* nimmt **Werte** *xi* mit **Frequenzen** an *fi, i = 1, 2,...,h,* die zentrierte vierte Ordnung **Der Zeitpunkt** der **Probe** ergibt sich aus der Formel:

$$m4 = \frac{1}{N}\cdot \sum_{i=1}^{H} fi \cdot (xi\ ÿ\ \bar{x})^4\ .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Für **eine Normalverteilung beträgt der** Koeffizient von Kurtosis ist gleich 3. Daher wird eine Kurve als platikurtisch bezeichnet werden (was flacher als der bedeutet). Normalverteilung), wenn der Kurtosis-Koeffizient kleiner als 3 ist. Es ist leptokurtisch (d. h. schärfer als die Normalverteilung), wenn ÿ2 größer als 3 ist.

Lassen Sie uns nun beweisen, dass der Kurto-Sisis-Koeffizient für die **Normalverteilung** gleich 3 ist . Wir wissen, dass $ÿ2 = \frac{ÿ4}{ÿ4}$, was bedeutet, dass die zentrierter **Moment** vierter Ordnung wird geteilt durch die **Standardabweichung** auf den vierten erhöht Leistung. Es kann bewiesen werden, dass das zentrierte etw Das Ordnungsmoment, bezeichnet als *ÿs,* erfüllt die folgende Beziehung für eine **Normalverteilung:**

$$ÿs = (s\ ÿ\ 1)ÿ2 \cdot ÿsÿ2\ .$$

Diese Formel ist eine rekursive Formel, die drückt **Momente** höherer Ordnung als Funktion von Momenten niedrigerer Ordnung aus.

Vorausgesetzt, dass ÿ0 = 1 (das **Moment** nullter Ordnung einer **beliebigen Zufallsvariablen** ist gleich 1, da es ist der erwartete **Wert** dieser **Variablen** zur Potenz Null) und dass ÿ1 = 0 (das zentrierte **Moment** erster Ordnung ist für jede Zufallsvariable Null), ergibt sich:

$$ÿ2 = ÿ2$$

$$ÿ3 = 0$$

Machine Translated by Google

$$ÿ4 = 3ÿ4$$

usw. .

Der Kurtosis-Koeffizient ist dann gleich:

$$ÿ2 = \frac{ÿ4}{ÿ4} = \frac{3ÿ4}{ÿ4} = 3 .$$

BEISPIELE

Wir wollen die Kurtosis der Verteilung berechnen

Beitrag zum Tagesumsatz von 75 Bäckereien. Lassen Wir berechnen den Kurtosis-Koeffizienten ÿ2

unter Verwendung folgender **Daten:**

Tabelle zur Kategorisierung der Tagesumsätze von 75 Bäckereien

Umsatzfrequenzen	
215–235	4
235–255	6
255–275	13
275–295	22
295–315	15
315–335	6
335–355	5
355–375	4

Das **Moment** vierter Ordnung der **Stichprobe** ist

gegeben von:

$$m4 = \frac{1}{N} \sum_{i=1}^{H} fi(xi ÿ \bar{x})^4 ,$$

wobei $n = 75$, $\bar{x} = 290{,}60$ und xi der Mittelpunkt des **Klassenintervalls** i ist. Wir können das zusammenfassen

Berechnungen in der folgenden Tabelle:

xi	xi x	fi	fi xi x)4
225	ÿ65,60	4	74075629.16
245	ÿ45,60	6	25942428.06
265	ÿ25,60	13	5583457.48
285	ÿ5,60	22	21635,89
305	14.40	15	644972.54
325	34.40	6	8402045.34
345	54,40	5	43789058.05
365	74,40	4	122560841.32
			281020067,84

Da $S = 33{,}88$, der Kurtosis-Koeffizient

ist gleich:

$$ÿ2 = \frac{\frac{1}{75}(281020067.84)}{(33.88)4} = 2{,}84 .$$

Da ÿ2 kleiner als 3 ist, können wir daraus schließen dass die Verteilung des Tagesumsatzes in

75 Bäckereien sind platikurtisch, was bedeutet, dass sie es sind flacher als die **Normalverteilung.**

WEITERLESEN

- **Maß der Kurtosis**
- **Formmaß**

VERWEISE

Siehe **Schiefekoeffizient** ÿ1 **de Pearson.**

Schiefekoeffizient

Der Schiefekoeffizient misst die Schiefe einer Verteilung. Es basiert auf der Vorstellung vom **Zeitpunkt** der Verteilung. Dieser Koeffizient ist eines der **Maße von Schiefe.**

GESCHICHTE

Zwischen dem Ende des neunzehnten Jahrhunderts

Zu Beginn des 20. Jahrhunderts untersuchte **Pearson und Karl** große Mengen von **Daten ,** die zum Teil deutlich von der Normalität abwichen und eine erhebliche Schiefe aufwiesen.

Er verwendete zunächst den folgenden Koeffizienten als ein **Maß für die Schiefe:**

$$\text{Schiefe} = \frac{\bar{x} ÿ \text{Modus}}{S} ,$$

wobei $\bar{x}$ **das arithmetische Mittel** darstellt und S ist die **Standardabweichung.**

Dieses Maß ist gleich Null, wenn die Daten vorhanden sind symmetrisch verteilt.

Machine Translated by Google

Er entdeckte empirisch, dass für eine mäßig asymmetrische Verteilung (das Gamma Verteilung):

$$\text{Mo} \ddot{y} \bar{x} \ddot{y} 3 \cdot (Md \ddot{y} \bar{x}) ,$$

wobei Mo und Md den Modus bezeichnen und der Median des Datensatzes. Indem man dies ersetzt Ausdruck in den vorherigen Koeffizienten, den erhält man folgende Alternativformel:

$$\text{Schiefe} = \frac{3 \cdot (\bar{x} \ddot{y} \text{Md})}{S} .$$

Im Anschluss daran: **Pearson, K.** (1894, 1895) führte einen Schiefekoeffizienten ein, bekannt als ÿ1- Koeffizient, basierend auf Berechnungen der zentrierten **Momente.** Dieser Koeffizient ist schwieriger zu berechnen, aber es ist mehr beschreibender und besser an eine große Anzahl von **Beobachtungen angepasst.**

Pearson, K. hat auch den **Koeffizienten von** erstellt **Kurtosis** (ÿ2), die zur Messung verwendet wird Abgeflachtheit einer Kurve. Dieser Koeffizient ist auch basierend auf den **Momenten** der Verteilung studiert werden.

Tabellen mit den Grenzwerten der Koeffizienten ÿ1 und ÿ2 finden Sie in den Werken von Pearson und Hartley (1966, 1972). Wenn die Die Stichprobe schätzt einen Unterschreitungsgrenzwert für ÿ1, ÿ2, wir schließen daraus, dass die Bevölkerung ist deutlich gebogen oder schief.

MATHEMATISCHE ASPEKTE

Der Schiefekoeffizient basiert auf dem cen tertes **Moment** dritter Ordnung der Verteilung in Frage, was gleich ist:

$$\ddot{y}3 = E(X \ddot{y} \ddot{y})3 .$$

Um einen Schiefekoeffizienten zu erhalten Unabhängig von der Maßeinheit wird das **Moment** dritter Ordnung durch das **Normal** dividiert **Abweichung** der **Grundgesamtheit** ÿ zum dritte Macht. Der erhaltene Koeffizient, bezeichnet mit $\overline{\ddot{y}\ddot{y}1}$, ist gleich:

$$\overline{\ddot{y}1} = \frac{\ddot{y}3}{\ddot{y}3} .$$

Der **Schätzer** dieses Koeffizienten, berechnet für eine **Stichprobe** (x1, *x2,...xn),* wird mit bezeichnet $\overline{\ddot{y}b1}$. Es ist gleich:

$$\overline{b1} = \frac{m3}{S3} ,$$

wobei m3 die zentrierte dritte Ordnung ist **Zeitpunkt** der Probe, gegeben durch:

$$m3 = \frac{1}{N} \cdot \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^3 .$$

Dabei ist $\bar{x}$ das **arithmetische Mittel,** *n* die Summe Anzahl der **Beobachtungen** und S3 ist die **Standardabweichung** erhöht auf die dritte Potenz. Für den Fall, dass eine **Zufallsvariable** *X* nimmt **Werte** *xi* mit **Frequenzen** *fi, i* = an *1, 2,...,h,* das zentrierte **Moment** dritter Ordnung der **Probe** ergibt sich aus der Formel:

$$m3 = \frac{1}{N} \cdot \sum_{i=1}^{H} fi \cdot (xi \ddot{y} \bar{x})^3 .$$

Wenn der Koeffizient positiv ist, ist die Verteilung breitet sich nach rechts aus. Wenn es negativ ist, wird die Die Verteilung erstreckt sich nach links. Ifitis closeto Null, die Verteilung ist annähernd symmetrisch.

Wenn die **Stichprobe** einer **Normalbevölkerung** entnommen wird , ergibt sich in etwa die **Statistik** $\overline{\ddot{y}b1}$ eine **Normalverteilung** mit einem **Mittelwert** von 0 und eine **Standardabweichung** von $\sqrt{\frac{6}{N}}$. Wenn die **Größe von** Ist **die Stichprobe** *n* größer als 150, kann die **Normaltabelle** zum Testen der Schiefe verwendet werden **Hypothese.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Dieser Koeffizient ist (ähnlich wie die anderen **Schiefemaße)** nur dann von Interesse, wenn er vorhanden ist

Machine Translated by Google

kann verwendet werden, um die Formen von zwei oder zu vergleichen weitere Distributionen.

BEISPIELE

Angenommen, wir möchten die Formen vergleichen der ermittelten Tagesumsatzverteilungen für 75 Bäckereien für zwei verschiedene Jahre. Wir Berechnen Sie dann den Schiefekoeffizienten in beide Fälle.

Die **Daten** sind in der folgenden Tabelle kategorisiert:

Umsatzfrequenzen	Frequenzen für Jahr 1	Frequenzen für Jahr 2
215–235	4	25
235–255	6	15
255–275	13	9
275–295	22	8
295–315	15	6
315–335	6	5
335–355	5	4
355–375	4	3

Das **Moment** dritter Ordnung der **Stichprobe** ist gegeben von:

$$m3 = \frac{1}{N} \cdot \sum_{i=1}^{H} fi \cdot (xi - \bar{x})^3 .$$

Für Jahr 1 ist $n = 75$, $\bar{x} = 290{,}60$ und xi ist Mittelpunkt jedes Klassenintervalls **i .** Die Kalkulation

Die Funktionen sind in der folgenden Tabelle zusammengefasst:

xi	xi x	fi	fi xi x 3
225	-65,60	4	-1129201.664
245	-45,60	6	-568912,896
265	-25,60	13	-218103.808
285	-5,60	22	-3863,652
305	14.40	15	44789.760
325	34.40	6	244245.504
345	54,40	5	804945.920
365	74,40	4	1647323.136
			821222.410

Da $S = 33{,}88$, der Schiefekoeffizient ist gleich:

$$\overline{b1} = \frac{\frac{1}{75}(821222.41)}{(33.88)3} = \frac{10949.632}{38889.307}$$

$$= 0{,}282 .$$

Für Jahr 2 ist $n = 75$ und $\bar{x} = 265{,}27$. Der Die Berechnungen sind im Folgenden zusammengefasst Tisch:

xi	xi x	fi	fi xi x 3
225	-40,27	25	-1632213,81
245	-20,27	15	-124864,28
265	-0,27	9	-0,17
285	19.73	8	61473,98
305	39,73	6	376371.09
325	59,73	5	1065663,90
345	79,73	4	2027588.19
365	99,73	3	2976063.94
			4750082.83

Da $S = 42{,}01$, der Schiefekoeffizient ist gleich:

$$\overline{b1} = \frac{\frac{1}{75}(4750082{,}83)}{(42.01)3} = \frac{63334.438}{74140.933}$$

$$= 0{,}854 .$$

Der Schiefekoeffizient für Jahr 1 beträgt nahe Null ($\overline{b1} = 0{,}282$), also die tägliche Umsatzverteilung der 75 Bäckereien für Jahr 1 ist nahezu symmetrisch Verteilung. Für Jahr 2 ist der Schiefekoeffizient höher; das heißt, die Verteilung breitet sich nach rechts aus.

WEITERLESEN

Formmaß
Maß für die Schiefe

Machine Translated by Google

VERWEISE

Pearson, ES, Hartley, HO: Biometrie
Tabellen für Statistiker, Bde. I und II.
Cambridge University Press, Cambridge
(1966,1972)

Pearson, K.: Beiträge zur mathematischen
Evolutionstheorie. I. In: Karl
Pearsons frühe statistische Arbeiten. Cambridge
University Press, Cambridge,
S. 1–40 (1948). Erstveröffentlichung 1894
as: Zur Zerlegung asymmetrischer
Frequenzkurven. Philos. Trans. Roy. Soc.
London. Ser. A **185,** 71–110

Pearson, K.: Beiträge zur mathematischen
Evolutionstheorie. II: Schräglage
Variation in homogenem Material. In:
Karl Pearsons frühe statistische Arbeiten.
Cambridge University Press, Cambridge,
S. 41–112 (1948). Erstveröffentlichung in
1895 in Philos. Trans. Roy. Soc. London.
Ser. A **186,** 343–414

Variationskoeffizient

Der Variationskoeffizient ist ein **Maß für relative Streuung.** Sie beschreibt die **Standardabweichung** als Prozentsatz des **arithmetischen Mittels.**

Dieser Koeffizient kann zum Vergleich verwendet werden Streuungen **quantitativer Variablen ,** die werden nicht in denselben Einheiten ausgedrückt (z. B. beim Vergleich der Gehälter in verschiedenen Ländern, die in verschiedenen Währungen angegeben werden), oder die Streuungen von Variablen, die sehr verschiedene **Mittel.**

MATHEMATISCHE ASPEKTE

Der Variationskoeffizient *CV* ist definiert als das Verhältnis der **Standardabweichung** zur **arithmetisches Mittel** für eine Reihe von **Beobachtungen;** mit anderen Worten:

$$CV = \frac{S}{\bar{X}} \cdot 100$$

für ein **Beispiel,** wobei:

S ist die Standardabweichung der Stichprobe, Und

$\bar{X}$ ist das arithmetische Mittel der Stichprobe,

oder:

$$CV = \frac{\bar{\sigma}}{\bar{y}} \cdot 100$$

für eine **Bevölkerung,** wo

$\bar{\sigma}$ ist die Standardabweichung der Grundgesamtheit und

$\bar{y}$ ist der Mittelwert der Grundgesamtheit.

Dieser Koeffizient ist unabhängig von der Einheit von Maßeinheit, die für die Variable verwendet wird.

BEISPIELE

Lassen Sie uns die Gehaltsverteilungen für zwei untersuchen Unternehmen aus zwei verschiedenen Ländern. Laut einer **Umfrage** ist die **Arithmetik Mittelwerte** und die **Standardabweichungen** der Die Gehälter sind wie folgt:

Unternehmen A:

$$\bar{x} = 2500 \text{ CHF} \quad ,$$

$$S = 200 \text{ CHF} \quad ,$$

$$CV = \frac{200}{2500} \cdot 100 = 8\ \% \ .$$

Unternehmen B:

$$\bar{x} = 1000 \text{ CHF} \quad ,$$

$$S = 50 \text{ CHF} \quad ,$$

$$CV = \frac{50}{1000} \cdot 100 = 5\ \% \ .$$

Die **Standardabweichung** beträgt 8 % von das **arithmetische Mittel** für Unternehmen A und 5 % für Unternehmen B. Die Gehaltsverteilung ist für Unternehmen B etwas homogener als für Unternehmen A.

Machine Translated by Google

WEITERLESEN

Arithmetisches Mittel

Maß der Streuung
Standardabweichung

VERWEISE

Johnson, NL, Leone, FC: Statistik und experimentelles Design in den Ingenieurwissenschaften und den Naturwissenschaften. Wiley, New York (1964)

Kollinearität

Es ist bekannt, dass Variablen mathematisch kollinear sind, wenn eine von ihnen ein linearer Kom ist

Kombination der anderen Variablen. Sie werden als statistisch kollinear bezeichnet, wenn eine von ihnen annähernd eine Linearkombination anderer Variablen ist. Im Fall eines Regressionsmodells, bei dem die erklärenden Variablen stark miteinander korrelieren, spricht man von Kollinearität (oder Multikollinearität) zwischen den erklärenden Variablen. Im ersten Fall ist es einfach unmöglich, Kleinste-Quadrate-Schätzer zu definieren, und im zweiten Fall können diese Schätzer eine erhebliche Varianz aufweisen.

GESCHICHTE

Der Begriff „Kollinearität“ wurde in der Mathematik erstmals zu Beginn des 20. Jahrhunderts verwendet, aufgrund der Wiederentdeckung des Satzes von Pappus von Alexandria (einem Mathematiker aus dem 3. Jahrhundert). Seien A, B, C drei Punkte auf einer Geraden und A', B', C' drei Punkte auf einer anderen Geraden. Wenn wir die Paare mithilfe von AB'.A'B,CA'.AC' und BC'.B'C in Beziehung setzen, werden ihre Schnittpunkte in einer Linie auftreten; mit anderen Worten, die drei Schnittpunkte sind kollinear.

MATHEMATISCHE ASPEKTE

Im Fall einer Matrix erklärender Variablen X bedeutet Kollinearität, dass eine der Spalten von **X** (näherungsweise) eine Linearkombination der anderen Spalten ist. Das

impliziert, dass **XX** fast singulär ist. Folglich wird der durch die ÿ1- Methode der kleinsten Quadrate erhaltene **Schätzer ÿ** = **XXXY** durch Invertieren einer nahezu singulären **Matrix erhalten,** was dazu führt, dass ihre Komponenten instabil werden. Um diese Kollinearitätsprobleme zu lösen, wurde die **Ridge-Regressionstechnik** entwickelt.

Eine kollineare Beziehung zwischen mehr als zwei Variablen sind nicht immer das Ergebnis der Beobachtung der paarweisen Korrelationen zwischen den Variablen. Einen besseren Hinweis auf das Vorliegen eines Kollinearitätsproblems liefern Varianzinflationsfaktoren, *VIF*. Der Varianzinflationsfaktor einer erklärenden Variablen *Xj* ist definiert durch:

$$VIFj = \frac{1}{1 \ddot{y} R2_j},$$

wobei R2 *das* **Bestimmtheitsmaß** des **Modells** ist

$$Xj = \ddot{y}0 + \sum_{k=j} \ddot{y}kXk + \ddot{y}.$$

Der Koeffizient VIF nimmt Werte zwischen 1 und 1 an undÿ. Wenn die Xj mathematisch kollinear mit anderen Variablen sind, erhalten wir R2 = 1 und *VIFj* = ÿ. Wenn andererseits die *Xj* reziprok unabhängig sind, gilt R2 = 0 und *VIFj* = 1. In der Praxis gehen wir davon aus, dass es ein echtes Problem mit der Kollinearität gibt, wenn *VIFj* größer als 100 ist, was einem R2 entspricht, der größer als 100 ist 0,99.

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Invertieren einer singulären Matrix, ähnlich wie das Invertieren von 0, ist keine gültige Operation. Verwendung der

Machine Translated by Google

Nach dem gleichen Prinzip ähnelt das Invertieren einer fast singulären Matrix dem Invertieren einer sehr kleinen Nummer. Einige der Elemente der Matrix muss daher sehr groß sein. Folglich, wenn die erklärenden Variablen kollinear sind, einige Elemente der Matrix **(X** X)ÿ1 von **ÿ** wird wahrscheinlich sehr groß sein. Deshalb Kollinearität führt zu instabilen Regressionsschätzern. Abgesehen von diesem Problem führt die Kollinearität auch zu einem Berechnungsproblem; es ist Es ist schwierig, die Umkehrung von genau zu berechnen eine nahezu singuläre Matrix.

BEISPIELE

Es werden dreizehn Portionen Zement untersucht im folgenden Beispiel. Jede Portion enthält vier Zutaten, wie in der Beschreibung beschrieben Tisch. Ziel des Experiments ist es herauszufinden, wie die Größen X1, X2, X3 und X4, entsprechend den Mengen dieser vier Inhaltsstoffe wirken sich auf die Wärmemenge *Y* aus, die beim Aushärten des Zements abgegeben wird.

Yi- Wärmemenge, die während der abgegeben wird Aushärtung des i-ten Anteils (in Joule);

Xi1 Menge der Zutat 1 (Tricalciumaluminat) in der i-ten Portion;

Xi2 Menge des Inhaltsstoffs 2 (Tricalciumsilikat) in der i-ten Portion;

Xi3 Menge des Inhaltsstoffs 3 (Tetracalcium-Aluminoferrit) im i-ten Teil;

Xi4 Menge der Zutat 4 (Dicalciumsilikat) in der i-ten Portion.

Tabelle: Wärmeabgabe der Zementportionen beim Aushärten

Portion	Ingredient	Zutat	Zutat	Ingredient	Hitze
...	1 $X_{1\,2}$	$X_{2\,3}$	$X_{3\,4}$	X_4	Y
1	7	26	6	60	78,5
2	1	29	12	52	74,3
3	11	56	8	20	104.3
4	11	31	8	47	87,6
5	7	54	6	33	95,9
6	11	55	9	22	109.2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93.1
10	21	48	4	26	115,9
11	1	40	23	34	83,9
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Quelle: Birkes & Dodge (1993)

Wir beginnen mit einer **einfachen linearen Regression.** Das für die lineare Regression verwendete Modell Ist:

$$Y = \ddot{y}0 + \ddot{y}1X1 + \ddot{y}2X2 + \ddot{y}3X3 + \ddot{y}4X4 + \ddot{y}\ .$$

Wir erhalten folgende Ergebnisse:

Variable	ÿ^...	Sd	tc
Konstant 62,41		**70.07**	**0,89**
X1	1.5511	0,7448	2.08
X2	0,5102	0,7238	0,70
X3	0,1019	0,7547	0,14
X4	ÿ0,1441	0,7091	*ÿ0,20*

Wir können sehen, dass nur der Koeffizient X1 deutlich größer als Null ist; in den anderen Fällen *tc < t(ÿ/2,nÿ2)* (Wert aus Student Tabelle) für ein **Signifikanzniveau** von ÿ = 0,1. Darüber hinaus sind die Standardabweichungen der geschätzten Koeffizienten ÿ 2, ÿ 3 und ÿ 4 größer dann die Koeffizienten selbst. Wenn das

Machine Translated by Google

Es ist bekannt, dass Variablen stark korrelieren
dass die Wirkung von einem die Wirkung von maskieren kann
andere. Aus diesem Grund können die Koeffizienten
scheinen sich unwesentlich davon zu unterscheiden
null.

Um das Vorhandensein von *Multikollinearität* zu überprüfen
Für ein paar Variablen berechnen wir die entsprechende
Beziehungsmatrix.

	x_1	x_2	x_3	x_4
X1	1	0,229 ÿ0,824		0,245
X2	0,229	1	ÿ0,139 ÿ0,973	
X3 ÿ0,824 ÿ0,139			1	0,030
X4	0,245 ÿ0,973		0,030	1

Wir stellen fest, dass eine starke negative Korrelation besteht
(ÿ0,973) liegt zwischen X2 und X4. Wenn wir uns die
Daten ansehen, können wir den Grund dafür erkennen
Das. Abgesehen von den Portionen 1 bis 5 ist die Gesamtsumme
Die Menge an Silikaten *(X2* + X4) beträgt nahezu
über die Portionen hinweg konstant und beträgt ca. 77;
daher ist X4 ungefähr
77 ÿ X2. Diese Situation lässt es uns nicht zu
zwischen den einzelnen Wirkungen zu unterscheiden
von X2 und denen von X4. Wir sehen zum Beispiel
dass die vier größten Werte von X4 (60, 52, 47
und 44) entsprechen kleineren *Y*- Werten
dann beträgt die mittlere Wärme 95,4. Deshalb zunächst
Auf den ersten Blick scheint es, dass große Mengen
der Zutat 4 zu geringen Hitzemengen führen.

Allerdings stellen wir auch fest, dass die vier größten
Die Werte von X4 entsprechen den vier kleinsten
Werten von X2, was eine negative Korrelation zwischen
X2 und X4 ergibt. Das deutet darauf hin
Die alleinige Einnahme von Inhaltsstoff 4 hat eine geringe Wirkung
auf Variable Y, und dass die kleinen Mengen von 2
allein genommen die kleinen erklären können
abgegebene Wärmemenge. Daher das Lineare
Die Abhängigkeit zwischen zwei erklärenden Variablen
(X2 und X4) macht es komplizierter, die Wirkung jeder
Variablen einzeln zu erkennen
die Antwortvariable Y.

WEITERLESEN

Multiple lineare Regression

Ridge-Regression

VERWEISE

Birkes, D., Dodge, Y.: Alternative Methoden
der Regression. Wiley, New York (1993)

Bock, RD: Multivariate statistische Methoden in der
Verhaltensforschung. McGraw
Hill, New York (1975)

Kombination

Eine Kombination ist eine ungeordnete Sammlung
einzigartiger Elemente oder Objekte.
Eine k-Kombination ist eine Teilmenge mit *k* Elementen.
Die Anzahl der k-Kombinationen aus einer Menge von
n Elemente ist die Anzahl der **Anordnungen.**

GESCHICHTE

Siehe **kombinatorische Analyse.**

MATHEMATISCHE ASPEKTE

1. *Kombination ohne Wiederholung*
   Kombination ohne Wiederholung beschreiben
   die Situation, in der sich jedes gezeichnete Objekt befindet
   nicht für die nächste Zeichnung zurückgelegt
   Objekt kann daher nur einmal vorkommen
   jede Gruppe.
   Die Anzahl der Kombinationen ohne Wiederholung
   von *k* Objekten unter *n* ist gegeben durch:

$$Ck_N = \binom{N}{k} = \frac{N!}{k! \cdot (n \ddot{y} k)!} .$$

2. *Kombination mit Wiederholungen*
   Kombination mit Wiederholungen (oder mit
   Überweisung) werden bei jeder Ziehung verwendet
   Das Objekt wird für die nächste Zeichnung
   zurückgelegt. Jedes Objekt kann dann mehrmals vorkommen
   jede Gruppe, *r = 0, 1,...,k.*

Machine Translated by Google

Die Anzahl der Kombinationen mit Wiederholungen von *k* Objekten unter *n* ist gegeben durch:

$$KR_N^k = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k! \cdot (n-1)!}.$$

BEISPIELE

1. *Kombinationen ohne Wiederholung*
   Betrachten Sie die Situation, in der wir müssen
   Wählen Sie ein Komitee aus drei Personen aus
   eine Versammlung von acht Personen. Wie viele
   Es könnten möglicherweise verschiedene Ausschüsse sein
   aus dieser Versammlung ausgewählt, wenn jede Person
   kann in jeder Gruppe nur einmal ausgewählt werden?
   Hier müssen wir die Zahl berechnen
   der möglichen Kombinationen von drei Personen
   ab acht:

$$C_N^k = \frac{N!}{k! \cdot (n-k)!} = \frac{8!}{3! \cdot (8-3)!} = \frac{40320}{6 \cdot 120} = 56$$

   Daher ist es möglich, aus einer Versammlung
   von acht Personen 56 verschiedene Ausschüsse
   mit jeweils drei Personen zu bilden.
2. *Kombinationen mit Wiederholungen* berücksichtigen
   eine Urne mit sechs nummerierten Kugeln. Wir
   vier aufeinanderfolgende Zeichnungen durchführen und
   Legen Sie die gezogene Kugel anschließend zurück in die Urne
   jede Zeichnung. Wie viele verschiedene
   Kombinationen könnten aus dieser Zeichnung entstehen?
   In diesem Fall wollen wir die Nummer finden
   von Kombinationen mit Wiederholung (weil
   Jede gezogene Kugel wird zurück in die Urne gelegt
   vor der nächsten Ziehung). Wir erhalten

$$KR_N^k = \frac{(n+k-1)!}{k! \cdot (n-1)!} = \frac{9!}{4! \cdot (6-1)!} = \frac{362880}{24 \cdot 120} = 126$$

   verschiedene Kombinationen.

WEITERLESEN

Anordnung
Binomialverteilung
Kombinatorische Analyse
Permutation

C

Kombinatorische Analyse

Die kombinatorische Analyse bezieht sich auf eine Gruppe von Techniken, die zur Bestimmung verwendet werden können Anzahl der Elemente in einer bestimmten Menge, ohne sie einzeln zählen zu müssen.
Die fraglichen Elemente könnten die Ergebnisse sein aus einem wissenschaftlichen **Experiment** oder etwas anderem mögliche Ergebnisse eines zufälligen **Ereignisses.**
Drei besondere Konzepte sind wichtig Kombinatorische Analyse:

- Permutationen;
- Kombinationen;
- Arrangements.

GESCHICHTE

Die kombinatorische Analyse interessiert Mathematiker seit Jahrhunderten. Laut Takacs (1982) geht eine solche Analyse auf die Antike zurück Griechenland. Doch die Hindus, die Perser (einschließlich des Dichters und Mathematikers). Khayyâm, Omar) und (besonders) die Chinesen untersuchten ebenfalls solche Probleme. Ein 3000 Das einjährige chinesische Buch „I Ging“ beschreibt die möglichen **Anordnungen** einer Menge von *n* Elemente, wobei *n* ÿ 6. Im Jahr 1303 veröffentlichte Chu, Shih chieh ein Werk mit dem Titel „Ssu Yuan Yü Chien“ (Kostbarer Spiegel der Vier). Elemente). Das Cover des Buches zeigt ein Dreieck, das die Kombinationen von zeigt *k* Elemente aus einer Menge der Größe *n* , wobei 0 ÿ *k* ÿ n. Dieses arithmetische Dreieck war auch von mehreren europäischen Mathematikern wie Stifel, Tartaglia und Hérigone erforscht, und insbesondere Pascal, der das „Trai-

Machine Translated by Google

té du Triangle Arithmétique“ (Abhandlung des arithmetisches Dreieck) im Jahr 1654 (obwohl es war erst nach seinem Tod im Jahr 1665 veröffentlicht). Ein weiteres Dokument zu **Kombinationen** war veröffentlicht 1617 von Puteanus, Erycius genannt „Erycii Puteani Pretatis Thaumata“. in Bernardi Bauhusii è Societate Jesu Pro teum Parthenium“. Allerdings kombinatorisch Die Analyse offenbarte erst mit die Werke von Fermat (1601–1665) und Pascal (1623–1662). Der Begriff „kombinatorisch Analyse“ wurde von Leibniz (1646–1646) eingeführt. 1716) im Jahr 1666. In seinem Werk „Dissertatio de Arte Combinatoria“ untersuchte er systematisch Probleme im Zusammenhang mit **Anordnungen, Permutationen** und **Kombinationen.**

Andere Arbeiten auf diesem Gebiet sollten erwähnt werden hier, wie die von Wallis, J. (1616–1703), berichtet in „The Doctrine of Permutations und Kombinationen“ (ein wesentlicher und grundlegender Teil der „Lehren des Zufalls“), Orthesen **von Bernoulli, J., Moivre, A.** de, Car dano, G. (1501–1576) und Galileo (1564–) 1642).

In der zweiten Hälfte des 19. Jahrhunderts Ry, Cayley (1829–1895) löste einige Probleme im Zusammenhang mit dieser Art der Analyse mithilfe von Grafiken, die er „Bäume“ nannte. Schließlich sollten wir es tun Erwähnen Sie auch das wichtige Werk von MacMahon (1854–1929), „Combinatory Analysis“. (1915–1916).

BEISPIELE

Siehe **Anordnung, Kombination** und **pro Mutation.**

WEITERLESEN

Arithmetisches Dreieck
Anordnung
Binomial
Binomialverteilung
Kombination
Permutation

VERWEISE

MacMahon, PA: Kombinatorische Analyse, Bände. I und II.CambridgeUniversity Press, Cambridge (1915–1916)

Stigler, S.: Die Geschichte der Statistik, die Messung der Unsicherheit vorher 1900. Belknap, London (1986)

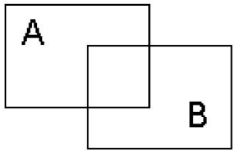
Takács, L.: Kombinatorik. In: Kotz, S., Johnson, NL (Hrsg.) Encyclopedia of Statistical Sciences, vol. 2. Wiley, New York (1982)

Kompatibilität

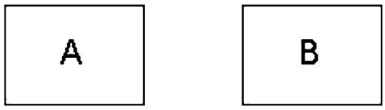
Zwei **Ereignisse** gelten als kompatibel, wenn die Das Eintreten des ersten Ereignisses verhindert nicht das Eintreten des zweiten (oder anderen). Wörter, wenn der **Schnittpunkt** zwischen den beiden Ereignisse ist nicht null):

$$P(A \ddot{y} B) = 0 .$$

Wir können zwei kompatible **Ereignisse** *A* darstellen und *B* schematisch in folgender Weise:



Zwei **Ereignisse** *A* und *B* sind inkompatibel (bzw sich gegenseitig ausschließen), wenn das Auftreten von *A* verhindert das Auftreten von B oder umgekehrt. Wir können dies folgendermaßen darstellen:



Machine Translated by Google

Dies bedeutet, dass die **Wahrscheinlichkeit ,** dass diese zwei Ereignisse gleichzeitig passieren, ist Null:

$$A \ddot{y} B = \ddot{y} \ddot{y}\ddot{y} \; P(A \ddot{y} B) = 0 .$$

MATHEMATISCHE ASPEKTE

Wenn zwei **Ereignisse** *A* und *B* kompatibel sind, ist die **Wahrscheinlichkeit ,** dass mindestens eines der Ereignisse auftritt, kann wie folgt ermittelt werden Formel:

$$P(A \ddot{y} B) = P(A) + P(B) \ddot{y} P(A \ddot{y} B) .$$

Wenn andererseits die beiden **Ereignisse** *A* und *B* sind inkompatibel, die **Wahrscheinlichkeit ,** dass die zwei Ereignisse *A* und *B* passieren gleichzeitig ist Null:

$$P(A \ddot{y} B) = 0 .$$

Die Wahrscheinlichkeit, dass mindestens eines der Ereignisse auftritt auftritt, kann durch einfaches Hinzufügen der erhalten werden individuelle Wahrscheinlichkeiten von *A* und B:

$$P(A \ddot{y} B) = P(A) + P(B) .$$

BEISPIELE

Stellen Sie sich ein **Zufallsexperiment** vor Dabei wird eine Karte aus einem Stapel gezogen 52 Karten. Uns interessieren die drei fol

Lowing- **Ereignisse:**

A = „Zeichne ein Herz“

B = „Ziehe eine Königin“

C = „Ziehe einen Verein“ .

Die mit jedem von ihnen verbundenen **Wahrscheinlichkeiten** Diese **Ereignisse** sind:

$$P(A) = \frac{13}{52}$$

$$P(B) = \frac{4}{52}$$

$$P(C) = \frac{13}{52} .$$

Die **Ereignisse** *A* und *B* sind kompatibel, weil Es ist möglich, sowohl ein Herz als auch eine Königin zu zeichnen gleichzeitig (die Königin der Herzen). Daher ist der Schnittpunkt zwischen *A* und *B* der Herzkönigin. Die Wahrscheinlichkeit dieses Ereignisses ist gegeben durch:

$$P(A \ddot{y} B) = \frac{1}{52} .$$

Die **Wahrscheinlichkeit** der Vereinigung der beiden **Ereignisse** *A* und *B* (entweder ein Herz zeichnen oder eine Königin) ist dann gleich:

$$P(A \ddot{y} B) = P(A) + P(B) \ddot{y} P(A \ddot{y} B)$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$

$$= \frac{16}{52} .$$

Andererseits die **Ereignisse** *A* und *C* sind inkompatibel, weil eine Karte es nicht sein kann Herz und Verein zugleich! Die **Kreuzung** zwischen *A* und *C* ist eine leere Menge.

$$A \ddot{y} C = \ddot{y} .$$

Die **Wahrscheinlichkeit** der Vereinigung der beiden Ereignisse *A* und *C* (Zeichnen eines Herzens oder einer Keule) ist einfach durch die Summe der Wahrscheinlichkeiten gegeben jeder Veranstaltung:

$$P(A \ddot{y} C) = P(A) + P(C)$$

$$= \frac{13}{52} + \frac{13}{52}$$

$$= \frac{1}{2} .$$

WEITERLESEN

Fall

Unabhängigkeit

Wahrscheinlichkeit

Machine Translated by Google

Komplementär

Betrachten Sie den **Probenraum** für ein **Zufallsexperiment.**

Für jedes **Ereignis** A kann ein Element , wir können ein neues Ereignis *B* bestimmen , das alle enthält die Elemente des **Probenraums ,** die nicht in A enthalten sind.

Dieses Ereignis *B* heißt das Komplement von *A* bzgl. und wird durch die Negation von A erhalten.

MATHEMATISCHE ASPEKTE

Betrachten Sie ein **Ereignis** A, das ein Element des **Probenraums** ist .

Das Komplement von *A* bezüglich des Ist bezeichneten A¯. Es ergibt sich durch die Negation von A:

$$A^{-} = \ddot{y}\ A$$

$$= \{w \ddot{y}\ ; w \ddot{y}/ A\}\ .$$

BEISPIELE

Stellen Sie sich ein **Zufallsexperiment** vor , bei dem eine Münze dreimal geworfen wird.

Der **Probenraum** dieses Experiments ist

$$= \{TTT, TTH, THT, THH,$$

$$HTT, HTH, HHT, HHH\}\ .$$

Betrachten Sie das **Ereignis**

$$A = \text{„Kopf } (H) \text{ kommt zweimal vor“}$$

$$= \{THH, HTH, HHT\}\ .$$

Das Komplement von *A* bezüglich gleich Ist

$$A^{-} = \{TTT, TTH, THT, HTT, HHH\}$$

$$= \text{„Kopf } (H) \text{ kommt nicht zweimal vor“}\ .$$

WEITERLESEN

- **Fall**
- **Zufälliges Experiment**
- **Probenraum**

Vollständige Verknüpfungsmethode

Die vollständige Verknüpfungsmethode ist eine hierarchische **Klassifizierungsmethode ,** bei der die Entfernung berücksichtigt wird zwischen zwei Klassen ist definiert als der größte Abstand, der bei Auswahl erreicht werden könnte ein Element aus jeder Klasse und Kennzahl der Abstand zwischen diesen Elementen. In anderen Mit anderen Worten, es ist der Abstand zwischen den am weitesten entfernten Elementen jeder Klasse.

Beispielsweise ist der **Abstand ,** der zum Erstellen der **Abstandstabelle** verwendet wird , der euklidische Abstand.

Bei Verwendung der vollständigen Verknüpfungsmethode wird der Abstand zwischen zwei Klassen durch gegeben Euklidischer Abstand zwischen den am weitesten entfernten Elemente (der maximale Abstand):

MATHEMATISCHE ASPEKTE

Siehe **Clusteranalyse.**

WEITERLESEN

- **Einstufung**
- **Clusteranalyse**
- **Dendrogramm**
- **Distanz**
- **Entfernungstabelle**

Vollständig Randomisiertes Design

Ein vollständig randomisiertes Design ist eine Art experimentelles Design, bei dem die experimentellen Einheiten zufällig den verschiedenen zugewiesen werden Behandlungen.

Es wird verwendet, wenn angenommen wird, dass die experimentellen Einheiten „einheitlich“ sind; das heißt, wenn es im Experiment keinen unkontrollierten **Faktor** gibt .

GESCHICHTE

Siehe **Versuchsplanung.**

Machine Translated by Google

BEISPIELE

Wir wollen fünf verschiedene Medikamente auf Basis von Aspirin testen. Dazu verteilen wir die fünf Medikamententypen nach dem Zufallsprinzip an 40 Patienten. Wenn wir die fünf Medikamente mit A, B, C, *D* und E bezeichnen, erhalten wir die folgende Zufallsverteilung:

A wird 10 Personen zugeschrieben;

B wird 12 Personen zugeschrieben;

C wird 4 Personen zugeschrieben;

D wird 7 Personen zugeschrieben;

E wird 7 Personen zugeschrieben.

Wir haben dann ein vollständig randomisiertes Design, bei dem die Behandlungen (Medikamente) nach dem Zufallsprinzip den Versuchseinheiten (Patienten) zugeordnet werden und jeder Patient nur eine Behandlung erhält. Wir gehen auch davon aus, dass die Patienten „einheitlich“ sind: dass es keine Unterschiede gibt zwischen ihnen. Darüber hinaus gehen wir davon aus, dass es keinen unkontrollierten **Faktor** gibt , der inter Venen während der Behandlung.

In diesem Beispiel handelt es sich bei dem vollständig randomisierten Design um ein **faktorielles Experiment ,** das nur einen Faktor verwendet: das Aspirin. Die fünf Arten von Aspirin weisen unterschiedliche Konzentrationen des Faktors auf.

WEITERLESEN

Versuchsplanung

Experiment

Zusammengesetzt

Indexnummer

Zusammengesetzte **Indexzahlen** ermöglichen es uns, mit einer einzigen Zahl die relativen Variationen innerhalb einer Gruppe von **Variablen** beim Übergang von einer Situation zur anderen zu messen.

Der Verbraucherpreisindex, der Großhandelspreisindex, der Beschäftigungsindex und der

Beispiele für zusammengesetzte Indexzahlen sind der Dow-Jones-Index.

Das Ziel der Verwendung zusammengesetzter Indexzahlen besteht darin, alle einfachen Indexzahlen, die in einer komplexen Zahl (einem aus einer Menge einfacher Werte gebildeten Wert) enthalten sind, in nur einem Index zusammenzufassen.

Die am häufigsten verwendeten zusammengesetzten Indexzahlen sind:

• Der **Laspeyres-Index;** • Der **Paasche-Index;** • Der **Fisher-Index.**

GESCHICHTE

Siehe **Indexnummer.**

MATHEMATISCHE ASPEKTE

Es gibt verschiedene Methoden zum Erstellen zusammengesetzter Indexnummern.

Um diese Methoden zu veranschaulichen, verwenden wir ein Szenario, in dem ein Preisindex für die aktuelle Periode *n* in Bezug auf eine Referenzperiode 0 ermittelt wird.

1. *Indexzahl des arithmetischen Mittels* (Summenmethode):

$$In/0 = \frac{Pn}{P0} \cdot 100 \quad ,$$

Dabei ist Pn die Summe der Artikelpreise im aktuellen Zeitraum und P0 die Summe der Artikelpreise im Referenzzeitraum.

2. *Arithmetisches Mittel der einfachen Indexzahl Berser:*

$$In/0 = \frac{1}{N} \cdot \frac{Pn}{P0} \cdot 100 \quad ,$$

Dabei ist *N* die Anzahl der betrachteten Güter und *Pn* die **einfache Indexnummer** P0 jedes Artikels.

Bei diesen beiden Methoden hat jedes Element die gleiche Bedeutung. Das ist eine Situation

Machine Translated by Google

was oft nicht der Realität entspricht.

3. *Indexnummer der gewichteten Arithmetik bedeutet* (die gewichtete Summenmethode):
Die allgemeine Formel für eine Indexzahl berechnet nach der Methode der gewichteten Summe folgendermaßen:

$$\text{In/0} = \frac{Pn \cdot Q}{P0 \cdot Q} \cdot 100 .$$

Wählen Sie die Menge Q für jeden Artikel betrachtet könnte sich hier als problematisch erweisen: Bei der Berechnung muss Q sowohl für den Zähler als auch für den Nenner gleich sein ein Preisindex.
Im **Laspeyres-Index** ist der Wert von Q dem Bezugsjahr entspricht gebraucht. Im **Paasche-Index ist der** Wert von Q für das laufende Jahr verwendet. Andere Statistiker haben vorgeschlagen, den Wert von Q zu verwenden für ein bestimmtes Jahr.

BEISPIELE

Betrachten Sie die folgende Tabelle die (fiktiven) Preise von drei Verbrauchern Waren im Referenzjahr (1970) und deren aktuelle Preise.

	Preis (Franken)	
Waren	**1970 (P_0)**	**Jetzt (Pn)**
Milch	0,20	1,20
Brot	0,15	1.10
Butter	0,50	2,00

Anhand dieser Zahlen untersuchen wir nun die drei Hauptmethoden zur Konstruktion zusammengesetzter **Indexzahlen.**

1. *Indexzahl der arithmetischen Mittelwerte* (die Summenmethode):

$$\text{In/0} = \frac{Pn}{P0} \cdot 100$$
$$= \frac{4.30}{0,85} \cdot 100 = 505,9 .$$

Nach dieser Methode der Preisindex ist um 405,9 % gestiegen (505,9 ÿ 100) zwischen dem Bezugsjahr und dem aktuellen Jahr Jahr.

2. *Arithmetisches Mittel der einfachen Indexzahl Berser:*

$$\text{In/0} = \frac{1}{N} \cdot \frac{Pn}{P0} \cdot 100$$
$$= \frac{1}{3} \cdot \frac{1,20}{0,20} + \frac{1.10}{0,15} + \frac{2,00}{0,50} \cdot 100$$
$$= 577,8 .$$

Diese Methode ergibt etwas anderes Ergebnis aus dem vorherigen, da wir eine Steigerung von 477,8 % (577,8 ÿ 100) im Preisindex zwischen dem Referenzjahr und dem aktuellen Jahr.

3. *Indexnummer der gewichteten Arithmetik bedeutet* (die gewichtete Summenmethode):

$$\text{In/0} = \frac{Pn \cdot Q}{P0 \cdot Q} \cdot 100 .$$

Diese Methode wird in Verbindung mit verwendet der **Laspeyres-Index** oder der **Paasche Index.**

WEITERLESEN

Fisher-Index
Indexnummer
Laspeyres-Index
Paasche-Index
Einfache Indexnummer

Bedingte Wahrscheinlichkeit

Die Wahrscheinlichkeit eines Ereignisses angesichts eines anderen Es ist bekannt, dass das Ereignis stattgefunden hat.

Die bedingte Wahrscheinlichkeit wird bezeichnet P(A|B), was als „Wahrscheinlichkeit“ gelesen wird von *A* bedingt durch B.“

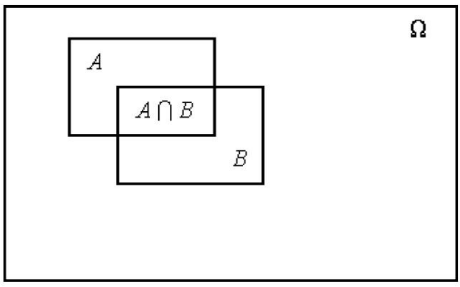
Machine Translated by Google

GESCHICHTE

Der Gedanke der **Unabhängigkeit** dominierte Wahrscheinlichkeitstheorie bis zum Anfang das zwanzigste Jahrhundert. Im Jahr 1933, Kolmogorov, Andrei Nikolajewitsch stellte das Konzept vor der bedingten Wahrscheinlichkeit; dieses Konzept jetzt spielt eine wesentliche Rolle in der Theorie und angewandte Wahrscheinlichkeit und Statistik.

MATHEMATISCHE ASPEKTE

Betrachten Sie ein **Zufallsexperiment,** für das Wir kennen den **Probenraum .** Betrachten Sie zwei Ereignisse A und B aus diesem Raum. **Die Wahrscheinlichkeit** von A, *P(A)* hängt von der Menge der möglichen Ereignisse im **Experiment** ().



Bedenken Sie nun, dass wir eine Ergänzung haben Informationen zum Experiment: das Das Ereignis *B* ist eingetreten. Die Wahrscheinlichkeit von Das auftretende Ereignis *A* ist dann eine Funktion des RaumesBratherals eine Funktion von. Der Die Wahrscheinlichkeit von A bedingt *durch B* wird wie folgt berechnet:

$$P(A|B) = \frac{P(A \ddot{y} B)}{P(B)} .$$

Wenn *A* und *B* zwei inkompatible Ereignisse sind, Der Schnittpunkt zwischen *A* und *B* ist ein Freiraum. Wir haben dann P(A|B) = P(B|A) = 0.

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Konzept der bedingten Wahrscheinlichkeitist eins der Wahrscheinlichkeitswichtigsten ory. Diese Bedeutung ist vor allem auf folgende Punkte zurückzuführen:

1. Wir sind oft daran interessiert, die zu berechnen Wahrscheinlichkeit eines **Ereignisses ,** wenn bereits einige Informationen über das Ergebnis bekannt sind. In diesem Fall beträgt die erforderliche Wahrscheinlichkeit eine bedingte Wahrscheinlichkeit.
2. Auch wenn unvollständige Informationen über die Wenn das Ergebnis nicht bekannt ist, können bedingte Wahrscheinlichkeiten bei der Berechnung nützlich sein Wahrscheinlichkeiten erforderlich.

BEISPIELE

Stellen Sie sich eine Gruppe von 100 verteilten Autos vor nach zwei Kriterien: Komfort und Geschwindigkeit. Wir werden folgende Unterscheidungen treffen:

ein Auto kann sein schnell / langsam ,

ein Auto kann sein komfortabel / unbequem .

Eine Aufteilung der 100 Autos anhand dieser Kriterien ist in der folgenden Tabelle dargestellt:

	schnell	langsam	gesamt
komfortabel	40	10	50
unbequem	20	30	50
gesamt	60	40	100

Betrachten Sie die folgenden Ereignisse:

A = „ein schnelles Auto wird gewählt“

und *B* = „Es wird ein komfortables Auto gewählt.“

Die Wahrscheinlichkeiten dieser beiden Ereignisse sind:

$$P(A) = 0{,}6 \quad ,$$

$$P(B) = 0{,}5 .$$

Die Wahrscheinlichkeit, sich für ein schnelles Auto zu entscheiden, beträgt dann von 0,6 und die Wahl eines komfortablen Autos ist 0,5.

Machine Translated by Google

Stellen Sie sich nun vor, wir erhalten eine Zusatzinformation: Es wurde ein schnelles Auto ausgewählt. Wie hoch ist dann die Wahrscheinlichkeit, dass es sich um dieses Auto handelt? auch bequem?

Wir berechnen die Wahrscheinlichkeit, dass *B* es weiß *dass A* eingetreten ist, oder die bedingte Wahrscheinlichkeit von *B* in Abhängigkeit von A: Das finden wir in der Tabelle

$$P(A \ddot{y} B) = 0{,}4 .$$

$$\ddot{y}\ P(B|A) = \frac{P(A \ddot{y} B)}{P(A)} = \frac{0{,}4}{0{,}6} = 0{,}667 .$$

Die Wahrscheinlichkeit, dass das Auto komfortabel ist, Da wir wissen, dass es schnell ist, gibt es 2 vorn $\frac{}{3}$.

WEITERLESEN

Fall

Wahrscheinlichkeit

Zufälliges Experiment

Probenraum

VERWEISE

Kolmogorov, AN: Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin Heidelberg New York (1933)

Kolmogorov, AN: Grundlagen der Wahrscheinlichkeitstheorie. Chelsea Publishing Unternehmen, New York (1956)

Konfidenzintervall

Ein Konfidenzintervall ist jedes um einen **Schätzer** herum konstruierte **Intervall ,** das mit einer bestimmten **Wahrscheinlichkeit** das Wahre enthält **Wert** des entsprechenden **Parameters** von eine **Bevölkerung.**

GESCHICHTE

Laut Desrosières, A. (1988) war Bowley, AL einer der ersten, die sich dafür interessierten das Konzept des Konfidenzintervalls. Bowley präsentierte der Royal Statistical Society im Jahr 2010 seine ersten Berechnungen mit Vertrauensintervallen 1906.

MATHEMATISCHE ASPEKTE

Um ein Konfidenzintervall zu erstellen die den wahren **Wert** des **Parameters** ÿ mit einer gegebenen **Wahrscheinlichkeit enthält,** eine Gleichung folgender Form zu lösen:

$$P(Li \ddot{y} \ddot{y} \ddot{y} Ls) = 1 \ddot{y} \ddot{y} \quad ,$$

Wo

ÿ ist der zu schätzende Parameter,

Li ist die untere Grenze des **Intervalls,**

Ls ist die Obergrenze des Intervalls und

1 ÿ ÿ ist die gegebene Wahrscheinlichkeit, genannt **Konfidenzniveau** des Intervalls.

Die Wahrscheinlichkeitÿ misst das **Fehlerrisiko** von das Intervall, also die Wahrscheinlichkeit, dass das Das Intervall enthält nicht den wahren **Wert** des Parameter ÿ.

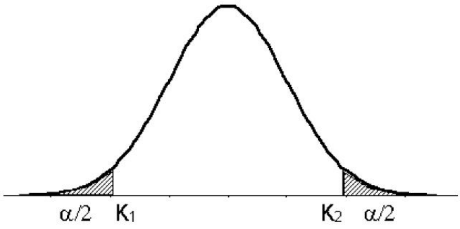
Um diese Gleichung zu lösen, muss eine Funktion verwendet werden *f(t,* ÿ) muss definiert werden, wobei *t* ein **Schätzer** von ÿ ist, für den die **Wahrscheinlichkeitsverteilung** bekannt ist.

Das Definieren dieses **Intervalls** für *f(t,* ÿ) erfordert Schreiben der Gleichung:

$$P(k1 \ddot{y} f(t, \ddot{y}) \ddot{y} k2) = 1 \ddot{y} \ddot{y} \quad ,$$

wobei die Konstanten k1 und k2 gegeben sind durch die **Wahrscheinlichkeitsverteilung** der Funktion *f(t,* ÿ). Im Allgemeinen wird das Fehlerrisiko ÿ bei der Verteilung in zwei gleiche Teile geteilt $\frac{\ddot{y}}{2}$ jede Seite der Verteilung von *f(t,* ÿ). Wenn wegen Beispielsweise folgt die Funktion *f(t,* ÿ) einer zentrierten und reduzierten **Normalverteilung,** der Konstantensk1 undk2sind symmetrisch und können durch $\ddot{y}z$ ÿ dargestellt werden $\frac{}{2}$ und $+z$ ÿ $\frac{}{2}$, wie gezeigt in der folgenden Abbildung.

Machine Translated by Google



Nachdem die Konstanten k1 und k2 gefunden wurden, muss der **Parameter** ÿ in der oben angegebenen Gleichung isoliert werden. Auf diese Weise wird das Konfidenzintervall ÿ für das **Konfidenzniveau** 1 ÿ ÿ ermittelt.

BEREICHE UND EINSCHRÄNKUNGEN

Bei der Interpretation eines Konfidenzintervalls sollte man sehr vorsichtig sein. Wenn wir für ein **Konfidenzniveau** von 95 % ein Konfidenzintervall für einen **Mittelwert** von ÿ finden, bei dem die Unter- und Obergrenzen k1 bzw. k2 sind , können wir (zum Beispiel) Folgendes schlussfolgern: „Auf der Grundlage der untersuchten. **In der Stichprobe** können wir bestätigen, dass es wahrscheinlich ist, dass der **Mittelwert** der **Grundgesamtheit** in dem festgelegten **Intervall** gefunden werden kann .“ Es wäre nicht richtig, daraus den Schluss zu ziehen ist eine 95-prozentige Chance, den **Mittelwert** der **Grundgesamtheit** im **Intervall zu finden.** Da ÿ und die Grenzen sk1 und k2 des Intervalls konstant sind, kann das Intervall ÿ enthalten oder auch nicht. Wenn der Statistiker jedoch die Möglichkeit hat, das **Experiment** (das darin besteht, eine **Stichprobe** aus der Grundgesamtheit zu ziehen) mehrmals zu wiederholen, enthalten 95 % der erhaltenen Intervalle den wahren **Wert** von ÿ.

BEISPIELE

Ein Unternehmen, das Glühbirnen herstellt, möchte die durchschnittliche Lebensdauer seiner Glühbirnen testen. Die Verteilung der **Zufallsvariablen** X, die die Lebensdauer in Stunden darstellt, ist eine **Normalverteilung** mit **Mittelwert** ÿ und **Standardabweichung** ÿ = 30.

Um ÿ zu schätzen, brennt das Unternehmen *n* = 25 Glühbirnen aus.

Es erreicht eine durchschnittliche Lebensdauer von $\bar{x}$ = 860 Stunden. Es möchte Vertrauen schaffen Intervall um den **Schätzer** $\bar{x}$ bei einer **Konfi**

Denzwert von 0,95. Daher besteht der erste Schritt darin, eine Funktion *f(t,* ÿ) = *f(x,* $\bar{x}$ ÿ) für die bekannte Verteilung zu erhalten. Hier verwenden wir:

$$f(t, \ddot{y}) = f(x, \bar{x}\ \ddot{y}) = \frac{\bar{x}\ \ddot{y}\ \ddot{y}}{\ddot{y}\ddot{y}\ n},$$

das einem zentrierten und reduzierten **Nor** folgt **schlechte Verteilung.** Die Gleichung *P(k1* ÿ *f(t,* ÿ) ÿ k2) = 1 ÿ ÿ wird zu:

$$P\ \ddot{y}z0{,}025\ \ddot{y}\ \ddot{y}\ z0{,}025 = 0{,}95\ \ddot{y}\ddot{y}\ \frac{\bar{x}\ \ddot{y}\ \ddot{y}}{n},$$

weil das **Fehlerrisiko** ÿ geteilt wurde in zwei gleiche Teile bei $\frac{\ddot{y}}{2}$ = 0,025.

Der Tisch für den zentrierten und reduzierten **Nor**

Die Malverteilung, die **Normaltabelle,** ergibt z0,025 = 1,96. Deshalb:

$$P\ \ddot{y}1{,}96\ \ddot{y}\ \ddot{y}\ 1{,}96 = 0{,}95\ \ddot{y}\ddot{y}\ \frac{\bar{x}\ \ddot{y}\ \ddot{y}}{n}$$

Um das Konfidenzintervall für ÿ bei einem **Konfidenzniveau** von 0,95 zu erhalten, muss ÿ in der obigen Gleichung isoliert werden:

$$P\ \ddot{y}1{,}96\ \ddot{y}\ \ddot{y}\ 1{,}96 = 0{,}95\ \frac{\bar{x}\ \ddot{y}\ \ddot{y}}{\ddot{y}\ddot{y}\ n}$$

$$P\ \ddot{y}1{,}96\ \ddot{y}\ \bar{x}\ \ddot{y}\ \ddot{y}\ \ddot{y}\ 1{,}96\ \ddot{y}n\ \ddot{y}n\ \frac{\ddot{y}}{\ddot{y}} = 0{,}95$$

$$P\ \bar{x}\ \ddot{y}\ 1{,}96\ \ddot{y}\ \ddot{y}\ \ddot{y}\ \bar{x}\ +\ 1{,}96\ \ddot{y}n\ \ddot{y}n\ \frac{\ddot{y}}{\ddot{y}} = 0{,}95\ .$$

Machine Translated by Google

Indem wir $\bar{x}$, ÿ und *n* durch ihre jeweiligen **Werte** ersetzen , erhalten wir:

$$P\left(860 \ddot{y} 1{,}96 \ddot{y} \ddot{y} \frac{\ddot{y}}{\sqrt{30}} \ddot{y} 25\right)$$

$$+ 1{,}96 \frac{\ddot{y}}{\sqrt{30}} = 0{,}95$$

$$P(848{,}24 \ddot{y} \ddot{y} \ddot{y} \mathit{871{,}76}) = 0{,}95 .$$

Das Konfidenzintervall für ÿ beim Konfidenzniveau 0,95 beträgt daher:

$$[\mathit{848{,}24}, 871{,}76].$$

Das bedeutet, dass wir mit einer **Wahrscheinlichkeit** von 95 % bestätigen können, dass dieses **Intervall** den wahren **Wert** des **Parameters** ÿ enthält, der der durchschnittlichen Lebensdauer der Glühbirnen entspricht.

WEITERLESEN

- **Vertrauensniveau**
- **Einschätzung**
- **Schätzer**

VERWEISE

Bowley, AL: Ansprache des Präsidenten an die Wirtschaftsabteilung des British Asso Ziation. J. Roy. Stat. Soc. **69,** 540–558 (1906)

Desrosières, A. Die Partei für alles: Allgemeiner Kommentar? Schätzung und Sondagen. Fünf Beiträge zur Statistikgeschichte. Economica, Paris (1988)

Vertrauensniveau

Das Konfidenzniveau ist die **Wahrscheinlichkeit ,** um die herum das **Konfidenzintervall** konstruiert wurde. Ein **Schätzer** enthält den wahren **Wert** des entsprechenden **Parameter** der **Bevölkerung.**

Wir bezeichnen das Konfidenzniveau mit (1 ÿ ÿ), wobei ÿ dem Fehlerrisiko entspricht ; das heißt, auf die Wahrscheinlichkeit, dass das Konfidenzintervall nicht den wahren Wert des enthält Parameter.

GESCHICHTE Das erste Beispiel eines Konfidenzintervalls findet sich im Werk von Laplace (1812). Laut Desrosières, A. (1988) war Bowley, AL einer der ersten, der sich für das Konzept des Konfidenzintervalls interessierte.

Siehe **Hypothesentests.**

MATHEMATISCHE ASPEKTE

Sei ÿ ein Parameter, der einer Population zugeordnet ist. ÿ ist zu schätzen und *T* ist sein Schätzwert mator aus einer Zufallsstichprobe. Wir bewerten die Präzision von *T* als Schätzer von ÿ, indem wir ein **Konfidenzintervall** um den Schätzer herum konstruieren, das oft als Fehlermarge interpretiert wird.
Um dieses Konfidenzintervall zu konstruieren, gehen wir im Allgemeinen wie folgt vor. Aus der Verteilung des Schätzers T bestimmen wir ein Intervall, das mit hoher Wahrscheinlichkeit den wahren Wert des Parameters enthält. Bezeichnen wir dieses Intervall mit $(T \ddot{y} \ddot{y}, T + \ddot{y})$ und die Wahrscheinlichkeit, dass der wahre Wert des Parameters in diesem Intervall liegt, als (1 ÿ ÿ). Wir können dann sagen, dass die Fehlermarge ÿ durch die Wahrscheinlichkeit mit ÿ zusammenhängt:

$$P(T \ddot{y} \ddot{y} \ddot{y} \ddot{y} \ddot{y} T + \ddot{y}) = 1 \ddot{y} \ddot{y} .$$

Das mit einem Schätzintervall verbundene Wahrscheinlichkeitsniveau wird als Konfidenz bezeichnet Niveau oder Vertrauensgrad.
Das Intervall $T \ddot{y} \ddot{y} \ddot{y} \ddot{y} \ddot{y} T + \ddot{y}$ wird als Konfidenzintervall für ÿ bei der Konfidenz bezeichnet Ebene 1ÿÿ. Nehmen wir zum Beispiel ÿ = 5 %,

Machine Translated by Google

Dadurch wird das Konfidenzintervall des Parameters ÿ auf ein Wahrscheinlichkeitsniveau von 95 % festgelegt. Das bedeutet, dass, wenn wir *T* als Schätzer von ÿ verwenden, das angegebene Intervall im Durchschnitt 95 Mal von 100 Stichproben den wahren Wert des Parameters enthält und ihn 5 Mal nicht enthält.

Die Größe ÿ des Konfidenzintervalls entspricht der Hälfte der Länge des Intervalls.
Dieser Parameter gibt uns daher eine Vorstellung von der Fehlermarge des Schätzers. Für ein gegebenes Konfidenzniveau 1ÿÿ ist der Schätzer umso effizienter, je kleiner das Konfidenzintervall ist.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die am häufigsten verwendeten Konfidenzniveaus sind 90 %, 95 % und 99 %. Bei Bedarf können jedoch auch andere Ebenen verwendet werden. Obwohl wir gerne das höchste Konfidenzniveau verwenden würden, um das zu maximieren **Wahrscheinlichkeit ,** dass das Konfidenzintervall den wahren Wert des Parameters enthält , aber wenn wir das Konfidenzniveau erhöhen, vergrößert sich auch das **Intervall .** Was wir also an Vertrauen gewinnen, geht an Präzision verloren, wir müssen also einen Kompromiss finden.

BEISPIELE

Ein Unternehmen, das Glühbirnen herstellt, möchte die mittlere Lebensdauer seiner Glühbirnen untersuchen. Die Verteilung der **Zufallsvariablen** *X* , die die Lebensdauer in Stunden darstellt, ist eine **Normalverteilung** des Mittelwerts ÿ **und der Standardabweichung** ÿ = 30.

Um ÿ zu schätzen, brennt das Unternehmen eine Zufallsstichprobe **von** n = 25 Glühbirnen aus.
Das Unternehmen ermittelt eine durchschnittliche Lampenlebensdauer von $\bar{x}$ = 860 Stunden. Dann will es Konstruieren Sie ein 95 % **-Konfidenzintervall** für ÿ um den **Schätzer** $\bar{x}$.

Die Standardabweichung ÿ der Grundgesamtheit ist bekannt; der Wert von ÿ istzÿ/2 ·ÿ$\bar{X}$. Der Wert von zÿ/2 wird aus der Normaltabelle ermittelt und hängt von der Wahrscheinlichkeit ab, die dem Parameter ÿ zugeschrieben wird. Anschließend leiten wir das Konfidenzintervall des Schätzers von ÿ auf dem Wahrscheinlichkeitsniveau 1 ÿ ÿ ab:

$$\bar{X} \text{ ÿ } z\ddot{y}/2\ddot{y}\bar{X} \text{ ÿ ÿ ÿ } \bar{X} + z\ddot{y}/2\ddot{y}\bar{X} .$$

Aus der **Hypothese ,** dass die Glühbirnenlebensdauer *X* einer Normalverteilung mit Mittelwert ÿ und Standardabweichung ÿ = 30 folgt, leiten wir diesen Ausdruck ab

$$\frac{\bar{x} \text{ ÿ } \ddot{y}}{\ddot{y}\ddot{y}\sqrt{n}}$$

folgt einer Standardnormalverteilung.
Daraus erhalten wir:

$$P\ \ddot{y}z0{,}025 \text{ ÿ ÿ } z0{,}025 = 0{,}95 \quad ,$$

wobei das Fehlerrisiko **ÿ** durch zwei geteilt wird Teile, die beide gleich sind $\frac{\ddot{y}}{2}$ = 0,025.
Die Tabelle der Standardnormalverteilung (die **normale Tabelle)** ergibt z0,025 = 1,96. Wir haben dann:

$$P\ \ddot{y}1{,}96 \text{ ÿ ÿ } 1{,}96 = 0{,}95 \quad \ddot{y}\ddot{y}\ n$$

Um das Konfidenzintervall für ÿ zu erhalten, müssen wir ÿ in der folgenden Gleichung mithilfe der folgenden Transformationen isolieren:

$$P\ \ddot{y}1{,}96\ddot{y} \ \bar{x}\ \ddot{y}\ddot{y}\ddot{y} \ 1{,}96\ \ddot{y}n\ \ddot{y}n \qquad \frac{\ddot{y}}{=}$$

$$= 0{,}95 \quad ,$$

$$P\ \bar{x} \text{ ÿ } 1{,}96 \text{ ÿ ÿ } \ddot{y}\ \bar{x} \ 1{,}96\ \ddot{y}n\ \ddot{y}n \qquad \frac{\ddot{y}}{=}$$

$$= 0{,}95 .$$

Machine Translated by Google

Ersetzen Sie x¯, ÿ und *n* durch ihre jeweiligen Werte bewerten wir das Konfidenzintervall:

$$860 \ \ddot{y}\ 1{,}96 \cdot \frac{30}{\ddot{y}25}$$

$$\ddot{y}\ \ddot{y}\ \ddot{y}\ 860 + 1{,}96 \cdot \ddot{y}25 \quad \frac{30}{}$$

848,24 ÿ ÿ ÿ *871,76* .

Wir können dies mit einer Sicherheit von 95 % bestätigen. dass dieses **Intervall** den wahren Wert von enthält der **Parameter** ÿ, der dem entspricht mittlere Lebensdauer der Glühbirne.

WEITERLESEN

- **Konfidenzintervall**
- **Einschätzung**
- **Schätzer**
- **Hypothesentest**

Kontingenztabelle

Eine Kontingenztabelle ist eine Kreuztabelle, die verschiedene Attribute einer Grundgesamtheit oder enthält eine beobachtete Stichprobe. Die Kontingenttabellenanalyse besteht aus Entdecken und Studieren die Beziehungen (falls vorhanden) zwischen diesen Attribute.

Eine Kontingenztabelle kann eine zweidimensionale Tabelle mit *r* Zeilen und *c* Spalten sein Bezug auf zwei qualitative kategoriale Variablen besitzen jeweils die Kategorien *r* und *c* . Es kann auch mehrdimensional sein, wenn die Anzahl der qualitativen Variablen ist größer dann zwei: wenn zum Beispiel die Elemente von eine **Population** oder eine **Stichprobe** charakterisiert werden Durch drei Attribute hat die zugehörige kontingente Cytable die Dimensionen I×J×K, wobei *I* stellt die Anzahl der definierenden Kategorien dar das erste Attribut, *J* die Anzahl der Kategorien des zweiten Attributs und *K* die Anzahl davon die Kategorien des dritten Attributs.

GESCHICHTE

Der Begriff „Kontingenz“ wird in Bezug auf verwendet eine gekreuzte Tabelle mit kategorialen Daten, scheint es zu sein stammen von **Pearson, Karl** (1904), Wer benutzte den Begriff „Kontingenz“? ein Maß für die Gesamtabweichung **relativ** zu die **Unabhängigkeit.**

Siehe auch **Chi-Quadrat-Test der Unabhängigkeit.**

MATHEMATISCHE ASPEKTE

Wenn wir eine zweidimensionale Tabelle betrachten, die Einträge für zwei qualitative kategoriale Variablen X und *Y* enthält , die jeweils Folgendes haben: *r*- und *c*- Kategorien, die Kontingenztabelle lautet:

	Kategorien der Variablen Y			
	Y1	...	*Yc*	Gesamt
Kategorien des Variable X	X1	n11	...	n1c n1.
	...	...	... ...	...
	Xr	*nr1*	...	*nrc* *Nr.*
	Insgesamt	*Nr. 1*	...	*nc* *N..*

Wo

nij stellt die beobachtete **Häufigkeit** dar Kategorie *i* der **Variablen** *X* und Kategorie *j* der Variablen Y;

ni. stellt die Summe der Frequenzen dar beobachtet für Kategorie *i* der Variablen X,

nj stellt die Summe der beobachteten Häufigkeiten für die Kategorie jof VariableY und dar

n.. gibt die Gesamtzahl der Beobachtungen an.

Im Fall einer mehrdimensionalen Tabelle ist die Elemente der Tabelle werden mit *Nijk bezeichnet und* stellen die beobachtete Häufigkeit für die Kategorie i *der* Variablen X, die Kategorie *j* der Variablen *Y* und dar Kategorie *k* der Variablen Z.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die **Unabhängigkeit** zweier kategorialer qualitativer Variablen, die im Kontingent dargestellt werden,

Machine Translated by Google

Die Cy-Tabelle kann durch die Durchführung eines **Chi-Quadrat-Unabhängigkeitstests beurteilt werden.**

WEITERLESEN

Chi-Quadrat-Test der Unabhängigkeit

Häufigkeitsverteilung

VERWEISE

Fienberg, SE: The Analysis of Cross Classified Categorical Data, 2. Aufl. MIT Press, Cambridge, MA (1980)

Pearson, K.: Zur Kontingenztheorie und seine Beziehung zu Assoziation und Normalität Korrelation. Unternehmensforschung von Drapers Memoiren, Biometric Ser. I., S. 1–35 (1904)

Kontinuierliche Verteilung Funktion

Die **Verteilungsfunktion** einer stetigen **Eine Zufallsvariable** ist definiert als **die Wahrscheinlichkeit ,** dass die Zufallsvariable einen **Wert** annimmt kleiner oder gleich einer reellen Zahl.

GESCHICHTE

Siehe **Wahrscheinlichkeit.**

MATHEMATISCHE ASPEKTE

Die Funktion definiert durch

$$F(b) = P(X \ddot{y} b) = \int_{\ddot{y}\ddot{y}}^{B} f(x)dx .$$

heißt die **Verteilungsfunktion** eines con kontinuierliche **Zufallsvariable.**

Mit anderen Worten, die **Dichtefunktion** *f* ist die Ableitung der kontinuierlichen Verteilung Funktion.

Eigenschaften der kontinuierlichen Verteilung Funktion

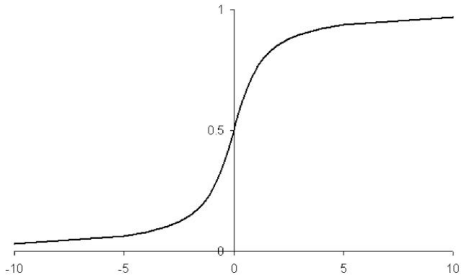
1. *F(x)* ist eine kontinuierlich wachsende Funktion für alle x;
2. *F* nimmt seine **Werte** im **Intervall** [0, 1] an;
3. $\lim_{b\ddot{y}\ddot{y}}$ 4. $F(b) = 0$;

$\lim_{b\ddot{y}\ddot{y}} F(b) = 1$;

5. *F(x)* ist stetig und differenzierbar Funktion.

Diese Verteilungsfunktion lässt sich grafisch auf einem Achsensystem darstellen. Die verschiedenen **Werte** der **Zufallsvariablen** *X* sind

Auf der Abszisse sind die entsprechenden Werte von F(x) und auf der Ordinate die entsprechenden Werte von *F(x)* aufgetragen.



DOMÄNEN UND EINSCHRÄNKUNGEN

Die **Wahrscheinlichkeit ,** dass die kontinuierliche **Zufallsvariable** *X* einen **Wert** im **Inter** annimmt **val]a,** b] für alle *a* < b, was bedeutet, dass P(a < *X* ÿ b) ist gleich *F(b)* ÿ *F(a),* wobei *F* ist die Verteilungsfunktion der Zufallsvariable fähig X.

Demonstration: Das **Ereignis** *{X* ÿ b} kann sein geschrieben als Vereinigung zweier sich gegenseitig ausschließender Ereignisse: *{X* ÿ a} und *{a* < *X* ÿ b}:

$$\{X \ddot{y} b\} = \{X \ddot{y} a\} \ddot{y} \{a < X \ddot{y} b\} .$$

Indem man die **Wahrscheinlichkeit** auf jeder Seite von ermittelt Mit der Gleichung erhalten wir:

$$P(X \ddot{y} b) = P(\{X \ddot{y} a\} \ddot{y} \{a < X \ddot{y} b\})$$
$$= P(X \ddot{y} a) + P(a < X \ddot{y} b) .$$

Machine Translated by Google

Die Summe der Wahrscheinlichkeiten ergibt sich aus der Tatsache, dass die beiden Ereignisse exklusiv sind.

Durch Subtrahieren von P(X ÿ a) auf jeder Seite erhalten wir:

$$P(a < X \ddot{y} b) = P(X \ddot{y} b) \ddot{y} P(X \ddot{y} a) .$$

Aus der Definition der **Verteilungsfunktion** erhalten wir schließlich:

$$P(a < X \ddot{y} b) = F(b) \ddot{y} F(a) .$$

BEISPIELE

Betrachten Sie eine kontinuierliche **Zufallsvariable** *X* für die die **Dichtefunktion** gegeben ist durch:

$$f(x) = \begin{cases} 1, \text{ wenn } 0 < x < 1 \\ 0 \text{ wenn nicht} \end{cases} .$$

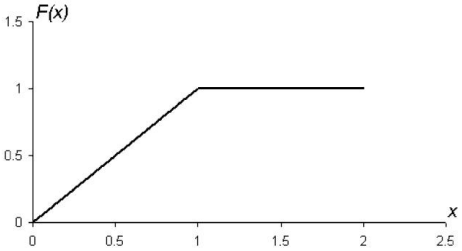
Die **Wahrscheinlichkeit ,** dass *X* einen **Wert** im **Intervall** *[a,* b] mit 0 < *a* und *b* < 1 annimmt, ist wie folgt:

$$\begin{aligned} P(a \ddot{y} X \ddot{y} b) &= \int_A^B f(x)dx \\ &= \int_A^B 1 \, dx \\ &= x \Big|_a^b \\ &= b \ddot{y} a . \end{aligned}$$

Daher lautet die Verteilungsfunktion für 0 < *x* < 1:

$$\begin{aligned} F(x) &= P(X \ddot{y} x) \\ &= P(0 \ddot{y} X \ddot{y} x) \\ &= x . \end{aligned}$$

Diese Funktion ist in der folgenden Abbildung dargestellt:



WEITERLESEN

Dichtefunktion

Wahrscheinlichkeit

Zufälliges Experiment

Zufällige Variable

Wert

Kontinuierliche Wahrscheinlichkeit Verteilung

Jede **Zufallsvariable** hat eine entsprechende **Häufigkeitsverteilung.** Für eine kontinuierliche Zufallsvariable ist diese Verteilung ebenfalls kontinuierlich.

Eine kontinuierliche Wahrscheinlichkeitsverteilung ist ein **Modell ,** das die **Häufigkeitsverteilung** einer kontinuierlichen **Variablen** darstellt

der beste Weg.

MATHEMATISCHE ASPEKTE

Die **Wahrscheinlichkeitsverteilung** einer kontinuierlichen **Zufallsvariablen** *X* ist durch ihre **Dichtefunktion** *f(x)* oder ihre **Verteilungsfunktion** *F(x) gegeben.*

Es kann im Allgemeinen durch seinen **Erwartungswert charakterisiert werden:**

$$E[X] = \int_D x \cdot f(x) \, dx = \ddot{y}$$

und seine **Varianz:**

$$\begin{aligned} \text{Var}(X) &= \int_D (x \ddot{y} \ddot{y})^2 \cdot f(x)dx \\ &= E(X \ddot{y} \ddot{y})^2 \\ &= E X2 \ddot{y} E[X]^2 , \end{aligned}$$

Machine Translated by Google

Dabei stellt *D* das **Intervall** dar, das den **Wertebereich** abdeckt, den *X* annehmen kann.

Eine wesentliche Eigenschaft einer kontinuierlichen **Zufallsvariablen** besteht darin, dass die **Wahrscheinlichkeit** , dass sie einen bestimmten numerischen **Wert** annimmt, Null ist, wohingegen die Wahrscheinlichkeit, dass sie über ein **Intervall** (endlich oder unendlich) einen Wert annimmt, normalerweise ungleich Null ist.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die bekannteste kontinuierliche Wahrscheinlichkeitsverteilung ist die **Normalverteilung.** Kontinuierliche Wahrscheinlichkeitsverteilungen werden häufig verwendet, um **diskrete Wahrscheinlichkeitsverteilungen anzunähern.** Sie kommen im Modellbau ebenso zum Einsatz wie bei der Anwendung statistischer Verfahren.

WEITERLESEN

- Beta-Verteilung
- Cauchy-Verteilung
- Chi-Quadrat-Verteilung
- Kontinuierliche Verteilungsfunktion
- Dichtefunktion
- Diskrete Wahrscheinlichkeitsverteilung
- Erwarteter Wert
- Exponentialverteilung
- Fisher-Verteilung
- Gammaverteilung
- Laplace-Verteilung
- Lognormalverteilung
- Normalverteilung
- Wahrscheinlichkeit
- Wahrscheinlichkeitsverteilung
- Zufällige Variable
- Studentenverteilung
- Gleichmäßige Verteilung
- Varianz einer Zufallsvariablen

VERWEISE

Johnson, NL, Kotz, S.: Verteilungen in der Statistik: Kontinuierliche univariate Verteilung Butionen, Bde. 1 und 2. Wiley, New York (1970)

C

Kontrast

In der Varianzanalyse ist ein Kontrast eine lineare Kombination der **Beobachtungen** oder Faktoren Niveaus oder Behandlungen in einem **faktoriellen Experiment,** wobei die Summe der Koeffizienten ist null.

GESCHICHTE

Laut Scheffé, H. (1953) war Tukey, JW (1949 und 1951) der erste, der eine Methode zur gleichzeitigen **Schätzung** aller Kontraste vorschlug.

MATHEMATISCHE ASPEKTE

Betrachten Sie T1, *T2,...,Tk,* die die Summen von n1, *n2,...,nk* **Beobachtungen sind.** Die lineare Funktion

$$cj = c1j \cdot T1 + c2j \cdot T2 + \cdots + ckj \cdot Tk$$

ist genau dann ein Kontrast, wenn

$$\sum_{i=1}^{k} ni \cdot cij = 0 .$$

Wenn jeweils $ni = n$, d. h. wenn Ti die Summe derselben Anzahl von **Beobachtungen ist,** reduziert sich die Bedingung auf:

$$\sum_{i=1}^{k} cij = 0 .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Bei den meisten **Experimenten** mit mehreren **Behandlungen** ist es für den Experimentator interessant, Vergleiche zwischen den verschiedenen Behandlungen anzustellen. Der Statistiker verwendet Kontraste, um diese Art von Vergleich durchzuführen Sohn.

Machine Translated by Google

BEISPIELE

Wenn eine **Varianzanalyse** durchgeführt wird für einen dreistufigen **Faktor** einige Kontraste von Interesse sind:

$$c1 = T1 \ddot{y} T2$$

$$c2 = T1 \ddot{y} T3$$

$$c3 = T2 \ddot{y} T3$$

$$c4 = T1 \ddot{y} 2 \cdot T2 + T3 .$$

WEITERLESEN

Varianzanalyse
Experiment
Faktorielles Experiment

VERWEISE

Ostle, B.: Statistik in der Forschung: Grundlegende Konzepte und Techniken für Forscher. Iowa State College Press, Ames, IA (1954)

Scheffé, H.: Eine Methode zur Beurteilung aller Kontraste in der Varianzanalyse.Biometri ka **40,** 87–104 (1953)

Tukey, JW: Vergleich einzelner Mittel in die Analyse der Varianz. Biometrie5, 99–114 (1949)

Tukey, JW: Schnelle und schmutzige Methoden in Statistiken. Teil II: Einfache Analysen für Standarddesigns. Quality Control Conference Papers 1951. American Society for Quality Control, New York, S. 189–197 (1951)

MATHEMATISCHE ASPEKTE

Verschiedene Arten stochastischer Konvergenz kann definiert werden. Sei {xn}nÿN eine Menge von Zufallsvariablen. Die wichtigsten Arten von stochastische Konvergenz sind:

1. {Xn}nÿ konvergiert *in der Verteilung* gegen eine Zufallsvariable *X* if

$$\lim_{n\ddot{y}\ddot{y}} FXn\,(z) = FX(z) \ddot{y}z \quad ,$$

wobei *FXn* und *FX* die Verteilung sind Funktionen von *Xn* bzw. *X.*
Diese Konvergenz ist einfach die Punktkonvergenz (in der Mathematik bekannt) von die Menge der Verteilungsfunktionen von der *Xn.*

2. {Xn}nÿ *Konvergenzwahrscheinlichkeit* gegen eine Zufallsvariable *X* , wenn:

$$\lim_{n\ddot{y}\ddot{y}} P\,(|Xn \ddot{y} X| > \ddot{y}) = 0 \quad ,$$

für jedes $\ddot{y} > 0$.

3. {Xn}nÿ weist *eine fast sichere* Konvergenz zu einer Zufallsvariablen *X* auf , wenn:

$$P\,w|\lim_{n\ddot{y}\ddot{y}} Xn(w) = X(w) = 1 \quad .$$

4. Angenommen, alle Elemente von *Xn* haben eine endliche **Erwartung.** Die *Menge {Xn}nÿ* konvergiert *im mittleren Quadrat* gegen *X* , wenn:

$$\lim_{n\ddot{y}\ddot{y}} E\,(Xn \ddot{y} X)^2 = 0 .$$

Beachten Sie, dass:

- Nahezu sichere Konvergenz und Mittelwert Quadratische Konvergenz impliziert beide eine Konvergenz der Wahrscheinlichkeit;
- Konvergenz in der Wahrscheinlichkeit (schwache Konvergenz) impliziert Konvergenz in der Verteilung.

Konvergenz

In der Statistik bezieht sich der Begriff „Konvergenz“ auf **die Wahrscheinlichkeitstheorie .** Diese statistische Konvergenz wird oft als stochastische Konvergenz bezeichnet, um sie von der klassischen zu unterscheiden Konvergenz.

Machine Translated by Google

BEISPIELE

Xi seien unabhängige Zufallsvariablen, die gleichmäßig über [0, 1] verteilt seien. Wir definieren den folgenden Satz von Zufallsvariablen aus *Xi:*

$$Zn = n \cdot \min_{i=1,\ldots,n} Xi \, .$$

Wir können zeigen, dass die Menge {Zn}nÿ in der Verteilung gegen eine **Exponentialverteilung** *Z* mit einem Parameter von 1 wie folgt konvergiert :

$$1 \ddot{y} FZn(t) = P(Zn > t)$$

$$= P \min Xi > {}_{i=1,\ldots,n} \frac{T}{N}$$

$$= P\,X1 > \frac{T}{N} \text{ und } X2 > \frac{T}{N} \text{ Und}$$

$$\ldots Xn > \frac{nt}{N}$$

$$\overset{\text{ind.}}{=} \prod_{i=1}^{N} P\,Xi > \frac{T}{N} = 1 \ddot{y} \frac{T}{N}^{N} .$$

Nun gilt für *limnÿÿ,* 1 ÿ da: $\frac{T}{N}^{N} = \exp(\ddot{y}t)$

$$\lim_{n\ddot{y}\ddot{y}} FZn = \lim_{n\ddot{y}\ddot{y}} P(Zn \ddot{y} t)$$

$$= 1 \ddot{y} \exp(\ddot{y}t) = FZ \, .$$

Schließlich sei *Sn* die Anzahl der Erfolge, die während *n* Bernoulli-Versuchen mit einer Erfolgswahrscheinlichkeit p erzielt werden . **Der Satz von Bernoulli** sagt uns, dass *Sn* in der Wahrscheinlichkeit gegen eine „zufällige“ Variable konvergiert, die den Wert *p* mit der Wahrscheinlichkeit 1 annimmt.

WEITERLESEN

- **Satz von Bernoulli**
- **Zentraler Grenzwertsatz**
- **Konvergenzsatz**
- **De Moivre-Laplace-Theorem**
- **Gesetz der großen Zahlen**
- **Wahrscheinlichkeit**
- **Zufällige Variable**
- **Stochastischer Prozess**

VERWEISE

Le Cam, LM, Yang, CL: Asymptotik in der Statistik: Einige grundlegende Konzepte. Springer, Berlin Heidelberg New York (1990)

Staudte, RG, Sheeater, SJ: Robuste Schätzung und Prüfung. Wiley, New York (1990)

Konvergenzsatz

Der Konvergenzsatz führt zu den wichtigsten theoretischen Ergebnissen der Wahrscheinlichkeitstheorie. Darunter finden wir das **Gesetz der großen Zahlen** und den **zentralen Grenzwert orem.**

BEISPIELE

Der **zentrale Grenzwertsatz** und das **Gesetz von Große Zahlen** sind beide Konvergenz Oreme.

Das Gesetz der großen Zahlen besagt, dass der **Mittelwert** einer Summe identisch verteilter Zufallsvariablen mit ihrem gemeinsamen **mathematischen Erwartungswert konvergiert.** Andererseits besagt der zentrale Grenzwertsatz, dass die Verteilung der Summe einer Suf

Eine ausreichend große Anzahl von Zufallsvariablen tendiert dazu, sich der **Normalverteilung anzunähern.**

WEITERLESEN

- **Zentraler Grenzwertsatz**
- **Gesetz der großen Zahlen**

Korrelationskoeffizient

Der einfache Korrelationskoeffizient ist ein Maß für die Stärke des linearen Zusammenhangs zwischen zwei **Zufallsvariablen.**

Machine Translated by Google

Der Korrelationskoeffizient kann **den Wert** annehmen **Werte ,** die im Intervall [ÿ1; 1]. Der zwei Extremwerte dieses Intervalls darstellen eine vollkommen lineare Beziehung zwischen den **Variablen,** „positiv“ im ersten Fall und „negativ“ im anderen. Der Wert 0 (Null) impliziert das Fehlen einer linearen Beziehung.

Der hier dargestellte Korrelationskoeffizient beträgt auch Bravais-Pearson-Korrelation genannt Koeffizient.

GESCHICHTE

Das Konzept der Korrelation hat seinen Ursprung im 1880er Jahre mit den Werken von **Galton, F..** In seinem AutobiographieMemories *of My Life* (1890), Er schreibt, dass ihm dieses Konzept bei einem Spaziergang auf dem Gelände von Naworth Castle in den Sinn kam. als ein Regenschauer ihn zwang, Schutz zu suchen.

Laut Stigler, SM (1989) handelte es sich bei Porter, TM (1986) um historische Arbeiten Recherche, als er einen vergessenen Artikel fand geschrieben von Galton im Jahr 1890 in The North American Review unter dem Titel „Verwandtschaft und Korrelation". In diesem Artikel, den er gleich nach seiner Entdeckung veröffentlichte, beschreibt Galton (1908) erklärte die Natur und die Bedeutung von das Konzept der Korrelation.

Diese Entdeckung stand im Zusammenhang mit früheren Arbeiten des Mathematikers, insbesondere die auf Vererbung und lineare Regression. Galton hatte interessiere mich seitdem für dieses Studienfach 1860. Er veröffentlichte ein Werk mit dem Titel „Natural Erbschaft“ (1889), das den Anfang bildete Punkt für seine Gedanken zur Korrelation. Im Jahr 1888 wurde in einem Artikel an die Royal Statistical Society mit dem Titel „Co-relations and their Messung hauptsächlich aus der Anthropometrie Daten“, Galton verwendete dafür den Begriff „Korrelation“. das erste Mal, obwohl er immer noch zwischen den Begriffen „Ko-Beziehung“ und „Ko-Beziehung“ schwankte „Korrelation“ und er sprach von einer „Ko-Relation“. Index." Andererseits beruft er sich auf den Betrug Konzept einer negativen Korrelation. Entsprechend Stigler (1989) schien Galton lediglich darauf hinzuweisen, dass Korrelation eine positive Beziehung sei.

Pearson, Karl schrieb 1920, dass der Zusammenhang von Galton entdeckt worden sei, dessen Arbeit „Natürliche Vererbung“ (1889) vorangetrieben ihn, zusammen mit zweien, auch dieses Konzept zu studieren andere Forscher, Weldon und Edgeworth. Pearson und Edgeworth entwickelten dann das Korrelationstheorie.

Weldon dachte an den Korrelationskoeffizienten sollte als „Galton-Funktion“ bezeichnet werden. Wie Edgeworth ersetzte Galtons Begriff „Co-Relation-Index“ und Weldons Begriff „Galton“.

Funktion“ durch den Begriff „Korrelationskoeffizient“ bezeichnet.

Nach Mudholkar (1982) systematisierte **Pearson, K.** die Korrelationsanalyse und etablierte eine Korrelationstheorie für drei **Variablen.** Forscher der Uni Universität, insbesondere sein Assistent Yule, GU, waren auch an der Entwicklung multipler Korrelationen interessiert.

Spearman veröffentlichte die erste Studie zum Rang Korrelation im Jahr 1904.

Unter den durchgeführten Arbeiten In diesem Bereich lohnt es sich, diese hervorzuheben von Yule, der in einem Artikel mit dem Titel „Why bekommen wir manchmal Unsinnskorrelationen? zwischen Zeitreihen“ (1926) diskutierte das Problem der Interpretation der Korrelationsanalyse. Schließlich war die Korrelationsrobustheit untersucht von Mosteller und Tukey (1977).

MATHEMATISCHE ASPEKTE

Einfacher linearer Korrelationskoeffizient

Man spricht hier von einfacher linearer Korrelation beschreiben eine lineare Abhängigkeit zwischen zwei **quantitative Variablen** *X* und *Y* (siehe **einfach lineare Regression).**

Machine Translated by Google

Wenn *X* und *Y* **Zufallsvariablen** sind, folgt:

Wenn eine unbekannte **gemeinsame Verteilung vorliegt,** ist der einfache lineare Korrelationskoeffizient gleich der **Kovarianz** zwischen *X*- und *Y* -Teilung berechnet durch das Produkt ihrer **Standardabweichungen:**

$$ÿ = \frac{\text{Cov}(X, Y)}{ÿX ÿY} .$$

Hier ist Cov(X, Y) die gemessene **Kovarianz** zwischen *X* und Y; *ÿX* und *ÿY* sind die jeweiligen **Standardabweichungen** von *X* und Y.

Gegeben sei eine Stichprobe der Größe n, (X1, Y1), (X2, *Y2),...,(Xn, Yn)* aus der gemeinsamen Verteilung, die Menge

$$r = \frac{\sum_{i=1}^{N} (Xi ÿ X^-)(Yi ÿ Y^-)}{\sqrt{\sum_{i=1}^{N} (Xi ÿ X^-)2 \sum_{i=1}^{N} (Yi ÿ Y^-)2}}$$

ist eine **Schätzung** von ÿ; es ist die **Sampling** -Cor-Relation.

Wenn wir *(Xi* ÿ X¯) mit *xi* und *(Yi* ÿ Y¯) mit *yi* bezeichnen , können wir diese Gleichung wie folgt schreiben:

$$r = \frac{\sum_{i=1}^{N} xiyi}{\sqrt{\sum_{i=1}^{N} x2 \sum_{i=1}^{N} y2}} .$$

Test der Hypothese

Um die **Nullhypothese** zu testen

$$H0 : ÿ = 0$$

gegen die **Alternativhypothese**

$$H1 : ÿ = 0 \quad ,$$

Wir berechnen die **Statistik** t:

$$t = \frac{R}{Sr} ,$$

wobei *Sr* die **Standardabweichung** des **Schätzers** r ist :

$$Sr = \sqrt{\frac{1 ÿ r2}{n ÿ 2}} .$$

Unter H0 folgt die **Statistik** *t* einer **Student-Verteilung mit nÿ2 Freiheitsgraden.**

Für ein gegebenes Signifikanzniveau ÿ wird H0 abgelehnt, wenn $|t| ÿ t_{ÿ/2,nÿ2}$ ist, wobei $t_{ÿ/2,nÿ2}$ der **Kritischer Wert** des in der Student-Tabelle angegebenen Tests .

Multipler Korrelationskoeffizient

Bekannt als Determina-Koeffizient

Die mit R2 bezeichnete Funktion bestimmt, ob die aus einer **multiplen linearen Regression** geschätzte Hyperebene korrekt an die Datenpunkte angepasst ist.

Der **Wert** des Mehrfachbestimmungskoeffizienten **R2** ist gleich:

$$R2 = \frac{\text{Erklärte Variation}}{\text{Totale Variation}} = \frac{\sum_{i=1}^{N} (Yˆi ÿ Y^-)^2}{\sum_{i=1}^{N} (Yi ÿ Y^-)^2} .$$

Es entspricht dem Quadrat des multiplen Korrelationskoeffizienten. Beachte das

$$0 ÿ R2 ÿ 1 \quad .$$

Im Falle einer **einfachen linearen Regression** lässt sich folgender Zusammenhang ableiten:

$$r = \text{sign}(ÿˆ 1) \sqrt{R2} \quad ,$$

Machine Translated by Google

wobei ŷˆ 1 der **Schätzer** der **Regression** ist
Koeffizient ÿ1, und er ist gegeben durch:

$$\ddot{y}\hat{}\ 1 = \frac{\sum_{i=1}^{N} (Xi \ddot{y} \bar{X})(Yi \ddot{y} \bar{Y})}{\sum_{i=1}^{N} (Xi \ddot{y} \bar{X})^2}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Wenn zwischen zwei Variablen ein linearer Zusammenhang besteht
fähig, der Korrelationskoeffizient ist gleich
1 oder ÿ1.

Eine positive Beziehung (+) bedeutet, dass die beiden Variablen variieren in die gleiche Richtung. Wenn die Einzelpersonen erzielen im ersten Fall hohe Punktzahlen Variable (zum Beispiel die **unabhängige variabel),** werden sie dazu neigen hohe Punktzahlen in der zweiten Variable erzielen (die **abhängige Variable).** Das Gegenteil ist der Fall auch wahr.

Eine negative Beziehung (ÿ) bedeutet, dass die Personen, die in der ersten Runde hohe Punktzahlen erzielen, erhalten **Die Variable** neigt dazu, niedrig zu werden punktet im zweiten und umgekehrt.

Beachten Sie, dass, wenn die Variablen unabhängig sind, die
Der Korrelationskoeffizient ist gleich Null. Der
Die gegenseitige Schlussfolgerung ist nicht unbedingt wahr.
Die Tatsache, dass zwei oder mehr Variablen miteinander in Beziehung stehen
Eine statistische Analyse reicht nicht aus, um den Schluss zu ziehen, dass ein Zusammenhang zwischen Ursache und Wirkung besteht.
Das Bestehen einer statistischen Korrelation ist
kein Beweis für Kausalität.
Statistiken liefern zahlreiche Zusammenhänge Koeffizienten. Die Wahl, wofür verwendet werden soll

Ein bestimmter Datensatz **hängt** von unterschiedlichen Daten ab Faktoren wie:

- Die Art der Skala, die zum Ausdrücken der Vari verwendet wird fähig;
- Die Art der zugrunde liegenden Verteilung (kontinuierlich oder diskret);
- Die Merkmale der Verteilung von die Scores (linear oder nichtlinear).

BEISPIELE

Die **Daten** für zwei **Variablen** *X* und *Y* sind in der folgenden Tabelle dargestellt:

Anzahl Befehl	X	Y	X = X	j = Y	xy	x²	j²
1	174	64	ÿ1,5	ÿ1,3	1,95	2,25	1,69
2	175	59	ÿ0,5	ÿ6,3	3,14	0,25	36,69
3	180	64	4,5	ÿ1,3	ÿ5,85	20,25	1,69
4	168	62	ÿ7,5	ÿ3,3	24,75	56,25	10,89
5	175	51	ÿ0,5	ÿ14,3	7,15	0,25	204,49
6	170	60	ÿ5,5	ÿ5,3	29,15	30,25	28,09
7	170	68	ÿ5,5	2,7	ÿ14,85	30,25	7.29
8	178	63	2,5	ÿ2,3	ÿ5,75	6,25	5.29
9	187	92	11,5	26,7	307,05	132,25	712,89
10	178	70	2.5	4,7	11,75	6,25	22.09
Insgesamt	1755	653			358,5	284,5	1034,1

$$\bar{X} = 175{,}5 \text{ und } \bar{Y} = 65{,}3 .$$

Wir führen nun die notwendigen Berechnungen durch um den Korrelationskoeffizienten zwischen zu erhalten die beiden Variablen. Anwendung der Formel gibt:

$$r = \frac{\sum_{i=1}^{N} xiyi}{\sqrt{\sum_{i=1}^{N} x2 \sum_{i=1}^{N} y2}} = \frac{358{,}5}{\sqrt{284{,}5 \cdot 1034{,}1}} = 0{,}66 .$$

Test der Hypothese

Wir können den geschätzten **Standard** berechnen **Abweichung** von r:

$$Sr = \sqrt{\frac{1 \ddot{y} r2}{n \ddot{y} 2}} = \sqrt{\frac{0{,}56}{10 \ddot{y} 2}} = 0{,}266 .$$

Machine Translated by Google

Die Berechnung der **Statistik** *t* ergibt:

$$t = \frac{r \ddot{y} 0}{Sr} = \frac{0{,}66}{0{,}266} = 2{,}485 .$$

Wenn wir ein Signifikanzniveau ÿ von 5 % wählen,
Der **Wert** aus der **Student-Tabelle,** t0.025,8, beträgt gleich 2,306.
Da /*t*/ = 2,485 > t0,025,8 = 2,306, die **Nullhypothese**

$$H0 : \ddot{y} = 0$$

ist abgelehnt.

WEITERLESEN

Bestimmtheitsmaß
Kovarianz
Abhängigkeit
Kendall-Rangkorrelationskoeffizient
Multiple lineare Regression
Regressionsanalyse
Einfache lineare Regression
Rangkorrelationskoeffizient nach Spearman

VERWEISE

Galton, F.: Ko-Beziehungen und ihre Messung, hauptsächlich aus der Anthropologie Daten. Proz. Roy. Soc. London. **45,** 135–145 (1888)

Galton, F.: Natürliche Vererbung. Macmillan, London (1889)

Galton, F.: Verwandtschaft und Korrelation. Norden Bin. Rev. **150,** 419–431 (1890)

Galton, F.: Erinnerungen an mein Leben. Methuen, London (1908)

Mosteller, F., Tukey, JW: Datenanalyse und Regression: Ein zweiter Kurs in Statistiken. Addison-Wesley, Reading, MA (1977)

Mudholkar, GS: Multipler Korrelationskoeffizient. In: Kotz, S., Johnson, NL (Hrsg.) Enzyklopädie der statistischen Wissenschaften, Bd. 5. Wiley, New York (1982)

Pearson, K.: Studien zur Geschichte der Statistik und Wahrscheinlichkeit. Biometrie **13,** 25–45 (1920). Nachdruck in: Pearson, ES, Kendall, MG (Hrsg.) Studies in the History of Statistics and Probability, vol. I. Grif fin, London

Porter, TM: Der Aufstieg des statistischen Denkens, 1820–1900. Princeton Universität Press, Princeton, NJ (1986)

Stigler, S.: Francis Galtons Bericht über die Erfindung der Korrelation. Stat. Wissenschaft. 4, 73–79 (1989)

Yule, GU: Zur Korrelationstheorie. J. Roy. Stat. Soc. **60,** 812–854 (1897)

Yule, GU (1926) Warum tun wir das manchmal? Unsinn-Korrelationen zwischen Zeitreihen erhalten? Eine Studie über Probenahme und Natur von Zeitreihen. J. Roy. Stat. Soc. **(2)** 89, 1–64

Korrespondenzanalyse

Die Korrespondenzanalyse ist eine **Datenanalyse** Technik, die zur Beschreibung **von Kontingenztabellen** (oder Kreuztabellen) verwendet wird. Diese Analyse erfolgt in Form einer **grafischen Darstellung** der Zusammenhänge und der „Entsprechung“ zwischen Zeilen und Spalten.

GESCHICHTE

Die theoretischen Prinzipien der Korrespondenz Die Analysen gehen auf die Arbeiten von Hartley, HO (1935) (veröffentlicht unter seinem ursprünglichen Namen Hirschfeld) und von **Fisher, RA** zurück (1940) über **Kontingenztabellen.** Sie sind erstmals im Rahmen der Inferenzstatistik vorgestellt.

Zuerst der Begriff „Korrespondenzanalyse“. erschien im Herbst 1962 und das erste Präsentation dieser Methode, auf die verwiesen wird

Machine Translated by Google

Dieser Begriff wurde von Benzécri, JP im Winter 1963. 1976 die Werke von Ben zécri, JP, der zwölf Jahre seines Lebens nachzeichnete Laborarbeiten wurden veröffentlicht und seitdem dann haben die algebraischen und geometrischen Eigenschaften dieses deskriptiven Analysewerkzeugs bekannter und verbreiteter werden.

MATHEMATISCHE ASPEKTE

Betrachten Sie eine Kontingenztabelle für zwei **kategoriale qualitative Variablen** *X* und *Y* die jeweils *r*- und *c* **-Kategorien haben:**

	Y_1	Y_2	...	Yc	Gesamt
X1	n11	n12	⋯	n1c	n1.
X2	n21	n22	⋯	n2c	n2.
...	...	...	...	...	...
Xr	*Nr.1*	*Nr.2*	⋯	*nrc*	*Nr.*
Gesamt	*Nr. 1*	*Nr. 2*	⋯	*nc*	*N..*

Wo

nij stellt die **Häufigkeit** dar , die Kategorie *i* hat der **Variablen** *X* und der Kategorie *j* der Variablen *Y* wird beobachtet,

ni. stellt die Summe der beobachteten Häufigkeiten für Kategorie *i* der Variablen *X dar,*

nj stellt die Summe der beobachteten Häufigkeiten für die Kategorie *j* der Variablen Y dar,

n.. stellt die Gesamtzahl der **Beobachtungen dar.**

Wir gehen davon aus, dass *r* ÿ c; Wenn nicht, nehmen wir die **Transponierung** der Ausgangstabelle und deren Verwendung Dies wird als **neue Kontingenztabelle umgesetzt.** Die Korrespondenzanalyse einer Kontingenztabelle mit mehr Zeilen als Spalten wird wie folgt durchgeführt:

1. Es werden Tabellen mit Zeilenprofilen *XI* und Spaltenprofilen *XJ* erstellt.
   Für eine feste Linie (Spalte) ist das Linien-(Spalten-)Profil die erhaltene Linie (Spalte).

indem Sie jedes Element in dieser Zeile (Spalte) durch die Summe der Elemente in der Zeile dividieren (Spalte).

Das Linienprofil der Zeile *i* wird erhalten durch Division jedes Termes der Zeile *i* durch *ni.,* was ist die Summe der beobachteten **Frequenzenin** die Reihe.

Die Tabelle der Reihenprofile wird erstellt durch Ersetzen jeder Zeile der **Kontingenz Tabelle** mit ihrem Profil:

	Y	Y	...	Yc	Gesamt
X1	1 n11 / n1.	2 n12 / n1.	...	n1c / n1.	1
X2	n21 / n2.	n22 / n2.	...	n2c / n2.	1
...	...	...	...	...	...
Xr	*Nr.1 / Nr.*	*Nr.2 / Nr.*	...	*nrc / Nr.*	1
Insgesamt	$n_{.1}$	$N_{.2}$	...	$n_{.C}$	*R*

Es ist auch üblich, zur Umrechnung jedes Element der Tabelle mit 100 zu multiplizieren die Begriffe in Prozentsätze umwandeln und umwandeln Summe der Terme in jeder Zeile 100 %.

Die Stützenprofilmatrix wird erstellt auf ähnliche Weise, aber dieses Mal wird jede Spalte der **Kontingenztabelle** ersetzt mit seinem Profil: Das Spaltenprofil von Spalte *j* erhält man durch Division jedes Termes von Spalte *j* durch *nj,* das ist die Summe der **Häufigkeiten,** die für die dieser Spalte entsprechende Kategorie beobachtet wurden.

	Y	Y	...	Yc	Gesamt
X1	1 n11 / *Nr. 1*	2 n12 / *Nr.2*	...	n1c / *nc*	$N_{1.}$
X2	n21 / *Nr. 1*	Nr.22 / *Nr. 2*	...	n2c / *nc*	$N_{2.}$
...	...	...	...	...	...
Xr	*Nr.1 / Nr.1*	*Nr.2 / Nr.2*	...	*nrc / nc*	$N_{R.}$
Gesamt	1	1	. . .	1	*C*

Die Tabellen der Zeilenprofile und Spaltenprofile entsprechen einer Transformation der **Kontingenztabelle ,** die zur Erstellung verwendet wird die Zeilen und Spalten vergleichbar.

Machine Translated by Google

2. Bestimmen Sie die **Trägheitsmatrix** V.

Dies geschieht auf folgende Weise: • Der gewichtete Mittelwert der *r*- Spalte Koordinaten werden berechnet:

$$gj = \sum_{i=1}^{R} \frac{ni.}{N..} \cdot \frac{nij}{ni.} = \overline{nj}\,;\ j = 1,...,c.\ N..$$

• Die *c* erhaltenen Werte *gj* werden *r* -mal in die Zeilen einer Matrix G geschrieben ; • Die Diagonalmatrix *DI* wird mit Diagonalelementen von *ni konstruiert. ;* $\overline{N..}$

• Abschließend wird die **Trägheitsmatrix** *V* nach folgender Formel berechnet:

$$V = (XI\ \ddot{y}\ G) \cdot DI \cdot (XI\ \ddot{y}\ G)\ .$$

3. Verwendung der **Matrix** M, die aus *n besteht.* $\overline{NJ}$ Terme auf ihrer Diagonale und Nullterme anderswo bestimmen wir die Matrix C:

$$C = \ddot{y}\ \overline{M} \cdot V \cdot \ddot{y}\ \overline{M}\ .$$

4. Finden Sie die **Eigenwerte** (mit *kl bezeichnet)* und die **Eigenvektoren** (mit *vl bezeichnet)* davon Matrix C.

Die *c* **Eigenwerte** *kc, kcÿ1,...,k1* (zehn in absteigender Reihenfolge geschrieben) sind die **Trägheit.** Die entsprechenden **Eigenvektoren** werden **Fakultätsachsen (oder** Trägheitsachsen) genannt .

Für jeden **Eigenwert** berechnen wir die entsprechende Trägheit, die durch die **Faktorachse erklärt wird.** Die erste Faktorachse erklärt beispielsweise:

$$\frac{100 \cdot k1}{\sum_{l=1}^{c} kl}$$ (in %) der **Trägheit.**

Auf die gleiche Weise erklären die beiden ersten **Fakultätsachsen :**

$$\frac{100 \cdot (k1 + k2)}{\sum_{l=1}^{c} kl}$$ (in %) der **Trägheit.**

Wenn wir zum Beispiel die Anzahl der **Eigenwerte** und damit die Fakultätsachsen wissen wollen, die mindestens 3/4 der Gesamtträgheit erklären, summieren wir die erklärte Trägheit von jedem der Eigenwerte, bis wir 75 % erhalten.

Anschließend berechnen wir die Hauptträgheitsachsen aus diesen Faktorachsen.

5. Die Hauptträgheitsachsen , mit *ul* bezeichnet , sind dann gegeben durch:

$$\overline{ul = M\ddot{y}1 \cdot vl\ ,}$$

was bedeutet, dass seine j-te Komponente ist:

$$ujl = \frac{\overline{nj}}{N..} \cdot vjl\ .$$

6. Anschließend berechnen wir die mit *yk* bezeichneten Hauptkomponenten , die die orthogonalen Projektionen der Zeilenkoordinaten auf die Hauptträgheitsachsen sind: die i-te **Koordinate** von

Die l-te Hauptkomponente nimmt folgenden **Wert an:**

$$yil = xi \cdot M \cdot ul\ ,$$

bedeutet, dass

$$yil = \sum_{j=1}^{c} \frac{nij}{ni.} \cdot \frac{\overline{N..}}{nj} \cdot vjl$$

ist die **Koordinate** der Zeile *i* auf der l-ten Achse.

7. Nachdem die Hauptkomponenten *yl* (der Zeilenkoordinaten) berechnet wurden, ermitteln wir mithilfe der Transaktionsformeln die Hauptkomponenten der Spaltenkoordinaten (bezeichnet mit *zl)* mithilfe von *yl :*

$$zjl = \frac{1}{\ddot{y}kl} \sum_{i=1}^{R} \frac{nij}{\cdot} \cdot yil\ ,\ nj$$

für *j = 1, 2,...,c;*

$$yil = \frac{1}{\ddot{y}kl} \sum_{i=1}^{c} \frac{nij}{ni.} \cdot zjl\ ,$$

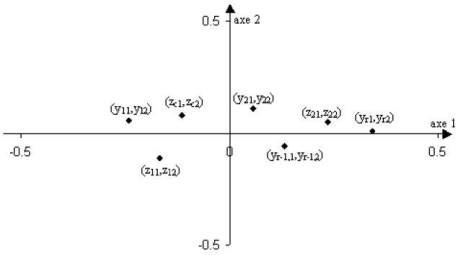
für *i = 1, 2,...,r.*

Machine Translated by Google

zjl ist die **Koordinate** der Spalte *j* auf die l-te **Fakultätsachse.**

8. Die gleichzeitige Darstellung der Zeilenkoordinaten und die Spaltenkoordinaten

Wird auf einem **Streudiagramm** mit zwei **Fakultätsachsen angezeigt und** gibt der Achse eine Bedeutung abhängig von den Punkten, auf die es sich bezieht. Die Qualität der Darstellung hängt davon ab zum Anteil der erklärten Trägheit durch die beiden verwendeten Hauptachsen. Je näher die Die erklärte **Trägheit** ist 1, je besser Qualität.



9. Es kann sinnvoll sein, zusätzliche Punktzeilen oder Spaltenkoordinaten einzufügen (zur Veranschaulichung). **Variablen)** zusammen mit den aktiven Variablen für die Korrespondenzanalyse verwendet. Betrachten Sie eine zusätzliche Zeilenkoordinate:

$$(ns1, ns2, ns3,...,nsc)$$

der Profilreihe

$$\frac{ns1}{ns.}, \frac{ns2}{ns.},..., \frac{nsc}{ns.}.$$

Die l-te Hauptkomponente des zusätzlichen Zeilenpunkts wird durch die zweite Formel von angegeben Punkt 7):

$$yil = \frac{1}{\ddot{y}kl} \sum_{j=1}^{c} \frac{nsj}{ns.} \cdot zjl ,$$

für $i = 1, 2,...,r.$

Für eine Spalte gehen wir genauso vor koordinieren, aber wenden Sie die erste Formel von an stattdessen Punkt 7.

DOMÄNEN UND EINSCHRÄNKUNGEN

Bei gleichzeitiger Darstellung der Zeile Koordinate und die Spaltenkoordinate, it Es ist nicht ratsam, eventuelle Ähnlichkeiten zwischen Zeilen und Spalten zu interpretieren wenn die beiden Punkte nicht identisch sind Anfangsraum. Andererseits ist es interessant, die Position einer Zeilenkoordinate zu interpretieren, indem man sie mit der Menge der Spalten vergleicht Koordinate (und umgekehrt).

BEISPIELE

Betrachten Sie das Beispiel eines Unternehmens, das möchte herausfinden, wie gesund seine Mitarbeiter sind. Eins Eines der im Fragebogen behandelten Themen betrifft die Anzahl der Arztbesuche (Zahnärzte eingeschlossen) pro Jahr. Drei möglich Antworten werden vorgeschlagen:

- Zwischen 0 und 6 Besuche;
- Zwischen 7 und 12 Besuchen;
- Mehr als 12 Besuche pro Jahr.

Die befragten Mitarbeiter werden auf fünf verteilt **Kategorien:**

- Führungskräfte über 40;
- Führungskräfte unter 40 Jahren;
- Mitarbeiter über 40;
- Mitarbeiter unter 40;
- Büropersonal.

Die Ergebnisse sind in der folgenden **Kontingenztabelle aufgeführt,** zu der Margen hinzugefügt wurden hinzugefügt:

	Anzahl der Besuche pro Jahr			
	0 bis 6	7 bis 12	> 12	Gesamt
Managerial Mitarbeiter > 40	5	7	3	15
Managerial Mitarbeiter < 40	5	5	2	12

Machine Translated by Google

	Anzahl der Besuche pro Jahr			
	0 bis 6	7 bis 12	> 12	Gesamt
Mitarbeiter > 40	2	12	6	20
Mitarbeiter < 40	24	18	12	54
Büro Personal	12	8	4	24
Gesamt	48	50	27	125

Anhand dieser **Daten** werden wir die Acht beschreiben
Die wichtigsten Schritte, die befolgt werden sollten, um sie zu erhalten
eine **grafische Darstellung** des Mitarbeiters
Gesundheit durch Korrespondenzanalyse.

1. Wir ermitteln zunächst die Tabelle der Leitungsprofile
MatrixXI, indem jedes Element durch dividiert wird
Summe der Elemente der Zeile, in der
es befindet sich:

0,333 0,467 *0,2* 1
0,417 0,417 0,167 1
0,1 *0,6* *0,3* 1
0,444 0,333 0,222 1
0,5 0,333 0,167 1
1,794 2,15 1,056 5

und die Tabelle der Spaltenprofile, indem jedes
Element durch die Summe der Elemente in der
entsprechenden Spalte dividiert wird:

0,104	0,14	0,111	0,355
0,104	0,1	0,074	0,278
0,042	0,24	0,222	0,504
0,5	0,36	0,444	1,304
0,25	0,16	0,148	0,558
1	1	1	3

2. Anschließend berechnen wir die Trägheitsmatrix *V*
für die **Frequenzen** (gegeben durch $\frac{ni.}{N..}$ für val
Ues von *i* von 1 bis 5), indem Sie in der fortfahren
Folgendermaßen:

- Wir beginnen mit der Berechnung der Gewichtung
Mittelwert der fünf Strichpunkte:

$$gj = \sum_{i=1}^{5} \frac{ni.}{N..} \cdot \frac{nij}{ni.} = \frac{NJ}{N..}$$

für *j = 1,* 2 und 3;

- Diese drei Werte schreiben wir dann fünf
timesinthematrixG, wie unten gezeigt:

$$G = \begin{pmatrix} 0{,}384 & 0{,}400 & 0{,}216 \\ 0{,}384 & 0{,}400 & 0{,}216 \\ 0{,}384 & 0{,}400 & 0{,}216 \\ 0{,}384 & 0{,}400 & 0{,}216 \\ 0{,}384 & 0{,}400 & 0{,}216 \end{pmatrix}.$$

- Anschließend konstruieren wir eine **Diagonalmatrix**
DI enthält das *ni.* $\frac{}{N..}$;
- Abschließend berechnen wir die **Matrix von**
Trägheit V, wie im Folgenden angegeben
Formel:

$$V = (XI - G) \cdot DI \cdot (XI - G),$$

was in diesem Fall ergibt:

$$V = \begin{pmatrix} 0{,}0175 & -0{,}0127 & -0{,}0048 \\ -0{,}0127 & 0{,}0097 & 0{,}0029 \\ -0{,}0048 & 0{,}0029 & 0{,}0019 \end{pmatrix}.$$

3. Wir definieren eine quadratische **Matrix** dritter Ordnung *M*
das enthält *n..* $\frac{}{NJ}$ Begriffe auf seiner Diagonale und
Nullbegriffe überall sonst:

$$M = \begin{pmatrix} 2{,}604 & 0 & 0 \\ 0 & 2.5 & 0 \\ 0 & 0 & 4.630 \end{pmatrix}.$$

Die Quadratwurzel von M, bezeichnet als $\sqrt{M}$, ist
erhalten, indem man jeweils die Quadratwurzel zieht
Diagonalelement von M. Mit diesem neuen
Matrix bestimmen wir

$$C = \sqrt{M} \cdot \overline{V} \cdot \sqrt{M}$$

$$C = \begin{pmatrix} 0{,}0455 & -0{,}0323 & -0{,}0167 \\ -0{,}0323 & 0{,}0243 & 0{,}0100 \\ -0{,}0167 & 0{,}0100 & 0{,}0087 \end{pmatrix}.$$

Machine Translated by Google

4. Die **Eigenwerte** von *C* werden durch Diagonalisierung der Matrix erhalten . Ordnet man diese Werte in absteigender Reihenfolge an, erhält man:

k1 = 0,0746

k2 = 0,0039

k3 = 0 .

Für jeden dieser Werte wird die erklärte Trägheit ermittelt. Zum Beispiel:

$$k1 : = \frac{0,0746}{95,03\ \% \cdot 0,0746 + 0,0039} ,$$

$$k2 : = \frac{0,0039}{4,97\ \% \cdot 0,0746 + 0,0039}$$

Für den letzten, k3, ist die erklärte Trägheit Null.

Die ersten beiden **Fakultätsachsen, die** den **Eigenwerten** k1 und k2 zugeordnet sind , erklären die gesamte Trägheit. Da der dritte Eigenwert Null ist, ist es nicht erforderlich, den zugehörigen **Eigenvektor** zu berechnen. Wir konzentrieren uns dann auf die Berechnung der ersten beiden normalisierten Eigenvektoren:

$$v1 = \begin{pmatrix} 0,7807 \\ -0,5576 \\ -0,2821 \end{pmatrix} \text{ Und}$$

$$v2 = \begin{pmatrix} 0,0807 \\ 0,5377 \\ -0,8393 \end{pmatrix} .$$

5. Anschließend berechnen wir die **Hauptträgheitsachsen** von:

$$ui = \overline{M^{-1}} \cdot vi , \qquad \text{für } i = 1 \text{ und } 2 ,$$

wobei $\overline{M^{-1}}$ durch Invertieren der Diagonalelemente von $\overline{M}$ erhalten wird.

Wir finden:

$$u1 = \begin{pmatrix} 0,4838 \\ -0,3527 \\ -0,1311 \end{pmatrix} \text{ Und}$$

$$u2 = \begin{pmatrix} 0,0500 \\ 0,3401 \\ -0,3901 \end{pmatrix} .$$

6. Anschließend berechnen wir die Hauptkomponenten, indem wir die Reihen auf die Hauptträgheitsachsen projizieren. Wir konstruieren eine **Hilfsmatrix** *U* , die aus den beiden **Vektoren** u1 und u2 gebildet wird, und definieren:

$$Y = XI \cdot M \cdot U$$

$$= \begin{pmatrix} -0,1129 & 0,0790 \\ 0,0564 & 0,1075 \\ -0,5851 & 0,0186 \\ 0,1311 & -0,0600 \\ 0,2349 & 0,0475 \end{pmatrix} .$$

Wir können die Koordinaten der fünf Reihen sehen horizontal in dieser **Matrix** geschrieben Y; Beispielsweise gibt die erste Spalte die Komponenten an, die sich auf die erste **Fakultätsachse** beziehen , und die zweite zeigt die Komponenten an, die sich auf die zweite Achse beziehen.

7. Wir verwenden dann die Koordinaten der Zeilen bis Finden Sie diejenigen der Spalten über die folgenden als Matrizen geschriebenen Übergangsformeln:

$$Z = K \cdot Y \cdot XJ ,$$

Wo:

K ist die **Diagonalmatrix** zweiter Ordnung , die $1/\sqrt{\lambda ki}$ -Terme auf ihrer Diagonale und null Terme an anderer Stelle hat;

Y ist die Transponierte der Matrix, die die Koordinaten der Zeilen enthält, und;

XJ ist die Spaltenprofilmatrix.

Machine Translated by Google

Wir erhalten die Matrix:

$$Z = \begin{matrix} 0{,}3442 \ \ddot{y} 0{,}2409 \ \ddot{y} 0{,}1658 \\ 0{,}0081 \ 0{,}0531 \ \ddot{y} 0{,}1128 \end{matrix},$$

wobei jede Spalte die Komponenten einer der drei Spaltenfarben enthält

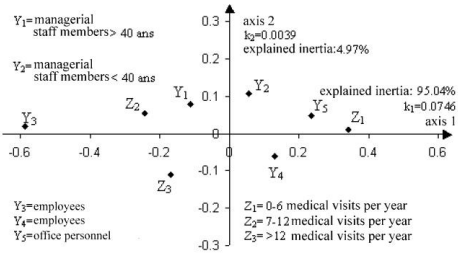
ditiert; Beispielsweise entspricht die erste Linie jeder Koordinate auf der ersten **Gesichtsachse** und die zweite Linie jeder Koordinate Koordinate auf der zweiten Achse.

Wir können die Übergangsformel überprüfen ergibt *Y* aus Z:

$$Y = XI \cdot Z \cdot K$$

unter Verwendung der gleichen Notation wie zuvor.

8. Wir können nun die fünf **Kategorien** darstellen Anzahl der befragten Personen und die drei Kategorien der darauf vorgeschlagenen Antworten Faktordiagramm:



Wir können dieses Faktordiagramm auf drei verschiedenen Analyseebenen untersuchen, abhängig von:

- Die Menge der **Kategorien** für die befragten Personen;
- Die Modalitäten für den medizinischen Besuch es ist;
- Beides gleichzeitig.

Im ersten Fall besteht die unmittelbare Nähe zwischen Zwei Zeilen **(zwischen zwei** Personalkategorien) weisen auf ähnliche Arztbesuchsprofile hin. Im faktoriellen Diagramm ist dies der Fall für Mitarbeiter unter 40 Jahren (Y4) und das Büropersonal (Y5). Wir können dies anhand der Tabelle überprüfen Linienprofile, die die **Prozentsätze** für diese angeben zwei Kategorien sind in der Tat sehr ähnlich. Ebenso die Nähe zwischen zwei Spalten (die zwei **Kategorien** darstellen , die sich auf die Anzahl der Arztbesuche beziehen) geben an ähnliche Verteilungen von Menschen innerhalb der Geschäft für diese Kategorien. Das kann sein gesehen für die Modalitäten Z2 (von 7 bis 12). Arztbesuche pro Jahr) und Z3 (mehr als 12 Besuche pro Jahr).

Wenn wir die Zeilen und Spalten betrachten gleichzeitig (und nicht getrennt wie wir). (zuvor) wird es möglich, Ähnlichkeiten zwischen Kategorien für bestimmte Modalitäten zu identifizieren. Zum Beispiel die Mitarbeiter unter 40 (Y4) und das Büropersonal (Y5) scheinen das gleiche Verhalten gegenüber zu haben Gesundheit: Hohe Anteile davon (0,44 und 0,5 bzw. 0,5) weniger als zum Arzt gehen 6 Mal im Jahr (Z1).

Abschließend steht Achse 1 auf Eins Seite mit den Kategorien, die ein durchschnittliches Alter oder eine hohe Anzahl von Besuchen anzeigen *(Z2* und Z3) – die Mitarbeiter oder Führungskräfte über 40 *(Y1* und Y3) – und auf der anderen Seite mit den Modalitäten, die mit einer geringen Anzahl verbunden sind Anzahl der Besuche *(Z1):Y2,Y4 undY5.* Der erste Faktor kann dann als Wichtigkeit interpretiert werden ärztliche Kontrolle je nach Alter.

WEITERLESEN

Kontingenztabelle
Datenanalyse
Eigenwert
Eigenvektor
Faktorielle Achse
Grafische Darstellung
Trägheitsmatrix
Matrix
Streudiagramm

Machine Translated by Google

VERWEISE

Benzécri, JP: L'Analyse des Données. Bd. 1: La Taxinomie. Bd. 2: L'Analyse Factorielle des Correspondances. Dunod, Paris (1976)

Benzécri, JP: Histoire et préhistoire de l'analyse des données, les cahiers de l'analyse des données 1, no. 1–4. Dunod, Paris (1976)

Fisher, RA: Die Präzision von Diskriminanzfunktionen.Ann.Eugen. (London)10.422–429 (1940)

Greenacre, M.: Theorie und Anwendungen der Korrespondenzanalyse. Akademisch, London (1984)

Hirschfeld, HO: Ein Zusammenhang zwischen Korrelation und Kontingenz. Proz. Camb. Philos. Soc. **31,** 520–524 (1935)

Lebart, L., Salem, A.: Analyze Statistique des Données Textuelles. Dunod, Paris (1988)

Saporta, G.: Wahrscheinlichkeit, Analyse von Données und Statistiken. Technip, Paris (1990)

Kovarianz

Die Kovarianz zwischen zwei **zufälligen Vari Die Werte** *X* und *Y* sind das Maß dafür, wie viel zwei Zufallsvariablen variieren gemeinsam. Wenn *X* und *Y* unabhängige **Zufallsvariablen sind,** ist die Kovarianz von *X* und *Y* Null. Das Gegenteil ist jedoch nicht der Fall.

MATHEMATISCHE ASPEKTE

Betrachten Sie *X* und Y, zwei **Zufallsvariablen,** die im selben **Stichprobenraum** definiert sind . Die *Kovarianz von X und Y, bezeichnet* mit Cov(X, Y), ist definiert durch

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] ,$$

wobei *E[.]* der erwartete Wert ist .

Die Entwicklung der rechten Seite der Gleichung ergibt:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[XY - E[X]Y - XE[Y] \\ &\quad + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] \\ &\quad - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] .\end{aligned}$$

Eigenschaften der Kovarianz

Betrachten Sie X, *Y* und Z, die im selben **Stichprobenraum** definierte **Zufallsvariablen** sind , , und a, b, *c* und d, die Konstanten sind. Wir finde das:

1. Cov(X, Y) = Cov(Y, X)
2. Cov(X, c) = 0 3.
   Cov(aX + *bY,* Z) = *a* Cov(X, Z) + bCov(Y, Z)
4. Cov(X, *cY* + *dZ)* = *c* Cov(X, Y) + dCov(X, Z)
5. Cov(aX + *b, cY* + d) = *ac* Cov(X, Y).

Konsequenzen der Definition 1.

Wenn *X* und *Y* unabhängige **Zufallsvariable** sind **fähig,**

$$\mathrm{Cov}(X, Y) = 0 .$$

Tatsächlich ist *E[XY] = E[X]E[Y],* was bedeutet, dass:

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0 .$$

Das Gegenteil ist im Allgemeinen nicht der Fall: Cov(X, Y) = 0 bedeutet nicht unbedingt, dass *X* und *Y* unabhängig sind.

2. Cov(X, X) = Var(X) wobei Var(X) die **Varianz** von X darstellt.

In der Tat:

$$\begin{aligned}\mathrm{Cov}(X, X) &= E[XX] - E[X]E[X] \\ &= E[X^2] - (E[X])^2 \\ &= \mathrm{Var}(X) .\end{aligned}$$

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Betrachten Sie zwei **Zufallsvariablen** *X* und Y und ihre Summe *X* + Y. Wir haben dann:

$$E[X + Y] = E[X] + E[Y] \text{ und}$$
$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) .$$

Wir zeigen nun diese Ergebnisse für diskrete Vari

fähig. Wenn $Pji = P(X = xi, Y = yj)$ gilt:

$$E[X + Y] = \sum_{J} (xi + yj)Pji .$$
$$= \sum_{J} xiPji + \sum_{ij} yjPji$$
$$= \sum_{xi} xi \sum_{Pji} Pji \sum_{yj} yj \sum_{Pji} Pji$$
$$= \sum_{J} xiPi + \sum_{J} yjPj$$
$$= E[X] + E[Y] .$$

Darüber hinaus:

$$Var(X + Y) = E[(X + Y)^2] - (E[X + Y])^2$$
$$= E[X^2] + 2E[XY] + E[Y^2] - (E[X + Y])^2$$
$$= E[X^2] + 2E[XY] + E[Y^2] - (E[X] + E[Y])^2$$
$$= E[X^2] + 2E[XY] + E[Y^2] - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2$$
$$= Var(X) + Var(Y) + 2(E[XY] - E[X]E[Y])$$
$$= Var(X) + Var(Y) + 2Cov(X, Y) .$$

Diese Ergebnisse können für *n* **Zufallsvariablen** X1, *X2,...,Xn* mit *xi* verallgemeinert werden mit einem **erwarteten Wert** gleich *E(Xi)* und einer **Varianz** gleich *Var(Xi)*. Wir haben dann:

$$E[X1 + X2 + \cdots + Xn]$$
$$= E[X1] + E[X2] + \cdots + E[Xn]$$
$$= \sum_{i=1}^{N} E[Xi] .$$

$$Var(X1 + X2 + \cdots + Xn)$$
$$= Var(X1) + Var(X2) + \cdots + Var(Xn) + 2[Cov(X1, X2) + \cdots + Cov(X1, Xn) + Cov(X2, X3) + \cdots + Cov(X2, Xn) + \cdots + Cov(Xn-1, Xn)]$$
$$= \sum_{i=1}^{N} Var(Xi) + 2 \sum_{i=1}^{n-1} \sum_{j>i} Cov(Xi, Xj) .$$

BEISPIELE

Betrachten Sie zwei nacheinander durchgeführte psychologische Tests. Jeder Proband erhält für den ersten Test eine Note *X* zwischen 0 und 3 und für den zweiten Test eine Note *Y* zwischen 0 und 2. Vorausgesetzt , dass die **Wahrscheinlichkeiten** dafür , dass

$$E[X] = 0 \cdot 0{,}16 + 1 \cdot 0{,}3 + 2 \cdot 0{,}41 + 3 \cdot 0{,}13 = 1{,}51$$
$$E[Y] = 0 \cdot 0{,}55 + 1 \cdot 0{,}32 + 2 \cdot 0{,}13 = 0{,}58$$
$$E[XY] = 0 \cdot 0 \cdot 0{,}16 \cdot 0{,}55 + 0 \cdot 1 \cdot 0{,}16 \cdot 0{,}32 + \ldots + 3 \cdot 2 \cdot 0{,}13 \cdot 0{,}13 = 0{,}88 .$$

Machine Translated by Google

Wir können dann die Kovarianz von *X* berechnen und Y:

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$
$$= 0{,}88 - (1{,}51 \cdot 0{,}58)$$
$$= 0{,}00428 .$$

WEITERLESEN

Korrelationskoeffizient

Erwarteter Wert

Zufällige Variable

Varianz einer Zufallsvariablen

Kovarianzanalyse

Die Kovarianzanalyse ist eine Methode zur Abschätzung und Prüfung der Wirkung von **Behandlungen.** Es prüft, ob zwischen den **Mitteln** mehrerer Behandlungen ein wesentlicher Unterschied besteht durch Berücksichtigung der beobachteten **Werte** der **Variablen** vor der Behandlung.

Die Kovarianzanalyse ist eine präzise Methode zur Durchführung von **Behandlungsvergleichen ,** da sie die Anpassung der Antwortvariablen *Y* an eine Begleitvariable X umfasst, die den vor der Behandlung beobachteten **Werten** entspricht.

GESCHICHTE Die Kovarianzanalyse geht auf das Jahr 1930 zurück. Sie wurde erstmals von **Fisher, RA** (1932) entwickelt. Danach wandten andere Autoren die Kovarianzanalyse auf landwirtschaftliche und medizinische Probleme an. Beispielsweise wandte Bartlett, MS (1937) die Kovarianzanalyse auf seine Studien zum Baumwollanbau in Ägypten und zu den Milcherträgen von Kühen im Winter an.

Delurry, DB (1948) verwendete Kovarianzanalyse, um die Wirkungen verschiedener Medikamente (Atropin, Chinidin, Atrophin) auf die Muskeln von Ratten zu vergleichen.

MATHEMATISCHE ASPEKTE

Wir betrachten hier eine Kovarianzanalyse eines **vollständig randomisierten Designs,** das einen **Faktor impliziert.**

Das lineare **Modell ,** das wir betrachten werden, ist das Folgendes:

$$Yij = \mu + \alpha i + \beta Xij + \varepsilon ij , \; i = 1, 2,\ldots,t, \; j = 1, 2,\ldots,ni$$

Wo

Yij stellt **die Beobachtung** j dar, **die Behandlung** i erhält , **μ** ist **der allgemeine Mittelwert** für alle Behandlungen,

αi ist die tatsächliche Auswirkung der Behandlung *i* auf die Beobachtung,

Xij ist der **Wert** der Begleitvariablen, Und

εij ist der experimentelle Beobachtungsfehler *Yij.*

Berechnungen

Um das *F* -Verhältnis zu berechnen, das uns hilft festzustellen, ob es einen signifikanten Unterschied zwischen **den Behandlungen gibt,** müssen wir Quadratsummen und Produktsummen berechnen. Wenn also $\bar{X}_{i.}$ und $\bar{Y}_{i.}$ sind jeweils die **Mittelwerte** der *X*- Werte und der *Y*- Werte für die Behandlung i, und wenn $\bar{X}_{..}$ und $\bar{Y}_{..}$ jeweils die Mittelwerte aller **Werte** von *X* und Y sind, erhalten wir die unten angegebenen Formeln.

1. Die Gesamtsumme der Quadrate für X:

$$SXX = \sum_{i=1}^{\tau} \sum_{j=1}^{ni} (Xij - \bar{X}_{..})^2 .$$

2. Die Gesamtsumme der Quadrate für *Y (SYY)* wird auf die gleiche Weise berechnet wie *SXX,* jedoch wird X durch Y ersetzt.

Machine Translated by Google

3. Die Gesamtsumme der Produkte von *X* und Y:

$$SXY = \sum_{i=1}^{T} \sum_{j=1}^{ni} (Xij - \bar{X}..)(Yij - \bar{Y}..) .$$

4. Die Summe der Quadrate der Behandlungen für X:

$$TXX = \sum_{i=1}^{T} \sum_{j=1}^{ni} (\bar{X}i. - \bar{X}..)^2 .$$

5. Die Summe der Quadrate der Behandlungen für Y *(TYY)*

wird auf die gleiche Weise berechnet *wie TXX,* jedochX wird durch Y ersetzt.

6. Die Summe der Produkte der Behandlungen von *X* und Y:

$$TXY = \sum_{i=1}^{T} \sum_{j=1}^{ni} (\bar{X}i. - \bar{X}..)(\bar{Y}i. - \bar{Y}..) .$$

7. Die Summe der Fehlerquadrate **für** X :

$$EXX = \sum_{i=1}^{T} \sum_{j=1}^{ni} (Xij - \bar{X}i.)^2 .$$

8. Die Summe der Fehlerquadrate für **Y:**

wird auf die gleiche Weise berechnet *wie EXX,* aber X wird durch Y ersetzt.

9. Die Summe der Produkte **der Fehler X und Y:**

$$EXY = \sum_{i=1}^{T} \sum_{j=1}^{ni} (Xij - \bar{X}i.)(Yij - \bar{Y}i.) .$$

Das Einsetzen geeigneter Werte und die Berechnung dieser Formeln entspricht **einer Analyse der Varianz** für X, *Y* und *XY.*

Die damit verbundenen **Freiheitsgrade** Diese verschiedenen Formeln lauten wie folgt:

1. Für die Gesamtsumme: $\sum_{i=1}^{T} ni - 1.$
2. Für die Summe der **Behandlungen:** $t - 1.$
3. Für die Summe der **Fehler:** $\sum_{i=1}^{T} ni - t.$

Die Anpassung der **Variablen** *Y* an die Begleitvariable *X* ergibt zwei neue Summen von Quadrate:

1. Die angepasste Gesamtsumme der Quadrate:

$$\text{SStot} = SYY - \frac{SXY^2}{SXX} .$$

2. Die angepasste Summe der Quadrate der Fehler:

$$\text{SSerr} = EYY - \frac{EXY^2}{EXX} .$$

Die neuen **Freiheitsgrade** für diese beiden Summen sind:

1. $\sum_{i=1}^{T} ni - 1;$
2. $\sum_{i=1}^{T} ni - t - 1,$ wobei aufgrund der Anpassung ein **Freiheitsgrad** abgezogen wird.

Die dritte angepasste Quadratsumme, die angepasste Summe der Quadrate der **Behandlungen,** beträgt gegeben von:

$$\text{SStr} = \text{SStot} - \text{SSerr} .$$

Dies hat die gleiche **Anzahl an Freiheitsgraden** $t - 1$ wie zuvor.

Kovarianzanalysetabelle

Wir haben jetzt alle notwendigen Elemente

Erstellen Sie die Kovarianzanalysetabelle. Der Summen der Quadrate dividiert durch die Anzahl der **Freiheitsgrade** geben die **Mittel** der Quadrate.

Quelle von variation	**Abschlüsse von kostenlos dom**	**Quadratsumme und von Produkten** $\sum_{i=1}^{T} x_i^2$	$\sum_{i=1}^{T} xiyi$	$\sum_{i=1}^{T} y_i^2$
Behandeln gen	$t - 1$	*TXX*	*TXY*	*TYY*
Fehler	$\sum_{i=1}^{T} ni - t$	*EXX*	*EXY*	*EYY*
Gesamt	$\sum_{i=1}^{T} ni - 1$	*SXX*	*SXY*	*SYY*

Machine Translated by Google

Hinweis: Die Zahlen im x2 und y2 Spalten dürfen nicht negativ sein; auf dem anderen Hand, die Zahlen in der Spalte $\sum_{i=1}^{T} xiyi$ kann negativ sein.

Einstellung

Quelle von var iation	Grade von Freiheit	Die Summe von Quadrate	Bedeutung von Quadrate
Behandeln gen	t ÿ 1	SStr	MCtr
Fehler	$\sum_{i=1}^{T}$ *niÿtÿ1* SSerr		MCerr
Gesamt	$\sum_{i=1}^{T}$ *ni* ÿ 2 TSS		

F Verhältnis: Testen der Behandlungen

Das *F*- **Verhältnis,** das zum Testen der **Nullhypothese** verwendet wird dass es einen signifikanten Unterschied gibt die **Mittel** der **Behandlungen** einmal angepasst zur Variablen X, ist gegeben durch:

$$F = \frac{\text{MCtr}}{\text{MCerr}} .$$

Das Verhältnis folgt **einer Fisher-Verteilung** mit tÿ1 undt $\sum_{i=1}$ **niÿtÿ1 Freiheitsgrade.**
Die **Nullhypothese**

$$H0 : ÿ1 = ÿ2 = \cdots = ÿt$$

wird auf dem Signifikanzniveau ÿ abgelehnt, wenn das *F*- Verhältnis ist größer oder gleich dem **Wert** der **Fisher-Tabelle,** mit anderen Worten, wenn:

$$F ÿ \text{Ftÿ1}, t_{\sum_{i=1}} \text{niÿtÿ1,ÿ} .$$

Es ist klar, dass wir bei der Durchführung der Kovarianzanalyse davon ausgehen, dass der ÿ- Koeffizient von Null verschieden ist. Wenn dies nicht der Fall ist, eine einfache **Varianzanalyse** reicht aus.

Test bezüglich der ÿ-Steigung

Wir würden also gerne wissen, ob es so ist ein signifikanter Effekt der begleitenden Vari vor der Anwendung der Behandlung.
Um diese **Hypothese zu testen,** gehen wir davon aus **Nullhypothese** zu sein

$$H0 : ÿ = 0$$

und die **Alternativhypothese** zu sein

$$H1 : ÿ = 0 .$$

Das *F*- Verhältnis kann ermittelt werden:

$$F = \frac{\text{E2 } XY/EXX}{\text{MCerr}} .$$

Es folgt eine **Fisher-Verteilung** mit 1 und $\sum_{i=1}^{T}$ *ni* ÿ *t* ÿ 1 **Freiheitsgrade.** Der **Die Nullhypothese** wird auf dem Signifikanzniveau ÿ abgelehnt, wenn das *F*- Verhältnis höher ist als oder gleich dem Wert in der **Fisher-Tabelle,** in mit anderen Worten, wenn:

$$F ÿ \text{F1}, t_{\sum_{i=1}} \text{niÿtÿ1,ÿ} .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Grundhypothesen , die vor Beginn einer Kovarianzanalyse aufgestellt werden müssen, sind dieselben wie die, die auch für **die Analyse** verwendet werden **Varianzanalyse** oder eine **Regressionsanalyse.** Diese sind die Hypothesen von Normalität, Homogenität (Homoskedastizität), **Varianzen** und Unabhängigkeit.

In der Kovarianzanalyse, wie in einer Analyse von Varianz, die **Nullhypothese** legt dies fest Die unabhängigen Stichproben stammen aus verschiedenen **Populationen** mit identischen **Mittelwerten.** Darüber hinaus gibt es immer Bedingungen im Zusammenhang mit einer statistischen Technik, diejenigen, die für die Kovarianzanalyse gelten, sind wie folgt folgt:

1. Die **Bevölkerungsverteilungen** müssen sein ungefähr normal, wenn nicht völlig normal.

Machine Translated by Google

2. Die Populationen, aus denen die **Proben** stammen genommen werden, müssen die gleiche **Varianz** haben ÿ2, was bedeutet:

$$ÿ2_1 = ÿ2_2 = \ldots = ÿ2_k,$$

wobei *k* die Anzahl der Populationen ist verglichen werden.

3. Die Stichproben müssen zufällig ausgewählt werden und alle Stichproben müssen unabhängig sein.

Wir müssen auch eine grundlegende **Hypothese** hinzufügen , die spezifisch für die Kovarianzanalyse ist, und zwar, dass die durchgeführte **Behandlungen** dürfen nicht erfolgen beeinflussendieWertederbegleitendenVariablen _ **fähig** X.

BEISPIELE

Stellen Sie sich ein **Experiment** vor , bei dem die Auswirkungen von drei verschiedenen Diäten verglichen werden eine **Population** von Kühen.
Die **Daten** werden in Form einer Tabelle dargestellt, Dazu gehören jeweils drei verschiedene Diäten davon wurde auf fünf Vorbauten verteilt. Der Anfangsgewichte werden durch die Begleitvariable X (in kg) und die Gewichtszunahme angegeben (nach **der Behandlung)** werden mit Y bezeichnet :

Diäten

123					
XY	XY	XY			
32	167	26	182	36	158
29	172	33	171	34	191
22	132	22	173	37	140
23	158	28	163	37	192
35	169	22	182	32	162

Wir berechnen zunächst die verschiedenen Quadratsummen und Produkte:

1. Die Gesamtsumme der Quadrate für X:

$$SXX = \sum_{i=1}^{3}\sum_{j=1}^{5} (Xij ÿ X¯..)^2$$

$$= (32 ÿ 29,8667)^2 + \ldots + (32 ÿ 29,8667)^2$$

$$= 453,73 .$$

2. Die Gesamtsumme der Quadrate für Y:

$$SYY = \sum_{i=1}^{3}\sum_{j=1}^{5} (Yij ÿ Y¯..)^2$$

$$= (167 ÿ 167,4667)^2 + \ldots + (162 ÿ 167,4667)^2$$

$$= 3885,73 .$$

3. Die Gesamtsumme der Produkte von *X* und Y:

$$SXY = \sum_{i=1}^{3}\sum_{j=1}^{5} (Xij ÿ X¯..)(Yij ÿ Y¯..)$$

$$= (32 ÿ 29,8667) \cdot (167 ÿ 167,4667) + \ldots + (32 ÿ 29,8667) \cdot (162 ÿ 167,4667)$$

$$= 158,93 .$$

4. Die Summe der Quadrate der Behandlungen für X:

$$TXX = \sum_{i=1}^{3}\sum_{j=1}^{5} (X¯i. ÿ X¯..)^2$$

$$= 5(28,2 ÿ 29,8667)^2 + 5(26,2 ÿ 29,8667)^2 + 5(35,2 ÿ 29,8667)^2$$

$$= 223,33 .$$

5. Die Summe der Quadrate der Behandlungen für Y:

$$TYY = \sum_{i=1}^{3}\sum_{j=1}^{5} (Y¯i. ÿ Y¯..)^2$$

Machine Translated by Google

$$= 5(159,6 - 167,4667)^2 + 5(174,2 - 167,4667)^2 + 5(168,6 - 167,4667)^2$$
$$= 542,53 .$$

6. Die Summe der Produkte der Behandlungen von X und Y:

$$TXY = \sum_{i=1}^{3}\sum_{j=1}^{5} (\bar{X}_{i.} - \bar{X}_{..})(\bar{Y}_{i.} - \bar{Y}_{..})$$
$$= 5(28,2 - 29,8667) \cdot (159,6 - 167,4667) + ... + 5(35,2 - 29,8667) \cdot (168,6 - 167,4667)$$
$$= -27,67 .$$

7. Die Summe der Fehlerquadrate fürX:

$$EXX = \sum_{i=1}^{3}\sum_{j=1}^{5} (X_{ij} - \bar{X}_{i.})^2$$
$$= (32 - 28,2)^2 + ... + (32 - 35,2)^2$$
$$= 230,40 .$$

8. Die Summe der Fehlerquadrate für Y:

$$EYY = \sum_{i=1}^{3}\sum_{j=1}^{5} (Y_{ij} - \bar{Y}_{i.})^2$$
$$= (167 - 159,6)^2 + ... + (162 - 168,6)^2$$
$$= 3343,20 .$$

9. Die Summe der Produkte der Fehler von X und Y:

$$EXY = \sum_{i=1}^{3}\sum_{j=1}^{5} (X_{ij} - \bar{X}_{i.})(Y_{ij} - \bar{Y}_{i.})$$
$$= (32 - 28,2)(167 - 159,6) + ... + (32 - 35,2)(162 - 168,6)$$
$$= 186,60 .$$

Die mit diesen verschiedenen Berechnungen verbundenen **Freiheitsgrade** sind wie folgt:

1. Für die Gesamtbeträge:

$$\sum_{i=1}^{3} n_i - 1 = 15 - 1 = 14 .$$

2. Für die Behandlungssummen :

$$t - 1 = 3 - 1 = 2 .$$

3. Für die Fehlersummen :

$$\sum_{i=1}^{3} n_i - t = 15 - 3 = 12 .$$

Die Anpassung der **Variablen** Y an die Begleitvariable X ergibt zwei neue Quadratsummen: 1. Die gesamte angepasste Quadratsumme:

$$SStot = SYY - \frac{SXY^2}{SXX}$$
$$= 3885,73 - \frac{158,932}{453,73}$$
$$= 3830,06 .$$

2. Die angepasste Summe der Quadrate der Fehler:

$$SSerr = EYY - \frac{EXY^2}{EXX}$$
$$= 3343,20 - \frac{186.602}{230,40}$$
$$= 3192,07 .$$

Die neuen **Freiheitsgrade** für diese beiden Summen sind:

1. $\sum_{i=1}^{3} n_i - 2 = 15 - 2 = 13;$
2. $\sum_{i=1}^{3} n_i - t - 1 = 15 - 3 - 1 = 11.$

Die angepasste Summe der Quadrate der **Behandlungen** ergibt sich aus:

$$SStr = SStot - SSerr$$
$$= 3830,06 - 3192,07$$
$$= 637,99 .$$

Machine Translated by Google

Dies hat die gleiche **Anzahl an Freiheitsgraden** wie zuvor:

$$t \ddot{y} 1 = 3 \ddot{y} 1 = 2 \quad .$$

Wir haben jetzt alle erforderlichen Elemente in um eine Kovarianzanalysetabelle zu erstellen. Die Summen der Quadrate dividiert durch die **Grade Die Freiheit** gibt die **Mittel** der Quadrate.

Quelle von variation	Abschlüsse von kostenlos dom	Quadratsumme und von Produkten $\sum_{i=1}^{3} x_i^2$	$\sum_{i=1}^{3} x_i y_i$	$\sum_{i=1}^{3} y_i^2$
Behandeln gen	2	223,33	*ÿ27,67*	543,53
Fehler	12	230,40	186,60	3343,20
Insgesamt	14	453,73	158,93	3885,73

Einstellung

Quelle von Variation	Abschlüsse von kostenlos dom	Die Summe von Quadrate	Bedeutung von Quadrate
Behandeln gen	2	637,99	318,995
Fehler	11	3192.07	290.188
Gesamt	13	3830.06	

Das *F*- **Verhältnis,** das zum Testen der **Null** verwendet wird **Hypothese, dass** es keinen signifikanten Unterschied zwischen den **Mitteln** der **Behandlungen** gibt einmal an die Variable Y angepasst , ist gegeben durch:

$$F = \frac{\text{MCtr}}{\text{MCerr}} = \frac{318.995}{290.188} = 1{,}099 \ .$$

Wenn wir ein Signifikanzniveau von ÿ = 0,05 wählen, Der Wert von *F* in der **Fisher-Tabelle** ist gleich:

$$F2,11,0,05 = 3{,}98 \ .$$

Da $F < F2,11,0,05$, können wir das nicht ablehnen **Nullhypothese** und daraus schließen wir Es gibt keinen signifikanten Unterschied zwischen den Reaktionen auf die drei Diäten, sobald die **Variable** *Y* an das Anfangsgewicht X angepasst wird.

WEITERLESEN

- **Varianzanalyse**
- **Versuchsplanung**
- **Fehlende Datenanalyse**
- **Regressionsanalyse**

C

VERWEISE

Bartlett, MS: Einige Beispiele für statistische Forschungsmethoden in der Landwirtschaft und Angewandte Biologie. J. Roy. Stat. Soc. (Ergänzung) 4, 137–183 (1937)

DeLury, DB: Die Analyse der Kovarianz. Biometrie 4, 153–170 (1948)

Fisher, RA: Statistische Methoden für Forschungsmitarbeiter. Oliver & Boyd, Edinburgh (1925)

Huitema, BE: Die Analyse der Kovarianz und Alternativen. Wiley, New York (1980)

Wildt, AR, Ahtola, O.: Analyse der Kovarianz (Sage University Papers Series zu Quantitative Anwendungen im Sozialen Wissenschaften, Aufsatz 12). Salbei, Tausend Oaks, Kalifornien (1978)

Kovariation

Gerade in der Wirtschaftswissenschaft ist es oft interessant, zwei **Zeitreihen zu vergleichen.** Da wir den **Grad der Abhängigkeit** zwischen zwei Variablen messen möchten , erinnert dies ein wenig an das Konzept der Korrelation. Allerdings in diesem Fall seit der **Zeit**

Reihen sind an eine dritte Variable gebunden, Zeit, Das Finden des **Korrelationskoeffizienten** würde nur zu einer künstlichen Beziehung führen. In der Tat, wenn zwei **Zeitreihen** betrachtet werden, *xt* und *yt,* die völlig unabhängige Phänomene darstellen und lineare Funktionen sind

Machine Translated by Google

von Zeit:

$$xt = a \cdot t + b \quad ,$$

$$yt = c \cdot t + d \quad ,$$

Wenn a, b, *c* und *d* Konstanten sind, ist es immer möglich, den Zeitfaktor zwischen den beiden Gleichungen zu eliminieren und eine funktionale Beziehung vom Typ $y = e$ zu erhalten. Diese gibt . Beziehung besagt, dass es ein lin $x + f$ Es besteht eine Ohrabhängigkeit zwischen den beiden Zeitreihen, was nicht der Fall ist.

Daher impliziert die Messung der Korrelation zwischen den Entwicklungen zweier Phänomene im Laufe der Zeit nicht die Existenz eines Zusammenhangs zwischen ihnen. Der Begriff Kovariation ist Daher wird diese Abhängigkeit anstelle einer Korrelation verwendet und mithilfe eines Kovariationskoeffizienten gemessen. Wir können unterscheiden zwischen: • dem linearen Kovariationskoeffizienten; • Der Tendenzkovariationskoeffizient.

GESCHICHTE

Siehe **Korrelationskoeffizient** und **Zeitreihen.**

MATHEMATISCHE ASPEKTE

Um zwei **Zeitreihen** *yt* und *xt zu vergleichen,* besteht der erste Schritt darin, sie in derselben Grafik darzustellen.
Allerdings ist ein visueller Vergleich im Allgemeinen schwierig. Folgende Variablenänderung **wird** durchgeführt:

$$Yt = \frac{yt \ddot{y} \bar{y}}{Sy} \quad \text{und} \quad Xt = \frac{xt \ddot{y} \bar{x}}{Sx} \ ,$$

Das sind die zentrierten und reduzierten Variablen wobei *Sy* und *Sx* die **Standardabweichungen** der jeweiligen **Zeitreihe** sind .
Wir können zwischen folgenden Kovariationskoeffizienten unterscheiden:

• *Der lineare Kovariationskoeffizient*

Die Form dieses Ausdrucks ist identisch mit der für den **Korrelationskoeffizienten** r, aber hier haben die Berechnungen das nicht dasselbe Verständnis, denn das Ziel besteht darin, die eventuelle Existenz einer Beziehung zwischen ihnen festzustellen Variationen, mit denen sie selbst in Zusammenhang stehen Zeit und um die Größenordnung zu messen

$$C = \frac{\sum_{t=1}^{N} (xt \ddot{y} \bar{x}) \cdot (yt \ddot{y} \bar{y})}{\sqrt{\sum_{t=1}^{N} (xt \ddot{y} \bar{x})2 \cdot \sum_{t=1}^{N} (yt \ddot{y} \bar{y})2}} \ .$$

Dies ergibt Werte zwischen ÿ1 und +1. Liegt er nahe bei ±1, besteht ein linearer **Zusammenhang** zwischen den zeitlichen Entwicklungen der beiden **Variablen.**
Beachte das:

$$C = \frac{\sum_{t=1}^{N} Xt \cdot Yt}{N} \ .$$

Dabei ist *n* die Anzahl der **Beobachtungen,** während *Yt* und *Xt* die zentrierten bzw. reduzierten Reihen sind, die durch eine Änderung der **Variablen** erhalten werden . • *Der Tendenzkovariationskoeffizient*
Der durch die **Mittel** ausgeübte Einfluss wird durch die Berechnung eliminiert:

$$K = \frac{\sum_{t=1}^{N} (xt \ddot{y} Txt) \cdot (yt \ddot{y} Tyt)}{\sqrt{\sum_{t=1}^{N} (xt \ddot{y} Txt)2 \cdot \sum_{t=1}^{N} (yt \ddot{y} Tyt)2}} \ .$$

Die Mittelwerte $\bar{x}$ und $\bar{y}$ wurden einfach durch die **Werte** der **säkularen Trends** *Txt* und *Tyt* jeder **Zeitreihe ersetzt.**
Der Tendenzkovariationskoeffizient *K* nimmt ebenfalls Werte zwischen ÿ1 und +1 an und je näher er ±1 kommt, desto stärker ist die Kovariation zwischen den Zeitreihen.

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Es gibt viele Beispiele für die Notwendigkeit
Vergleichen Sie zwei **Zeitreihen** in der Wirtschaftswissenschaft: für
Beispiel, wenn man die Entwicklung von vergleicht
der Preis eines Produkts zur Entwicklung von
die Menge des Produkts auf dem Markt,
oder die Entwicklung der Staatseinnahmen zu
die Entwicklung von Immobilientransaktionen. Es
Es ist wichtig zu wissen, ob es welche gibt
Art der Abhängigkeit zwischen den beiden Phänomenen,
die sich im Laufe der Zeit entwickeln: Das ist das Ziel
der Messung der Kovariation.
Der visuelle Vergleich zweier **Zeitreihen** ist ein
Wichtige Operation, die jedoch oft eine schwierige
Aufgabe ist, weil:

- Die verglichenen **Daten** können
  kommen aus sehr unterschiedlichen Domänen und
  weisen sehr unterschiedliche Größenordnungen
  auf, daher ist es vorzuziehen, die Abweichungen
  vom **Mittelwert zu untersuchen.**
- Die Höchst- und Tiefstwerte zweier Zeitreihen
  kann sehr unterschiedliche Amplituden haben; es ist
  Dann ist es vorzuziehen, die Dispersionen durch
  Rückverknüpfung der Variationen zu homogenisieren
  zur **Standardabweichung** der Zeit
  Serie.

Der visuelle Vergleich wird vereinfacht, wenn wir
Betrachten Sie die zentrierte und reduzierte **Vari**
Werte, die über die folgende Variable erhalten werden
Änderungen:

$$Yt = \frac{yt \ddot{y} \bar{y}}{Sy} \quad \text{und} \quad Xt = \frac{xt \ddot{y} \bar{x}}{Sx} .$$

Auch ähnlich wie bei der **Korrelation**
Koeffizient können nichtlineare Beziehungen bestehen
zwischen zwei Variablen, die einen *C* **-Wert** ergeben
das ist nahe Null.

Es ist daher wichtig, währenddessen vorsichtig zu sein
Deutung.
Der Tendenzkovariationskoeffizient wird vorzugsweise
verwendet, wenn die Beziehung zwischen
Zeitreihen sind nicht linear.

BEISPIELE

Lassen Sie uns die Kovariation zwischen zwei untersuchen
Zeitfolgen.

Die **Variable** *xt* stellt die Jahresproduktion eines
Agrarprodukts dar; Die Variable *yt* ist der
durchschnittliche jährliche Preis pro Einheit in
konstanter Euro.

T	xt	yt
1	320	5.3
2	660	3.2
3	300	2.2
4	190	3.4
5	320	2.7
6	240	3.5
7	360	2,0
8	170	2.5

$$\sum_{t=1}^{8} xt = 2560 \quad , \qquad \sum_{t=1}^{8} yt = 24{,}8 \quad ,$$

was $\bar{x} = 320$ und $\bar{y} = 3{,}1$ ergibt .

xt	x	(xt	x)2	yt	j	(yt	j)2
	0		02.2	4.84			
	340		115600		0,1		0,01
	ÿ20		400		*ÿ0,9*		0,81
	ÿ130		16900		0,3		0,09
	0		0		*ÿ0,4*		0,16
	ÿ80		6400		0,4		0,16
	40		1600		*ÿ1,1*		1.21
	ÿ150		22500		0,6		0,36

$$\sum_{t=1}^{8} (xt \ddot{y} \bar{x}) \; 2 = 163400 \quad ,$$

$$\sum_{t=1}^{8} (yt \ddot{y} \bar{y}) \; 2 = 7{,}64 \quad ,$$

was $\ddot{y}x = 142{,}9$ und $\ddot{y}y = 0{,}98$ ergibt .
Die zentrierten und reduzierten **Werte** *Xt* und *Yt*
werden dann berechnet.

Machine Translated by Google

Xt	Yt	Xt · Yt	Xt · Yt ÿ1
0,00	2,25	0,00	5.36
2,38	0,10	0,24	*ÿ0,01*
ÿ0,14	*ÿ0,92*	0,13	0,84
ÿ0,91	0,31	*ÿ0,28*	0,00
0,00	*ÿ0,41*	0,00	0,23
ÿ0,56	0,41	*ÿ0,23*	0,11
0,28	*ÿ1,13*	*ÿ0,32*	1.18
ÿ1,05	*ÿ0,61*	0,64	

Der lineare Kovariationskoeffizient ist dann cal kultiviert:

$$C = \frac{\sum_{t=1}^{8} Xt \cdot Yt}{8} = \frac{0,18}{8} = 0,0225 .$$

Wenn die **Beobachtungen** *Xt* verglichen werden mit Ytÿ1, was die Produktion in diesem Jahr bedeutet mit dem Vorjahr verglichen wird, wir erhalten:

$$C = \frac{\sum_{t=2}^{8} Xt \cdot Yt\ddot{y}1}{8} = \frac{7.71}{8} = 0,964 .$$

Der lineare Kovariationskoeffizient für eine Verschiebung von einem Jahr ist sehr stark. Es gibt eine starke (positive) Kovariation mit einer Verschiebung um ein Jahr zwischen den beiden **Variablen.**

WEITERLESEN

Korrelationskoeffizient

Gleitender Durchschnitt

Säkularen Trend

Standardabweichung

Zeitfolgen

VERWEISE

Kendall, MG: Zeitreihen. Griffin, London (1973)

Py, B.: Statistique Déscriptive. Economica, Paris (1987)

Cox, David R.

Cox, David R. wurde 1924 in Birmingham in England geboren. Er studierte Mathematik an der Universität Cambridge und promovierte 1949 in angewandter Mathematik an der University of Leeds.

Von 1966 bis 1988 war er Professor für Statistik am Imperial College London und Anschließend unterrichtete er von 1988 bis 1994 in Nuffield Hochschule, Oxford.

Cox, David ist ein hervorragender Statistiker. Er war 1982 von Königin Elizabeth II. zum Ritter geschlagen Dankbarkeit für seine Beiträge zur Statistik Wissenschaft und wurde von vielen Universitäten in England zum Doctor Hon oris Causa ernannt und anderswo. Von 1981 bis 1983 war er Präsident der Royal Statistical Society; er war auch Präsident der Bernoulli Soci ety von 1972 bis 1983 und Präsident des Internationales Statistisches Institut von 1995 bis 1997.

Aufgrund der Vielfalt der Fächer, die er studiert und entwickelt hat, hatte Professor Cox die Wahl ein tiefgreifender Einfluss auf sein Fachgebiet. Er wurde genannt Ehrendoktorwürde der Universität Neuenburg im Jahr 1992.

Cox, Sir David ist der Autor und Co-Autor von mehr als 250 Artikeln und 16 Büchern, und zwischen 1966 und 1991 war er Herausgeber von Biometrie.

Einige Hauptwerke und Artikel von Cox, Herr David:

1964 (und Box, GEP) Eine Analyse von Transformationen (mit Diskussion) J. Roy. Stat. Soc. Ser. B 26, 211–243.

1970 Die Analyse binärer Daten. Methuen, London.

Machine Translated by Google

1973 (und Hinkley, DV) Theoretisch
Statistiken. Chapman & Hall, London.

1974 (und Atkinson, AC) Planungsexperimente zur
Unterscheidung zwischen
Modelle. J. Roy. Stat. Soc. Ser. B 36,
321–348.

1978 (und Hinkley, DV) Probleme und
Lösungen in der theoretischen Statistik.
Chapman & Hall, London.

1981 (und Snell, EJ) Angewandte Statistik:
Prinzipien und Beispiele. Chapman
& Saal.

1981 Theorie und allgemeine Prinzipien in
Statistiken. J. Roy. Stat. Soc. Ser. A
144, 289–297.

2000 (und Reid, N.) Theory of Design
Experimente. Chapman & Hall, London.

Cox, Mary Gertrude

Cox, Gertrude Mary wurde 1900 geboren,
in der Nähe von Dayton, Iowa, USA. Ihr Ehrgeiz war
Menschen zu helfen, und so studierte sie zunächst
zwei Jahre lang ein sozialwissenschaftliches Studium.
Dann arbeitete sie im Waisenhaus für junge Menschen
Jungen in Montana für zwei Jahre. In Ordnung
um Direktorin des Waisenhauses zu werden, sie
beschloss, ihre Ausbildung in Iowa fortzusetzen
State College. Sie hat ihren Abschluss in Iowa gemacht
State College im Jahr 1929. Um ihr Studium zu
finanzieren, arbeitete Cox, Gertrude mit **Snedecor,
George Waddel,** ihr Professor, in seinem statistischen
Labor, was dazu führte, dass sie sich für Statistik
interessierte. Nach dem Schulabschluss,
Sie begann ein Doktoratsstudium in Psychologie. Im
Jahr 1933, bevor sie ihren Doktortitel abschloss,
arbeitete Snedecor, George, der damalige Direktor von
Er arbeitete am Iowa State Statistical Laboratory und
überzeugte sie, seine Assistentin zu werden
sie stimmte zu, obwohl sie in der Wohnung blieb
Bereich der Psychologie, weil sie daran gearbeitet hat
die Auswertung statistischer Tests in der Psychologie
und die Analyse psychologischer
Daten.

Am 1. November 1940 wurde sie zur
Direktor der Abteilung für Experimentelles
Statistiken des Bundesstaates North Carolina.

Im Jahr 1945 gab das General Education Board nach
ihre Erlaubnis, ein Institut für Statistik an der University
of North Carolina zu gründen, mit
eine Abteilung für mathematische Statistik an der
Kapellenhügel.
Sie war Gründungsmitglied des Journal of
der International Biometric Society im Jahr 1947,
und sie war dort von 1947 bis 1947 Direktorin
1955 und dessen Präsident von 1968 bis 1969.
Sie war 1956 Präsidentin der American Statistical
Association (ASA). Sie starb im Jahr
1978.

Hauptwerk von Cox, M. Gertrude:

1957 (und Cochran, W.) Experimentell
Entwürfe, 2. Aufl. Wiley, New York

Vgl Kriterium

Das Cpciterion ist ein Modellauswahlzitat
in linearer Regression. Für eine lineare Regression
Modell mit *p* Parametern einschließlich eines beliebigen
konstanten Termes, im Modell gilt eine Faustregel
um ein Modell auszuwählen, in dem der Wert von *Cp* ist
nahe an der Anzahl der Terme im Modell.

GESCHICHTE

Eingeführt von Mallows, Colin L. im Jahr 1964
Es wurde das Modellauswahlkriterium *Cp* verwendet
seitdem als Kriterium für die Bewertung der
Güte der Anpassung eines Regressionsmodells.

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Sei $Yi = ÿ0 + \sum_{j=1}^{pÿ1} Xjiÿj + ÿi, \quad ich = 1,...,n$ eine **multiple lineare Regression** Modell. Bezeichnen Sie den **mittleren quadratischen Fehler** als *MSE(yi)*. Das darin eingeführte Kriterium

Im Abschnitt können Sie das Modell auswählen, mit dem Sie auswählen möchten die minimale Summe der mittleren quadratischen Fehler:

$$MSE\ (yi) = E\ (yi ÿ\ yi)^2$$

$$ÿ\ ÿ2\ (1 ÿ\ 2hii)$$

$$= E\ (yi ÿ\ yi)^2$$

$$ÿ\ ÿ2(1 ÿ\ 2hii)$$

$$= E(RSS)\ ÿ\ ÿ2(n ÿ\ 2p)\ ,$$

Wo

n ist die Anzahl der Beobachtungen,

p ist die Anzahl der geschätzten Parameter,

hii sind die diagonalen Elemente des **Hutes Matrix** und

yi ist der Schätzer von *yi*.

Erinnern Sie sich an die folgende Eigenschaft des *hii:*

$$hii = p\ .$$

Definieren Sie den Koeffizienten

$$P = \frac{MSE\ (yi)}{ÿ2}$$

$$= \frac{E\ (RSS)\ ÿ\ ÿ2\ (n ÿ\ 2p)}{ÿ2}$$

$$= \frac{E\ (RSS)}{ÿ2}\ ÿ\ n + 2p\ .$$

Wenn das Modell korrekt ist, müssen wir Folgendes haben:

$$E\ (RSS) = (n ÿ\ p)\ ÿ2\ ,$$

was impliziert

$$P = p\ .$$

In der Praxis schätzen wir P von

$$Cp = ÿ\ \frac{RSS}{ÿ2}\ ÿ\ n + 2p\ ,$$

wobei ÿ2 ein Schätzer von ÿ2 ist . Hier schätzen wir ÿ2 anhand des s2 des vollständigen Modells. Für Mit diesem vollständigen Modell erhalten wir tatsächlich $Cp = p$, Das ist kein interessantes Ergebnis. Allerdings für alle anderen Modelle, die wir verwenden nur *eine Teilmenge* der erklärenden Variablen, die Der Koeffizient Cp kann von p verschiedene Werte haben. Von den Modellen, die integrieren nur eine Teilmenge der erklärenden Variablen, wir Wählen Sie dann diejenigen aus, für die der Wert von *Cp* gilt ist p am nächsten .

Wenn wir *k* erklärende Variablen haben, können wir Definieren Sie auch den Koeffizienten Cp für ein Modell, das enthält eine Teilmenge *X1,...,Xpÿ1, p* ÿ *k* von die *k* erklärenden Variablen im Folgenden Benehmen:

$$Cp = \frac{(n ÿ\ k)\ RSS\ X1,...,Xpÿ1}{RSS\ (X1,...,Xk)}$$

$$ÿ\ n + 2p\ ,$$

wobei RSS *X1,...,Xpÿ1* die Summe der ist Quadrate der Residuen, die sich auf das Modell mit *p* ÿ 1 erklärenden Variablen beziehen, und RSS *(X1,...,Xk) ist* die Summe der Quadrate von die auf das Modell bezogenen Residuen mit allen *k* erklärende Variablen. Von den beiden Modellen die pÿ1 erklärende Variablen beinhalten , wir wählen, je nach Kriterium, das Eine aus für den der Wert des Koeffizienten Cp ist am nächsten an p.

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Cp- Kriterium wird bei der Auswahl von Variablen verwendet. Bei Verwendung mit dem R2 -Kriterium ist dies Kriterium kann uns etwas über die Güte von sagen Passform des gewählten Modells. Die zugrunde liegende Idee eines solchen Verfahrens ist folgendes: statt

Machine Translated by Google

zu versuchen, eine Variable mit allen zu erklären
der verfügbaren erklärenden Variablen ist es
Manchmal ist es möglich, ein zugrunde liegendes
Modell mit einer Teilmenge dieser Variablen zu bestimmen.
und die Erklärungskraft dieses Modells ist
fast das gleiche wie das Modell enthalten
ing alle erklärenden Variablen. Ein weiterer Grund
dafür ist die Kollinearität von
die erklärenden Variablen nehmen tendenziell ab
Die Genauigkeit des Schätzers kann daher nützlich sein
einige oberflächliche Variablen zu löschen.

BEISPIELE

Betrachten Sie einige Daten im Zusammenhang mit Chicago
Brände von 1975. Wir bezeichnen die Variable cor
auf den Logarithmus der Zahl reagieren
von Bränden pro 1000 Haushalte pro Bezirk *i* von
Chicago im Jahr 1975 nach Y und die Variablen,
die dem Anteil der Haushalte entsprechen
gebaut vor 1940, auf die Anzahl
Diebstähle und zum mittleren Umsatz pro Bezirk
i durch X1, X2 bzw. X3 .
Da dieser Satz drei erklärende Variablen enthält,
haben wir 23 = 8 mögliche Modelle. Wir
Teilen Sie die acht möglichen Gleichungen in vier
Sätze:

1. Satz *A* enthält die einzige Gleichung ohne
   erklärende Variablen:

$$Y = ÿ0 + ÿ .$$

2. Satz *B* enthält drei Gleichungen mit einer
   erklärende Variable:

$$Y = ÿ0 + ÿ1X1 + ÿ$$

$$Y = ÿ0 + ÿ2X2 + ÿ$$

$$Y = ÿ0 + ÿ3X3 + ÿ .$$

3. Satz *C* enthält drei Gleichungen mit zwei
   erklärende Variablen:

$$Y = ÿ0 + ÿ1X1 + ÿ2X2 + ÿ$$

$$Y = ÿ0 + ÿ1X1 + ÿ3X3 + ÿ$$

$$Y = ÿ0 + ÿ2X2 + ÿ3X3 + ÿ .$$

4. Satz *D* enthält eine Gleichung mit drei
   erklärende Variablen:

$$Y = ÿ0 + ÿ1X1 + ÿ2X2 + ÿ3X3 + ÿ .$$

Bezirk	X_1	X_2	X_3	Y
„	xi 1	xi 2	xi 3	yi
1	0,604	29	11.744	1.825
2	0,765	44	9.323	2.251
3	0,735	36	9.948	2.351
4	0,669	37	10.656	2.041
5	0,814	53	9.730	2.152
6	0,526	68	8.231	3.529
7	0,426	75	21.480	2.398
8	0,785	18	11.104	1.932
9	0,901	31	10.694	1.988
10	0,898	25	9.631	2.715
11	0,827	34	7.995	3.371
12	0,402	14	13.722	0,788
13	0,279	11	16.250	1.740
14	0,077	11	13.686	0,693
15	0,638	22	12.405	0,916
16	0,512	17	12.198	1.099
17	0,851	27	11.600	1.686
18	0,444	9	12.765	0,788
19	0,842	29	11.084	1.974
20	0,898	30	10.510	2.715
21	0,727	40	9.784	2.803
22	0,729	32	7.342	2.912
23	0,631	41	6.565	3.589
24	0,830	147	7.459	3.681
25	0,783	22	8.014	2.918
26	0,790	29	8.177	3.148
27	0,480	46	8.212	2.501
28	0,715	23	11.230	1.723
29	0,731	4	8.330	3.082
30	0,650	31	5.583	3.073
31	0,754	39	8.564	2.197
32	0,208	15	12.102	1.281
33	0,618	32	11.876	1.609
34	0,781	27	9.742	3.353

Machine Translated by Google

Bezirk	X_1	X_2	X_3	Y
ich	xi 1	xi 2	xi 3	yi
35	0,686	32	7.520	2.856
36	0,734	34	7.388	2.425
37	0,020	17	13.842	1.224
38	0,570	46	11.040	2.477
39	0,559	42	10.332	2.351
40	0,675	43	10.908	2.370
41	0,580	34	11.156	2.380
42	0,152	19	13.323	1.569
43	0,408	25	12.960	2.342
44	0,578	28	11.260	2.747
45	0,114	3	10.080	1.946
46	0,492	23	11.428	1.960
47	0,466	27	13.731	1.589

Quelle: Andrews und Herzberg (1985)

Die resultierenden Modelle bezeichnen wir wie folgt: **1** für X1, **2** für X2, **3** für X3, **12** für X1 und X2, **13** für X1 und X3, **23** für X2 und X3 und **123** für das Vollmodell. Die folgende Tabelle stellt die Ergebnisse dar, die für erhalten wurden das *Cp*- Kriterium für jedes Modell:

Modell	Vgl
1	32.356
2	29.937
3	18.354
12	18.936
13	14.817
23	3.681
123	4.000

Wenn wir ein Modell aus Gruppe *B* betrachten (das eine erklärende Variable enthält), beträgt die Anzahl der geschätzten Parameter $p = 2$ und keiner der *Cp*- Werte für die drei Modelle Ansätze 2. Betrachten wir nun ein Modell aus Gruppe *C* (mit zwei erklärenden Variablen) ist *p* gleich 3 und der *Cp* von Modell **23** nähert sich diesem. Abschließend noch einmal zur Vollständigkeit Modell finden wir, dass $Cp = 4$ ist auch die Anzahl der geschätzten Parameter, aber Dies ist, wie bereits erläutert, kein interessantes Ergebnis. Daher das Vernünftigste Auswahl für das Modell scheint zu sein:

$$Y = ÿ0 + ÿ2X2 + ÿ3X3 + ÿ .$$

WEITERLESEN

Bestimmtheitsmaß

Kollinearität

Hutmatrix

Mittlere quadratische Fehler

Regressionsanalyse

VERWEISE

Andrews DF, Hertzberg, AM: Daten: Eine Sammlung von Problemen von vielen Felder für Studierende und wissenschaftliche Mitarbeiter. Springer, Berlin Heidelberg New York (1985)

Mallows, CL: Auswahl von Variablen in einer linearen Regression: Eine grafische Hilfe. Präsentiert auf der Zentralen Regionaltagung der Institut für Mathematische Statistik, Man Hattan, KS, 7.–9. Mai 1964 (1964)

Mallows, CL: Einige Kommentare zu *Cp.* Technometrics **15,** 661–675 (1973)

Cramér, Harald

Cramér, Harald (1893–1985) betrat die Universität Stockholm im Jahr 1912 in Auftrag gegeben Chemie und Mathematik studieren; Er wurde Schüler von Leffler, Mittag und Riesz, Marcel. Im Jahr 1919 wurde Cramér benannt Assistenzprofessor an der Universität Stockholm. Gleichzeitig arbeitete er als Aktuar für eine Versicherungsgesellschaft, Svenska Lebensversicherung, die ihm das Studium ermöglichte Wahrscheinlichkeit und Statistik.

Machine Translated by Google

Sein Hauptwerk in der Versicherungsmathematik ist *Kollektive Risikotheorie.* 1929 wurde er gebeten, eine neue Abteilung in Stockholm zu gründen, und er wurde der erste schwedische Professor für versicherungsmathematische und statistische Mathematik matik. Am Ende des Zweiten Weltkriegs

Er verfasste sein Hauptwerk *„Mathematical Methods of Statistics“,* das erstmals 1945 veröffentlicht und kürzlich (1999) erneut veröffentlicht wurde.

Einige Hauptwerke und Artikel von Cramér, Harald

1946 Mathematische Methoden der Statistik.
Princeton University Press, Princeton, NJ.

1946 Kollektive Risikotheorie: Ein Überblick über die Theorie aus der Sicht der Theorie stochastischer Prozesse.
Skandia Jubilee Volume, Stockholm.

Kriterium der Gesamtsumme

Mittlere quadratische Fehler

Das Kriterium des gesamten mittleren quadratischen Fehlers ist eine Möglichkeit, Schätzungen der Parameter eines voreingenommenen oder unvoreingenommenen Modells zu vergleichen.

MATHEMATISCHE ASPEKTE

Lassen

$$\ddot{y} = \ddot{y}\ 1,\ldots,\ \ddot{y}\ p\ddot{y}1$$

Ein Vektor von Schätzern für die Parameter eines Regressionsmodells. Wir definieren den gesamten mittleren quadratischen Fehler (TMSE) des Vektors $\ddot{y}$ von Schätzern als die Summe der mittleren quadratischen Fehler *(MSE)* seiner Komponenten.

Wir erinnern uns daran

$$MSE\ \ddot{y}\ j = E\ \ddot{y}\ j\ \ddot{y}\ \ddot{y}j^2$$
$$= V\ \ddot{y}\ j + E\ \ddot{y}\ j\ \ddot{y}\ \ddot{y}j^2,$$

Dabei sind *E (.)* und *V (.)* die üblichen Symbole für den **Erwartungswert** und die **Varianz.** Wir definieren den gesamten mittleren quadratischen Fehler als:

$$TMSE(\ddot{y}\ j) = \sum_{j=1}^{p\ddot{y}1} MSE\ \ddot{y}\ j$$
$$= \sum_{j=1}^{p\ddot{y}1} E\ \ddot{y}\ j\ \ddot{y}\ \ddot{y}j^2$$
$$= \sum_{j=1}^{p\ddot{y}1} \text{Var}\ \ddot{y}\ j + E\ \ddot{y}\ j\ \ddot{y}\ \ddot{y}j^2$$
$$= (p\ \ddot{y}\ 1)\ddot{y}2 \cdot \text{Spur}\ \mathbf{(V)} + \sum_{j=1}^{p\ddot{y}1} E\ \ddot{y}\ j\ \ddot{y}\ \ddot{y}j^2.$$

wobei Vis die **Varianz-Kovarianz-Matrix ist** von $\ddot{y}$ **.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Leider wollen wir den gesamten mittleren quadratischen Fehler berechnen

$$TMSE\ \ddot{y} = \sum_{j=1}^{p\ddot{y}1} E\ \ddot{y}\ j\ \ddot{y}\ \ddot{y}j^2$$

eines Vektors von Schätzern

$$\ddot{y} = \ddot{y}\ 1,\ldots,\ \ddot{y}\ p\ddot{y}1$$

für die Parameter des Modells

$$Yi = \ddot{y}0 + \ddot{y}1Xs_{i1} + \ldots + \ddot{y}p\ddot{y}1Xs_{ipy1} + \ddot{y}i\ ,$$

Wir müssen die Werte des Modells kennen Parameterÿj, die offensichtlich für alle realen Daten unbekannt sind. Die Notation *Xs* bedeutet, dass *j* die Daten für die j-te erklärende Variable standardisiert wurden (siehe **standardisierte Daten).** Um diese *TMSE* abzuschätzen , generieren wir daher auf ähnliche Weise Daten aus einer Struktur

Machine Translated by Google

Modell. Verwendung eines Pseudozufallsgenerators Zahlen können wir alle Daten simulieren Analysieren Sie das künstliche Modell mit verschiedenen Regressionsmodellen (z. B. einfach Regression oder **Ridge-Regression)** und berechnen Sie, was wir den *gesamten quadratischen Fehler nennen. TSE,* der Vektoren der Schätzer **ÿ** bezogen auf jede Methode:

$$TSE(\ddot{y}\hat{}) = \sum_{j=1}^{p\ddot{y}1} (\ddot{y}\, j\, \ddot{y}\, \ddot{y}j)^{2} .$$

Wir wiederholen diesen Vorgang 100 Mal und stellen so sicher, dass die 100 Datensätze pseudounabhängig sind. Für jedes Modell der Durchschnitt von 100 *TSE* gibt eine gute Schätzung *TMSE*. Beachten Sie, dass einige Statistiker dies bevorzugen das durch Auswahl von Variablen erhaltene Modell aufgrund seiner Einfachheit. Andererseits andere Statistiker bevorzugen die Ridge-Methode, weil Es nutzt alle verfügbaren Informationen.

BEISPIELE

Im Allgemeinen können wir Folgendes vergleichen Regressionsmethoden: **lineare Regression** nach mittlere Quadrate, Gratregression oder die Variablenauswahlmethode.

Im folgenden Beispiel dreizehn Portionen von Zement wurden untersucht. Jede Portion besteht aus vier Zutaten, die in der Tabelle angegeben sind Tisch. Ziel ist es herauszufinden, welchen Einfluss die Mengen xi1, xi2, *xi3* und *xi4* dieser vier Zutaten auf die Menge *yi*, die abgegebene Wärme, haben aufgrund der Aushärtung des Zements aus. Vom Zement abgegebene Wärme

Portion	Zutat 1234				Hitze
ich	xi 1	xi 2	xi 3	xi 4	yi
1	7	26	6 60		78,5
2	1	29 15 52			74,3
3	11	56	8 20 104,3		
4	11	31	8 47		87,6
5	7	52	6 33		95,9
6	11	55	9 22 109.2		
7	3	71	17	6 102,7	
8	1	31 22 44			72,5
9	2	54	18 22		93.1
10	21	47	4 26 115,9		
11	1	40 23 34			83,9
12	11	66	9 12 113.3		
13	10	68	8 12 109,4		

yi Wärmemenge, die aufgrund der abgegeben wird Aushärtung des i-ten Anteils (in Joule);
xi1 Menge der Zutat 1 (Tricalciumaluminat) in der i-ten Portion;
xi2 Menge der Zutat 2 (Tricalciumsilikat) in der i-ten Portion;
xi3 Menge der Zutat 3 (Tetracalcium). Alumino-Ferrit) im i-ten Teil;
xi4 Menge der Zutat 4 (Dicalciumsilikat) in der i-ten Portion.

In diesem Abschnitt vergleichen wir die durch die Methode der **kleinsten Quadrate** (LS) ermittelten Schätzer. Regression mit denen von Ridge Regression (R) und die mit der erhaltenen Variablenauswahlmethode (SV) über die Summe mittlerer quadratischer Fehler *TMSE*. Die drei aus jeder Methode erhaltenen Schätzvektoren Sind:

$$YLS = 95{,}4 + 9{,}12X1^{s} + 7{,}94X2^{s} + 0{,}65X3^{s} \ddot{y}\ 2{,}41X4^{s}$$

$$YR = 95{,}4 + 7{,}64X1^{s} + 4{,}67X2^{s} \ddot{y}\ 0{,}91X3^{s} \ddot{y}\ 5{,}84X4^{s}$$

$$YSV = 95{,}4 + 8{,}64X1^{s} + 10{,}3X2^{s}$$

Wir beachten hier, dass alle Schätzungen wurden aus standardisierten Erklärungen gewonnen

Machine Translated by Google

ry-Variablen. Wir vergleichen diese drei Schätzungen anhand der gesamten mittleren quadratischen Fehler der drei Vektoren:

$$\ddot{y}\hat{} \text{ LS} = \ddot{y}\hat{} \quad LS1 , \ddot{y}\hat{}LS2 , \ddot{y}\hat{}LS3 , \ddot{y}\hat{}_{LS4}$$

$$\ddot{y}\hat{} \text{ R} = \ddot{y}\hat{} \quad \text{R1} , \ddot{y}\hat{}\text{R2} , \ddot{y}\hat{}\text{R3} , \ddot{y}\hat{}_{\text{R4}}$$

$$\ddot{y}\hat{} \text{ SV} = \ddot{y}\hat{} \quad SV1 , \ddot{y}\hat{}SV2 , \ddot{y}\hat{}SV3 , \ddot{y}\hat{}SV4 .$$

Dabei entspricht der Index LS der Methode der kleinsten Quadrate, R der Ridge-Methode und SV der Variablenauswahl

Methode bzw. Für diese letztere Methode sind die Schätzungen für die Koeffizienten der uns Ausgewählte Variablen im Modell werden berücksichtigt Null sein. In unserem Fall haben wir:

$$\ddot{y}\hat{} \text{ LS} = (9{,}12,\ 7{,}94,\ 0{,}65,\ \ddot{y}2{,}41)\ \ddot{y}\hat{} \text{ R} =$$

$$(7{,}64,\ 4{,}67,\ \ddot{y}0{,}91,\ \ddot{y}5{,}84)\ \ddot{y}\hat{} \text{ SV} = (8{,}64,$$

$$10{,}3,\ 0,\ 0) .$$

Wir haben uns entschieden, den zugrunde liegenden Prozess, der zu den Zementdaten führt, durch die folgende Gleichung der kleinsten Quadrate anzunähern:

$$YiMC = 95{,}4 + 9{,}12\text{Xs} + 7{,}94\text{Xs}_{i1\ i2} + 0{,}65\text{Xs}\ \ddot{y}\ 2{,}41\text{Xs}_{i3\quad i4} + \ddot{y}i .$$

Das Verfahren besteht darin, 13 zufällige Fehlerterme $\ddot{y}1,...,\ \ddot{y}13$ 100-mal basierend auf einer Normalverteilung mit zu generieren

Mittelwert 0 und Standardabweichung 2,446 (denken Sie daran, dass 2,446 der Kleinste-Quadrate-Schätzer von ÿ für die Zementdaten ist). Anschließend berechnen wir $Y1LS,...,Y13LS$ anhand der *Xs*- Werte in der Datentabelle. Auf diese Weise generieren wir 100 $Y1LS,...,Y13LS$-Proben aus 100 $\ddot{y}1,...,\ \ddot{y}13$- Proben.

Die drei Methoden werden auf jede dieser 100 $Y1LS\text{-},...,Y13LS$- Proben (immer) angewendet unter Verwendung der gleichen Werte für *Xs* ij), was 100 Schätzer ÿ^ LS, ÿ^ R und ÿ^ SV ergibt . Beachten Sie, dass

Diese drei Methoden werden auf diese 100 Proben angewendet, ohne dass die Ergebnisse der Gleichungen Einfluss haben

$$YiLS = 95{,}4 + 9{,}12\text{Xs}_{i1} + 7{,}94\text{Xs}_{i2} + 0{,}65\text{Xs}_{i3}\ \ddot{y}\ 2{,}41\text{Xs}_{i4} ,$$

$$YiR = 95{,}4 + 7{,}64\text{Xs}_{i1} + 4{,}67\text{Xs}_{i2}\ \ddot{y}\ 0{,}91\text{Xs}\ \ddot{y}\ 5{,}84\text{Xs und i3 i4}$$

$$YiSV = 95{,}4 + 8{,}64\text{Xs}_{i1} + 10{,}3\text{Xs}_{i2}$$

für die Originalprobe erhalten. Trotz der Tatsache, dass die variable Auswahlmethode vorhanden ist wählte die Variablen *Xi1* und *Xi2* in $YiSV = 95{,}4 + 8{,}64\text{Xs} + 10{,}3\text{Xs}$ i1 i2, Es ist möglich, dass die Methode für eine dieser 100 Stichproben bessere *Xi2*- und *Xi3*- Variablen oder nur Xi3 oder die gesamte Teilmenge der vier verfügbaren Variablen ausgewählt hat. Auf die gleiche Weise, trotz der Tatsache, dass die Gleichung $YiR = 95{,}4 + 7{,}64\text{Xs} + 4{,}67\text{xXs}\ \ddot{y}\ 0{,}91\text{Xs}$ i2 i3 i1 ÿ 5,84Xs wurde i4 erhalten mit $k = 0{,}157$ wird der Wert von *k* für jede dieser 100 Stichproben neu berechnet (weitere Einzelheiten finden Sie im Beispiel der Ridge-Regression).

Aus diesen 100 Schätzungen von ÿ^ LS, ÿ^ SV und ÿ^ R können wir 100 *ETCs* für jede Methode berechnen, die wir wie folgt bezeichnen:

$$ETCLS = \ddot{y}\hat{} \quad LS1\ \ddot{y}\ 9{,}122\ LS2$$

$$\ddot{y}\ 0{,}652\ddot{y}\ 7{,}942 + \ddot{y}\hat{}\ LS3$$

$$+ \ddot{y}\hat{}\ 2{,}412$$

$$+ \ddot{y}\hat{} \quad LS4 +$$

$$ETCR = \ddot{y}\hat{}\ \text{R2}\ \ddot{y}\ \text{R1}\ \ddot{y}\ 9{,}122$$

$$0{,}652{,}942 + \ddot{y}\hat{}\ \text{R3}\ \ddot{y}$$

$$+ \ddot{y}\hat{} \quad + \ddot{y}R4 + 2{,}412$$

Machine Translated by Google

$$ETCSV = \ddot{y}\hat{} \quad SV1 \ddot{y} 9{,}122$$

$$\ddot{y}SV2 \ddot{y} 7{,}942 + \ddot{y}\hat{} \; SV3$$

$$+ \ddot{y}\hat{}652\ 2{,}412 .$$

$$+ \ddot{y}\hat{} \quad SV4 +$$

Die Mittelwerte der 100 Werte von ETCLS, ETCR und *ETCSV* sind die Schätzungen von *TMSELS, TMSESV* und TMSER: die TMSE- Schätzungen für die drei betrachteten Methoden.

Nachdem diese Simulation durchgeführt wurde, wurden die folgenden Schätzungen erhalten:

$$TMSELS = 270 \quad ,$$

$$TMSER = 75 .$$

$$TMSESV = 166 .$$

Daraus ergeben sich folgende Unterschiede:

$$TMSELS \ddot{y} TMSESV = 104 \quad ,$$

$$TMSELS \ddot{y} TMSER = 195 \quad ,$$

$$TMSESV \ddot{y} TMSER = 91 .$$

Da die Standardabweichungen von 100 ob

Da die ermittelten Unterschiede 350, 290 und 280 betragen, können wir die ungefähren 95 %-Konfidenzintervalle für die Unterschiede berechnen zenzen zwischen den *TMSEs* zweier Methoden

$$TMSELS \ddot{y} TMSESV = 104 \pm \frac{2 \cdot 350}{\ddot{y}100} ,$$

$$TMSELS \ddot{y} TMSER = 195 \pm \ddot{y}100 \frac{2 \cdot 290}{2 \cdot 280} ,$$

$$TMSESV \ddot{y} TMSER = 91 \pm \frac{}{\ddot{y}100} \cdot$$

Wir bekommen

$$34 < TMSELS \ddot{y} TMSESV < 174 \quad ,$$

$$137 < TMSELS \ddot{y} TMSER < 253 \quad ,$$

$$35 < TMSESV \ddot{y} TMSER < 147 \quad .$$

Wir können daher schlussfolgern (zumindest für das spezielle Modell, das zur Generierung der simulierten Daten verwendet wird, und unter Berücksichtigung unseres Ziels, einen kleinen TMSE zu erhalten), dass die Ridge-Methode die beste der betrachteten Methoden ist, gefolgt vom Variablenauswahlverfahren.

WEITERLESEN

Voreingenommenheit

Erwarteter Wert
Hutmatrix

Mittlere quadratische Fehler
Ridge-Regression
Standardisierte Daten
Varianz
Varianz-Kovarianz-Matrix

Gewichtete Methode der kleinsten Quadrate

REFERENZEN

Box, GEP, Draper, NR: Eine Grundlage für die Auswahl des Reaktionsflächendesigns. Marmelade. Stat. Assoc. **54,** 622–654 (1959)

Dodge, Y.: Analyse der angewandten Regression. Dunod, Paris (1999)

Kritischer Wert

Beim Testen **von Hypothesen** ist der kritische Wert der **Grenzwert ,** bei dem wir die Entscheidung treffen

Lehnen Sie **die Nullhypothese H0** für ein gegebenes **Signifikanzniveau ab.**

GESCHICHTE

Das Konzept eines kritischen Werts wurde 1928 von **Neyman, Jerzy** und **Pearson und Egon Sharpe** eingeführt .

MATHEMATISCHE ASPEKTE

Der kritische Wert hängt von der Art des verwendeten Tests ab **(zweiseitiger Test** oder **einseitiger Test).**

Machine Translated by Google

Test rechts oder links), die **Wahrscheinlichkeit Verteilung** und das **Signifikanzniveau** ÿ.

DOMÄNEN UND EINSCHRÄNKUNGEN

Der kritische Wert wird aus dem ermittelt **Wahrscheinlichkeitsverteilung** der **Statistik** mit dem Test verbunden. Es wird bestimmt durch Konsultieren Sie die **statistische Tabelle ,** die dieser Wahrscheinlichkeitsverteilung entspricht **(normal). Tisch, Studententisch, Fisher-Tisch, Chi-Quadrat-Tisch** usw.).

BEISPIELE

Ein Unternehmen produziert Stahlseile. Benutzen Anhand einer **Stichprobe** der Größe *n* = 100 soll überprüft werden, ob die Durchmesser der Kabel ausreichend genau dem geforderten Durchmesser entsprechen 0,9 cm im Allgemeinen.
Die **Standardabweichung** ÿ der **Grundgesamtheit** ist bekannt und beträgt 0,05 cm.
In diesem Fall handelt es sich um **Hypothesentests** ein **zweiseitiger Test.** Die Hypothesen lauten wie folgt:

Nullhypothese H0 : ÿ = *0,9*

Alternativhypothese H1 : ÿ = *0,9* .

Bei einem **Signifikanzniveau** von ÿ = 5 % stellen wir anhand der **Normaltabelle** fest, dass der kritische Wert *z* ÿ ist $\frac{}{2}$ entspricht 1,96.

WEITERLESEN

- **Konfidenzintervall**
- **Hypothesentest**
- **Signifikanzniveau**
- **Statistische Tabelle**

REFERENZ

Neyman, J., Pearson, ES: Über die Verwendung und Interpretation bestimmter Testkriterien für Zwecke der statistischen Schlussfolgerung, Teile I und II. Biometrie 20A, 175–240, 263–294 (1928)

C

Zyklische Schwankung

Zyklische Schwankungen sind ein gebräuchlicher Begriff beschreiben Schwingungen, die über einen längeren Zeitraum auftreten Perioden über der **säkularen Trendlinie** oder Kurve einer **Zeitreihe.**

GESCHICHTE

Siehe **Zeitreihen.**

MATHEMATISCHE ASPEKTE

Betrachten Sie *Yt,* eine **Zeitreihe ,** die durch ihre Komponenten gegeben ist; *Yt* kann geschrieben werden als:

- *Yt* = *Tt* ·*St* ·*Ct* ·*It* (multiplikatives **Modell),**

  oder;

- *Yt* = *Tt* + *St* + *Ct* + *It* (additives Modell).

Wo

Yt sind die **Daten** zum Zeitpunkt t;
Tt ist der **säkulare Trend** zum Zeitpunkt t;
St ist die **saisonale Variation** zum Zeitpunkt t;
Ct ist die zyklische Fluktuation zum Zeitpunkt t und;
Es ist die **unregelmäßige Variation** zum Zeitpunkt t.

Der erste Schritt bei der Untersuchung einer Zeitreihe ist immer der säkulare Trend *Tt zu bestimmen,* und dann die saisonalen Schwankungen zu bestimmen tion *St.* Es ist dann möglich, die Initiale anzupassen Daten der **Zeitreihe** *Yt* entsprechend zwei Komponenten:

- $\frac{Yt}{St \cdot Tt}$ = *Ct* · *It* (multiplikatives **Modell);**
- *Yt* ÿ *St* ÿ *Tt* = *Ct* + *It* (additives Modell).

Um konjunkturelle Schwankungen zu vermeiden, ist eine gewichtete **Der gleitende Durchschnitt** wird über einige festgelegt nur Monate. Die Verwendung gleitender Durchschnitte Ermöglicht die Glättung der **unregelmäßigen Schwankungen** *It* durch Beibehaltung der zyklischen Schwankungen *Ct.*

Machine Translated by Google

Die Wahl eines gewichteten **gleitenden Durchschnitts** ermöglicht es uns, den zentralen **Werten** im Vergleich zu den Extremwerten mehr Gewicht beizumessen , um zyklische Schwankungen genauer wiederzugeben. Daher groß

Die zentralen Werte werden gewichtet und kleine Gewichte bis zu den Extremwerten.

Beispielsweise gilt für einen **gleitenden Durchschnitt,** der für einen Zeitraum von fünf Monaten betrachtet wird, der Gewichte können *ÿ0,1,* 0,3, 0,6, 0,3 und *ÿ0,1* sein gebraucht; Da ihre Summe 1 ist, wird es keine geben Notwendigkeit einer Normalisierung.

Wenn die **Werte** von *Ct · It* (resultierend aus dem vorgenommene Anpassungen in Bezug auf **säkularen Trend** und den **saisonalen Schwankungen tionen)** werden mit *Xt* bezeichnet , dem Wert der Die zyklische Schwankung für den Monat *t* wird bestimmt durch:

$$Ct = ÿ0{,}1 \cdot Xtÿ2 + 0{,}3 \cdot Xtÿ1 + 0{,}6 \cdot Xt + 0{,}3 \cdot Xt+1 ÿ 0{,}1 \cdot Xt+2 .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Durch die Schätzung zyklischer Schwankungen können wir:

- Bestimmen Sie die Maxima oder Minima eine **Zeitreihe** erreichen kann.
- Erstellen Sie eine kurz- oder mittelfristige Prognose **ing.**
- Identifizieren Sie die zyklischen Komponenten.

Die Einschränkungen und Vorteile der Verwendung von gewichtete gleitende Durchschnitte bei der Auswertung zyklische Schwankungen sind folgende:

- Gewichtete gleitende Durchschnitte können glätten eine Kurve mit zyklischen Schwankungen, die Sie behalten immer noch den größten Teil der ursprünglichen Fluktuation bei, da sie die Amplituden der Zyklen auf genaue Weise bewahren.
- Die Verwendung einer ungeraden Anzahl von Monaten bis Die Ermittlung des gleitenden Durchschnitts erleichtert bessere Zentrierung der erhaltenen **Werte .**
- Aufgrund der Zyklen ist es schwierig, die zyklische Fluktuation einer **Zeitreihe** zu untersuchen Normalerweise variieren Länge und Amplitude. Das ist auf das Vorhandensein einer Vielzahl von Faktoren zurückzuführen, wobei die Auswirkungen dieser Faktoren auftreten können Wechsel von einem Zyklus zum anderen. Keiner der zur Erklärung und Vorhersage verwendeten Modelle Solche Schwankungen wurden festgestellt völlig zufriedenstellend.

BEISPIELE

Lassen Sie uns einen **gleitenden Durchschnitt** ermitteln, der über fünf Monate mit bereinigten **Daten** betrachtet wird je nach **säkularem Trend** und **saisonal Variationen.**

Verwenden wir auch die Gewichte ÿ0,1, 0,3, 0,6, 0,3, *ÿ0,1;* Da ihre Summe gleich 1 ist, gibt es sie keine Notwendigkeit einer Normalisierung.

Xi seien die angepassten **Werte** von *Ci · Ii:*

$$Ci = ÿ0{,}1 \cdot Xiÿ2 + 0{,}3 \cdot Xiÿ1 + 0{,}6 \cdot Xi + 0{,}3 \cdot Xi+1 ÿ 0{,}1 \cdot Xi+2 .$$

Die folgende Tabelle zeigt den Stromverbrauch von Straßenlaternen jeden Monat im Jahr Millionen Kilowattstunden im Laufe der Jahre 1952 und 1953. Die **Daten** wurden angepasst je nach **säkularem Trend** und **saisonal Variationen.**

Jahr-Monats-Daten		Xi	Gleitender Durchschnitt für 5 Monate Ci
1952 J		99,9	
	F	100,4	
	M	100,2	100,1
	A	99,0	99,1
	M	98,1	98,4
	J	99,0	98,7
	J	98,5	98,3
	A	97,8	98,3
	S	100,3	99,9
	Ö	101,1	101,3
	N	101,2	101,1
	D	100,4	100,7

Machine Translated by Google

Jahr-Monats-Daten	Xi	Gleitender Durchschnitt für 5 Monate Ci
1953 J	100,6	100,3
F	100,1	100,4
M	100,5	100,1
A	99,2	99,5
M	98,9	98,7
J	98,2	98,2
J	98,4	98,5
A	99,7	99,5
S	100,4	100,5
Ö	101,0	101,0
N	101.1	
D	101.1	

Dabei treten keine nennenswerten zyklischen Effekte auf Daten und nicht signifikante Auswirkungen bedeuten nicht keine Auswirkungen. Der Beginn einer wirtschaftlichen Der Zyklus wird oft gesucht, aber dieser erscheint nur alle 20 Jahre.

WEITERLESEN

- **Prognose**
- **Unregelmäßige Variation**
- **Gleitender Durchschnitt**
- **Saisonale Unterschiede**
- **Säkularen Trend**
- **Zeitfolgen**

REFERENZ

Box, GEP, Jenkins, GM: Zeitreihen Analyse: Prognose und Kontrolle (Reihe inTimeSeriesAnalysis).HoldenDay, San Francisco (1970)

Wold, HO (Hrsg.): Bibliographie zur Zeit Reihen und stochastische Prozesse. Oliver & Boyd, Edinburgh (1965)

Machine Translated by Google

Daniels, Henry E.

Daniels, Henry Ellis wurde in London geboren
Oktober 1912. Nach seinem Abschluss an der Edinburgh
University im Jahr 1933 setzte er sein Studium fort
Studium an der Universität Cambridge. Nach seiner
Promotion an der Universität Edinburgh war er
kehrte 1947 als Dozent für Mathematik nach Cambridge
zurück. 1957 Daniels, Henry
wurde der erste Professor für Mathematik an der
Birmingham University, eine Position, die er bis dahin innehatte
1978, bis zu seiner Pensionierung. Das Forschungsfeld
Was Daniels und Henry interessierte, waren
Inferenzstatistiken und Sattelpunktnäherungen
in Statistik, epidemiologischer Modellierung und
Statistische Theorie in ihrer Anwendung auf die
Textiltechnologie. Als er in Birmingham war, war er
gründete die Jahrestagung der Statistiker
in Gregynog Hall in Powys. Diese jährlichen Treffen
wurden nach und nach zu einem der
wird von englischen Statistikern am meisten angenommen.

Von 1974 bis 1975 war er Präsident
die Royal Statistical Society, die verliehen hat
ihm 1957 die Guy-Medaille in Silber
Gold im Jahr 1984. 1980 wurde er zum Mitglied der
Royal Society und 1985 zum Ehrenmitglied der
International Statistical Society gewählt
Gesellschaft. Er starb im Jahr 2000.

Einige Hauptwerke und Artikel von
Daniels, Henry E.:

1954 Sattelpunktnäherungen in der Statistik. Ann.
Mathematik. Stat. 25, 631–650.

1955 Diskussion eines Artikels von Box, GEP
und Anderson, SL. J. Roy. Stat. Soc.
Ser. B 17, 27–28.

1958 Diskussion des Aufsatzes von Cox, DR.
J. Roy. Stat. Soc. Ser. B 20, 236–238.

1987 Tail-Wahrscheinlichkeitsnäherungen. Int.
Stat. Rev. 55, 137–48.

Daten

Ein Datum (Pluraldaten) ist das Ergebnis eines
Beobachtung an einer **Population** oder an
eine **Probe.**
Das Wort „Datum“ ist lateinisch und bedeutet
„etwas Gegebenes“; es wird in der Mathematik verwendet
eine Information (nicht unbedingt numerisch) bezeichnen,
aus der eine Schlussfolgerung gezogen werden kann
kann gemacht werden.

Beachten Sie, dass eine Zahl (oder jede andere Form von
Beschreibung), die nicht unbedingt enthält
Alle Informationen sollten nicht mit verwechselt werden
ein Datum, das Informationen enthält.

Machine Translated by Google

Die aus Beobachtungen gewonnenen Daten sind bezogen auf die untersuchte **Variable .** Diese Daten sind quantitativ, qualitativ, diskret oder kontinuierlich, wenn die entsprechende Variable jeweils quantitativ, qualitativ, diskret oder kontinuierlich ist.

WEITERLESEN

- **Binärdaten**
- **Kategoriale Daten**
- **Unvollständige Daten**
- **Überwachung**
- **Bevölkerung**
- **Probe**
- **Geodaten**
- **Standardisierte Daten**
- **Wert**
- **Variable**

REFERENZEN

Federer, WT: Datenerfassung. In: Kotz, S., Johnson, NL (Hrsg.) Encyclopedia of Sta tistical Sciences, vol. 2. Wiley, New York (1982)

Datenanalyse

Einer der ersten Schritte in der wissenschaftlichen Forschung ist oft das Sammeln von **Daten.**

Diese Daten werden im Allgemeinen in zweidimensionalen Tabellen organisiert. Diese Organisation erleichtert es in der Regel, Informationen über die Daten zu extrahieren, sie also zu analysieren.

Im weitesten Sinne kann die Datenanalyse als das Wesen der **Statistik** betrachtet werden

mit dem alle anderen Aspekte des Themas verknüpft sind.

GESCHICHTE

Da **die Datenanalyse** viele verschiedene Methoden der statistischen Analyse umfasst, ist dies der Fall

Es ist schwierig, einen kurzen Überblick über die Geschichte der Datenanalyse zu geben. Dennoch können wir die Chronologie der grundlegendsten Aspekte des Themas schnell überblicken: • Die erste Veröffentlichung zur explorativen Datenanalyse stammt aus den Jahren 1970–1971 und wurde von Tukey, JW, verfasst. Dies war die erste Version seiner *Arbeit „Exploratory Data Analysis“,* die 1977 veröffentlicht wurde. • Die theoretischen Grundlagen der **Korrespondenzanalyse** gehen auf Hartley, HO (1935) (veröffentlicht unter seinem ursprünglichen deutschen Namen Hirschfeld) und auf **Fisher, RA** (1940) zurück. Die Theorie wurde jedoch größtenteils in den 1970er Jahren von Benzécri, JP, entwickelt • Die ersten Studien zur **Klassifizierung** waren in der Biologie und in der Zoologie durchgeführt. Die älteste Form der Typologie wurde von Galen (129–199 n. Chr.) konzipiert. Numerische Klassifizierungsmethoden leiten sich ab die Ideen von Adanson (im 18. Jahrhundert) und wurden unter anderem von Zubin (1938) und Thorndike (1953) entwickelt.

BEREICHE UND GRENZEN

Der Bereich der **Datenanalyse** umfasst viele verschiedene statistische Methoden.

Arten der Datenanalyse können wie folgt klassifiziert werden: 1. *Explorative Datenanalyse,* die (wie der Name schon sagt) die Untersuchung der Daten umfasst, durch: • **grafische Darstellung der Daten, • Datentransformation** (falls erforderlich) , • Erkennung **von Ausreißerbeobachtungen ,** • Ausarbeitung von Forschungshypothesen , die zu Beginn des **Experiments nicht vorgesehen waren, • Robuste Schätzung.**

Machine Translated by Google

2. *Erste Datenanalyse,* die Folgendes umfasst:
   - Auswahl der statistischen Methoden auf die Daten angewendet.
3. *Multivariate Datenanalyse,* die Folgendes umfasst:
   - Diskriminanzanalyse,
   - Datentransformation , die die Anzahl der Dimensionen und Einrichtungen reduziert
   - Tates-Interpretation,
   - Suche nach Struktur.
4. *Spezifische Formen der Datenanalyse* , die für unterschiedliche Analyseaufgaben geeignet sind; Zu diesen Formen gehören:
   - **Korrespondenzanalyse,**
   - Mehrfachkorrespondenzanalyse,
   - **Klassifizierung.**
5. *Bestätigungsdatenanalyse,* bei der Analyseergebnisse ausgewertet und getestet werden; Dazu gehören:
   - **Parameterschätzung,**
   - **Hypothesentests,** Die Verallgemeinerung und die Schlussfolgerung sion.

WEITERLESEN

Einstufung
Clusteranalyse
Korrespondenzanalyse
Daten
Explorative Datenanalyse
Hypothesentest
Statistiken
Transformation

VERWEISE

Benzécri, JP: L'Analyse des Données. Bd. 1: La Taxinomie. Bd. 2: L'Analyse Factorielle des Correspondances. Dunod, Paris (1976)

Benzécri, JP: Geschichte und Vorgeschichte der Analyse der Données, Les Cahiers de l'analyse des données 1, Nr. 1–4. Dunod, Paris (1976)

Everitt, BS: Clusteranalyse. Halstead, London (1974)

Fisher, RA: Die Präzision von Diskriminanzfunktionen.Ann.Eugen. (London)10.422–429 (1940)

Gower, JC: Klassifizierung, Geometrie und Datenanalyse. In: Bock, HH (Hrsg.) Klassifizierung und verwandte Methoden von Datenanalyse. Elsevier, Amsterdam (1988)

Hirschfeld, HO: Ein Zusammenhang zwischen Korrelation und Kontingenz. Proz. Camb. Philos. Soc. **31,** 520–524 (1935)

Thorndike, RL: Wer gehört in eine Familie? Psychometrika **18,** 267–276 (1953)

Tukey, JW: Explanatory Data Analysis, limitierte Vorabausgabe. Addison Wesley, Reading, MA (1970–1971)

Tukey, JW: Explorative Datenanalyse. Addison-Wesley, Reading, MA (1977)

Zubin, J.: Eine Technik zur Messung von Gleichgesinntheit. J. Abnorm. Sozialpsychol. **33,** 508–516 (1938)

Datensammlung

Beim Sammeln von **Daten** müssen wir mehrere Aspekte berücksichtigen. Zunächst muss festgelegt werden, warum wir die Daten erheben müssen und wofür die Daten verwendet werden. Zweitens müssen wir die Art der Daten berücksichtigen, die erfasst werden sollen: Es ist wichtig, dass die Daten erfasst werden stehen im Zusammenhang mit dem Ziel der Studie. Abschließend muss noch überlegt werden, wie die Daten erhoben werden sollen Hrsg. Es gibt drei Ansätze zur Datenerhebung:

Machine Translated by Google

1. Registrieren
2. **Probenahme** und **Volkszählung**
3. Experimentelle Forschung

1. *Registrieren*

Eine Form zugänglicher Daten sind registrierte Daten eins. Es handelt sich beispielsweise um registrierte Daten über Geburten, Todesfälle, Hochzeiten, Tagestemperaturen, monatliche Regenfälle und Autoverkäufe. Es gibt keine allgemeinen statistischen Methoden stellen sicher, dass wertvolle Schlussfolgerungen gezogen werden aus diesen **Daten.** Jeder Datensatz sollte sein für sich betrachtet betrachtet werden, und das sollte man auch tun Seien Sie vorsichtig, wenn Sie **Schlussfolgerungen** ziehen **Bevölkerung.**

2. *Probenahme und Volkszählung*

Die **Probenahmemethoden** können in zwei unterteilt werden Kategorien: Zufallsmethoden und nichtzufällige Methoden Methoden.

Nichtzufällige Methoden umfassen die Konstruktion, empirisch gesehen eine **Stichprobe ,** die der **Grundgesamtheit** ähnelt, aus der sie entnommen wird so viel wie möglich. Am häufigsten verwendet Die nichtzufällige Methode ist **die Quotenstichprobe.**

Zufällige Methoden verwenden ein probabilistisches Verfahren, um eine **Stichprobe aus der Grundgesamtheit abzuleiten.** Angesichts der Tatsache, dass die Wahrscheinlichkeit bekannt ist, dass eine bestimmte Einheit in der Stichprobe ausgewählt wird, ist die Der durch die Stichprobenentnahme verursachte Fehler kann dann berechnet werden. Die wichtigsten Zufallsmethoden sind **einfache Zufallsstichproben, geschichtete Stichproben, systematische Stichproben** und **Chargenstichproben.**

Bei einer **Volkszählung** werden alle Objekte einer **Population** beobachtet, wodurch Daten für die gesamte Bevölkerung gewonnen werden Bevölkerung. Ganz klar, diese Art von Untersuchung ist sehr kostspielig, wenn die Bevölkerung sehr groß ist groß. Aus diesem Grund werden keine Volkszählungen durchgeführt sehr oft.

3. *Experimentelle Forschung*

In einem **Experiment** erfolgt die Datenerfassung auf der Grundlage eines bestimmten Experiments AI-Design. Diese experimentellen Designs sind Wird auf **die Datenerfassung** in allen Forschungsbereichen angewendet Felder.

GESCHICHTE

Es ist wahrscheinlich, dass die älteste Form der **Datenerhebung** auf Volkszählungen zurückgeht, die in der Antike durchgeführt wurden.

Antille und Ujvari (1991) erwähnen die Existenz einer Stelle für einen amtlichen Statistiker in China während der Chow-Dynastie (111–211). v. Chr.).

Der römische Autor Tacitus sagt, dass Kaiser oder Augustus allen Soldaten Schiffe befahl und Reichtum im Reich zu zählen.

Belege für **Volkszählungen** finden sich auch in die Bibel; Der heilige Lukas berichtet, dass „Cäsar Augustus ordnete ein Dekret an, das eine Volkszählung der ganzen Welt vorschrieb (...) und alle gingen dazu eingeschrieben werden, jeder in seiner eigenen Stadt.“

Eine Form der Statistik kann auch während gefunden werden diesmal. Sein Name verrät seine Verwaltung Ursprung, weil es vom lateinischen Wort kommt „Status“: der Staat.

WEITERLESEN

- **Volkszählung**
- **Versuchsplanung**
- **Probenahme**
- **Umfrage**

VERWEISE

Antille, G., Ujvari, A.: Pratique de la Statis tique Inférentielle. PAN, Neuenburg, Schweiz (1991)

Boursin, J.-L.: Les structure du hasard. Seuil, Paris (1986)

Dezil

Dezile **sind berechnete** Positionsmaße auf einem Datensatz .

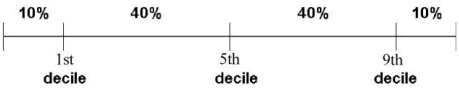
Machine Translated by Google

Die Dezile sind die **Werte ,** die eine Verteilung in gleiche Teile unterteilen, wobei jeder Teil die gleiche Anzahl an **Beobachtungen enthält.** Das Dezil gehört zur größeren Familie der **Quantile.**

Das x-te Dezil gibt den Wert an, bei dem *10x %* der Beobachtungen liegen darunter Wert und (100 ÿ *10x) %* der Beobachtungen liegen über diesem Wert. Beispielsweise ist das achte Dezil der Wert, bei dem 80 % der Beobachtungen unter diesem Wert liegen und 20 % auftreten über.

Das fünfte Dezil stellt den **Median dar.**

Für eine gegebene Verteilung gibt es also neun Dezile:



MATHEMATISCHE ASPEKTE

Der Prozess zur Berechnung von Dezilen ähnelt dem zur Berechnung des **Medians** oder

Quartile.

Wenn alle **Rohbeobachtungen** vorliegen

In der Lage ist der Prozess zur Berechnung von Dezilen wie folgt:

1. Organisieren Sie die *n* **Beobachtungen** in einem **Fre Frequenzverteilung**
2. Die Dezile entsprechen den **Beobachtungen ,** deren kumulative relative **Häufigkeit** 10 %, 20 %, 30 %, ..., 80 %, 90 % übersteigt.

Einige Autoren schlagen vor, die folgende Formel zu verwenden, um die **Werte** der verschiedenen Dezile genau zu bestimmen:

Berechnung des j-ten Dezils:

Sei *i* der ganzzahlige Teil von $j \cdot \frac{n+1}{10}$ und 10 ~~*k*~~ der Bruchteil von $j \cdot \frac{n+1}{10}$.

Nehmen Sie *xi* und xi+1 als die **Werte** der **Beobachtungen** an der i-ten und (i+1)-ten Position tionen (wenn die Beobachtungen in aufsteigender Reihenfolge angeordnet sind).

Das j-te Dezil ist dann gleich:

$$Dj = xi + k \cdot (xi+1 \text{ ÿ } xi) .$$

Wenn die **Beobachtungen** in Klassen gruppiert werden, werden die Dezile auf folgende Weise bestimmt: 1.

Bestimmen Sie die Klasse, die die enthält gewünschtes Dezil liegt:

- *Erstes Dezil:* die erste Klasse, für die die kumulierte relative **Häufigkeit** übersteigt 10 %.
- *Zweites Dezil:* die erste Klasse, für die die kumulierte relative **Häufigkeit** übersteigt 20 %.

  ...
- *Neuntes Dezil:* die erste Klasse, für die die kumulierte relative **Häufigkeit** übersteigt 90 %.

2. Berechnen Sie den **Wert** des Dezils basierend auf die **Hypothese ,** dass die **Beobachtungen** in jeder Klasse gleichmäßig verteilt sind:

$$\text{Dezil} = L1 + \frac{(N \cdot q) \text{ ÿ finf}}{\text{fdecile}} \cdot c .$$

Wo

L1 = untere Grenze der Klasse des Dezils

N = Gesamtzahl der **Beobachtungen**

q = 1/10 für das erste Dezil

q = 2/10 für das zweite Dezil

...

q = 9/10 für das neunte Dezil

finf = Summe der **Häufigkeiten** unterhalb der Klasse des Dezils

fdecile = Häufigkeit der Klasse des Dezils

c = Größe des **Intervalls** der Klasse des Dezils

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Berechnung von Dezilen hat nur einen Sinn für **eine quantitative Variable ,** die über ein bestimmtes Intervall **Werte** annehmen kann . Die Anzahl der **Beobachtungen** muss sein relativ hoch, da die Berechnung von Bei Dezilen wird die Menge der Beobachtungen in zehn Teile unterteilt. Dezile werden relativ häufig verwendet üben; zum Beispiel beim Dolmetschen die Verteilung der Einnahmen in einer Stadt bzw Zustand.

BEISPIELE

Betrachten Sie ein Beispiel, in dem sich die Dezile befinden berechnet für eine **Häufigkeitsverteilung** von eine kontinuierliche **Variable,** bei der die **Observa**

Die Funktionen sind in Klassen gruppiert.

Die folgende **Häufigkeitstabelle** stellt dar der Gewinn (in Tausend Euro) von 500 Bäckereien:

Gewinn (in Tausende in Euro)	Frei quen cies	Cumu verspätet Frequenz	Relativ kumuliert Frequenz
100–150	40	40	0,08
150–200	50	90	0,18
200–250	60	150	0,30
250–300	70	220	0,44
300–350	100	320	0,62
350–400	80	400	0,80
400–450	60	460	0,92
450–500	40	500	1,00
Gesamt	500		

Das erste Dezil liegt in der Klasse 150–200 (dies ist die kumulative relative **Häufigkeit** 10 % überschreitet).

Unter der Annahme, dass die **Beobachtungen** in jeder Klasse gleichmäßig verteilt sind, erhalten wir: der folgende **Wert** für das erste Dezil:

$$1.\ \text{Dezil} = 150 + \frac{(500 \cdot \frac{1}{10}) \ddot{y}\ 40}{50} \cdot 50$$

$$= 160\ .$$

Das zweite Dezil fällt in die Klasse 200–250.

Der Wert des zweiten Dezils ist gleich

$$2.\ \text{Dezil} = 200 + \frac{(500 \cdot \frac{2}{\ }) \ddot{y}\ 90}{10\ 60} \cdot 50$$

$$= 208{,}33\ .$$

Wir können die anderen Dezile im selben berechnen Weg, was ergibt:

Dezil	Klasse	Wert
1	150–200	160,00
2	200–250	208.33
3	200–250	250,00
4	250–300	285,71
5	300–350	315,00
6	300–350	340,00
7	350–400	368,75
8	350–400	400,00
9	400–450	441,67

Wir können zum Beispiel daraus schließen, dass 10 % von Die 500 Bäckereien erwirtschaften einen Gewinn von 1,5 % 100.000 und 160.000 Euro, 50 % machen einen Gewinn, der niedriger als 315.000 Euro ist, und 10 % machen einen Gewinn zwischen 441670 und 500000 Euro.

WEITERLESEN

- **Standortmaß**
- **Median**
- **Perzentil**
- **Quantil**
- **Quartil**

Machine Translated by Google

Freiheitsgrad

Die Anzahl der Freiheitsgrade beträgt ein **Parameter** aus der **Chi-Quadrat-Verteilung.** Es ist auch ein Parameter, der in anderen verwendet wird **Wahrscheinlichkeitsverteilungen** im Zusammenhang mit der Chi-Quadrat-Verteilung, z. B. **Student Verteilung** und die **Fisher-Verteilung.**

In einem anderen Zusammenhang die Gradzahl Die Freiheit bezieht sich auf die Anzahl der linearen

Bei der Berechnung der Quadratsumme auf der Grundlage von *n* unabhängigen **Beobachtungen sind sehr unabhängige Terme beteiligt.**

GESCHICHTE

Der Begriff „Freiheitsgrad“ wurde 1925 von **Fisher, RA** eingeführt .

MATHEMATISCHE ASPEKTE

Sei Y1, *Y2,...,Yn* eine Zufallsstichprobe von Größe *n* aus einer Population mit an unbekannter Mittelwert Y¯. Die Summe **der Abweichungen** von *n* **Beobachtungen** ist bezüglich ihres **arithmetischen Mittels** immer gleich Null:

$$\sum_{i=1}^{N} (Yi \ddot{y} Y^{-}) = 0 .$$

Diese Anforderung stellt eine Einschränkung für jede Abweichung *Yi ÿY¯* dar, die bei der Berechnung der **Vari** verwendet wird **ance:**

$$S2 = \frac{\sum_{i=1}^{N} (Yi \ddot{y} Y^{-})^{2}}{n \ddot{y} 1} .$$

Diese Einschränkung impliziert, dass *n* ÿ 1 Abweichungen die n-te Abweichung vollständig bestimmen. Die *n* Abweichungen (und auch die Summe von ihre Quadrate und die Varianz im S2 des **Stichprobe)** haben daher nÿ1 Freiheitsgrade.

WEITERLESEN

- **Chi-Quadrat-Verteilung**
- **Fisher-Verteilung**
- **Parameter**
- **Studentenverteilung**
- **Varianz**

VERWEISE

Fisher, RA: Anwendungen von „Student's“ Verteilung. Metron 5, 90–104 (1925)

Deming, W. Edwards

Deming, W. Edwards wurde 1900 geboren Sioux City, Iowa. Er studierte Naturwissenschaften an der Wyoming University, Abschluss 1921, und an der Colorado University, wo er seinen Master in Mathematik abschloss und Physik im Jahr 1924. Dann ging er nach Yale Universität, wo er promovierte in Physik im Jahr 1928. Nach seiner Promotion war er arbeitete zehn Jahre lang im Labor von das Ministerium für Landwirtschaft. Im Jahr 1939 wechselte Deming zum Census Bureau in Washington. Dort nutzte er seine theoretischen Kenntnisse, um die ersten Stichprobenerhebungen zu initiieren. Diese Volkszählungen verwendet Techniken, die später ein Beispiel dafür lieferten Ähnliche Volkszählungen wurden rund um den Rest durchgeführt der Welt. Er verließ das Census Bureau 1946Beratung im Statistikstudium ies und Professor für Statistik in New York Universität. Er starb in Washington D.C 1993.

Lange Zeit (bis 1980) galten die Theorien Von Deming wurden WIR von amerikanischen Unternehmen ignoriert. Im Jahr 1947 ging Deming nach Tokio als Berater zur Anwendung seiner Techniken der Probenahme. Ab 1950 japanisch Die Industrie übernahm die Managementtheorien von Deming. Innerhalb von zehn Jahren viele Japaner

D

Machine Translated by Google

Produkte wurden nach Amerika exportiert, weil sie besser und kostengünstiger waren als entsprechende in den Vereinigten Staaten hergestellte Produkte.

Einige Hauptwerke und Artikel von Deming, W. Edwards:

1938 Statistische Anpassung der Daten. Wiley, New York.

1950 Einige Sampling-Theorien. Wiley, New York.

1960 SampleDesigninBusinessResearch. Wiley, New York.

Demographie

Unter Demografie versteht man die Untersuchung der menschlichen **Bevölkerung.** Dabei werden Phänomene wie Geburten, Todesfälle, Migration, Ehen, Scheidungen, Fruchtbarkeitsrate, Sterblichkeitsrate und Alterspyramiden analysiert.
Diese Phänomene werden sowohl aus biologischer als auch aus sozioökonomischer Sicht behandelt.
Die in der Demografie verwendeten Methoden sind auch fortgeschrittene Mathematik und Statistik

so viele Bereiche der Sozialwissenschaften.

GESCHICHTE Die ersten demografischen Studien stammen aus der Mitte des 17. Jahrhunderts. In dieser Zeit analysierte Graunt, John, ein englischer Wissenschaftler, die einzigen verfügbaren Bevölkerungsdaten: eine Liste der Todesfälle im Großraum London, geordnet nach ihrer Ursache. Im Jahr 1662 veröffentlichte er eine Studie, in der er versuchte, die durchschnittliche Größe einer Familie, die Bedeutung der Migrationsbewegung und andere Elemente im Zusammenhang mit der Bevölkerungsstruktur zu bewerten. In Zusammenarbeit mit Petty und Sir William schlug Graunt ernsthaftere Studien vor Die Grundgesamtheit sollte erstellt werden und es sollte ein Zentrum geschaffen werden, wo statistische **Daten** gesammelt würden.

Im 18. Jahrhundert wurden die gleichen Arten von Analysen durchgeführt und schrittweise verbessert, aber erst im 19. Jahrhundert führten mehrere europäische Länder sowie die Vereinigten Staaten nationale Analysen durch

Volkszählungen und erstellte die regelmäßige Aufzeichnung von Geburten, Ehen und Todesfällen. Diese Berichte zeigen, dass verschiedene Regionen unterschiedliche Überlebenschancen für sie boten

Bewohner. Diese Schlussfolgerungen führten letztendlich zu verbesserten Arbeits- und Hygienebedingungen.

Anhand der gesammelten demografischen Daten wurden Vorhersagen möglich, und gegen Ende des Jahrhunderts erschienen die ersten demografischen Zeitschriften und Rezensionen. Dazu gehörten „Demografie“ in den Vereinigten Staaten, „Bevölkerung“ in Frankreich und „Bevölkerungsstudien“ in Großbritannien.

In der Mitte des 20. Jahrhunderts begannen die Bevölkerungsstudien, sich auf die Weltbevölkerung zu konzentrieren, als die demografischen Probleme der Dritten Welt zu einem wichtigen Thema wurden.

WEITERLESEN

Volkszählung

Bevölkerung

VERWEISE

Cox, PR: Demographie, 5. Aufl. Cambridge University Press, Cambridge (1976)

Hauser, PM, Duncan, OD: Das Studium der Bevölkerung: Eine Bestandsaufnahme und Bewertung. University of Chicago Press, Chicago, IL (1959)

Machine Translated by Google

Dendrogramm

Ein Dendrogramm ist eine grafische Darstellung verschiedener Aggregationen, die während einer **Clusteranalyse erstellt wurden.** Es besteht aus Knoten, die Gruppen entsprechen, und Zweigen, die repräsentieren die bei jedem Schritt hergestellten Assoziationen. Die Struktur des Dendrogramms wird durch die bestimmt Reihenfolge, in der die Aggregationen vorgenommen werden. Wenn dem Dendrogramm eine **Skala** hinzugefügt wird, ist es möglich, die **Entfernungen** darzustellen, über die es geht die Aggregationen fanden statt.

GESCHICHTE

Im allgemeineren Sinne ist ein Dendrogramm (von das griechische „dendron“, was Baum bedeutet) ist ein Baum Diagramm, das die Beziehungen veranschaulicht, die existieren zwischen den Mitgliedern einer Menge.

Die ersten Beispiele für Dendrogramme waren die phylogenetische Bäume, die von Systematikspezialisten verwendet werden. Der Begriff „Dendrogramm“ scheint es zu sein wurden in der Arbeit zum ersten Mal verwendet von Mayr et al. (1953).

MATHEMATISCHE ASPEKTE

Bei **der Clusteranalyse** einer Reihe von Objekten Aggregationen werden mit Hilfe von erreicht eine **Distanztabelle** und die gewählte Verknüpfung Methode **(Einzelverknüpfungsmethode, vollständig Verknüpfungsmethode** usw.).
Jeder Extrempunkt des Dendrogramms stellt eine andere automatisch erzeugte Klasse dar Einstufung.

Das Dendrogramm wird dann erstellt, indem jede Gruppe durch einen bei platzierten Knoten dargestellt wird eine bestimmte Position in Bezug auf die horizontale **Skala,** wobei die Position davon abhängt abhängig von der **Distanz ,** die das Aggregat zurücklegt gebildet.

Die Objekte werden auf der Nullebene des platziert Skala. Um die Zweige zu stoppen, die die verbinden

Um zu verhindern, dass sich Knoten verheddern, wird der Vorgang des Zeichnens eines Dendrogramms durchgeführt auf systematische Weise.

- Die ersten beiden gruppierten Elemente werden ersetzt, übereinander, auf der Nullebene der **Skala.** Eine horizontale Linie ist dann neben jedem Element gezeichnet, beginnend mit dem Nullpunkt Ebene zur Aggregationsentfernung für jeden Element. Die resultierende Klasse wird dann durch eine vertikale Linie dargestellt, die die Klasse verbindet Enden der Linien für die Elemente (bilden ein Knoten"). Die Mitte dieser vertikalen Linie Bietet den Ausgangspunkt für eine neue horizontale Linie für diese Klasse, wenn sie mit einer anderen Klasse zusammengefasst wird. Das Dendrogramm besteht daher aus Zweigkonstrukten aus horizontalen und vertikalen Linien gebildet.
- Zwei Sonderfälle können während des auftreten nächster Schritt:
  1. Zwei Klassen, die jeweils aus einer bestehen Element werden aggregiert; Diese Klassen werden auf der Nullebene platziert und die Der resultierende Knoten befindet sich im Aggregationsabstand, wie zuvor beschrieben.
  2. Ein Element wird zu einer Klasse aggregiert zuvor gebildet. Das Element ist aggregiert ab der Nullebene und wird zwischen bestehenden Klassen eingefügt Falls benötigt). Der Knoten repräsentiert Diese Aggregation wird im (horizontalen) **Abstand** platziert , der dem entspricht Aggregationsabstand zwischen der Klasse und das Element.

Dieser Vorgang wird bis zum gewünschten Ergebnis fortgesetzt Konfiguration erhalten wird. Falls noch welche vorhanden sind Klassen, die aus einem einzelnen Element bestehen Am Ende des Aggregationsprozesses werden sie einfach ab der Nullebene des Dendrogramms hinzugefügt. Wenn die Klassifizierung **abgeschlossen** ist, entspricht jede Klasse einem bestimmten Teil des Dendrogramms.

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Dendrogramm beschreibt den geordneten Pfad der Reihe von Operationen, die während ausgeführt werden **Clusteranalyse.** Es veranschaulicht diese Art von **Klassifizierung** sehr präzise vornehmen. Dieser streng definierte Ansatz zur Erstellung eines Dendrogramms wird manchmal aufgrund von Änderungen geändert an die Umstände. Zum Beispiel, wie wir es wollen Sehen Sie in einem der folgenden Beispiele die Aggregationsabstände von zwei oder mehr aufeinanderfolgender Die einzelnen Schritte und die Vorgehensweise können gleich sein muss dann geändert werden, um sicherzustellen, dass die Zweige des Dendrogramms verschränken sich nicht.

In dem Fall, den wir oben beschrieben haben, ist die Das Dendrogramm wird horizontal gezeichnet. Natürlich ist es auch möglich, ein Dendrogramm zu zeichnen von oben nach unten oder von unten nach oben. Die Zweige können sogar aus Diagonalen bestehen Linien.

BEISPIELE

Wir werden eine **Clusteranalyse** mit dem durchführen **Verwenden Sie die Single-Link-Methode** und erstellen Sie das entsprechende Dendrogramm, das die verschiedenen Aggregationen veranschaulicht, die auftreten können. Betrachten Sie die Noten von fünf Schülern bei Prüfungen für vier verschiedene Kurse: Englisch, Französisch, Mathematik und Physik. Diese fünf Studierenden möchten wir über den **einzigen Link** in zwei Gruppen aufteilen **Methode,** um eine neue Lehre zu testen Methode.

Diese Noten können Werte von 1 annehmen bis 6 sind in der folgenden Tabelle zusammengefasst:

	Englisch	Französisch	Mathematik	Phys
Alain	5.0	3.5	4,0	4.5
Jean	5.5	4,0	5,0	4.5
März	4.5	4.5	4,0	3.5
Paul	4.0	5.5	3.5	4,0
Pierre	4.0	4.5	3,0	3.5

Die **Distanztabelle** erhält man, indem man die euklidischen **Distanzen** zwischen den Schülern berechnet und sie zur folgenden Tabelle addiert:

	Alain	Jean	Marc	Paul	Pierre
Alain	0	1,22	1,5	2,35	2
Jean	1,22	0	1.8	2,65	2,74
März	1.5	1,8	0	1,32	1,12
Paulus	2,35	2,65	1,32	0	1.22
Pierre	2	2,74	1,12	1,22	0

Verwendung der **Single-Link-Methode,** basierend auf der **Mindestabstand** zwischen den Objekten In zwei Klassen gibt es die folgenden Partitionen bei jedem Schritt erhalten:

Erster Schritt: Marc und Pierre werden gruppiert eine Aggregationsdistanz von 1,12: {Alain}, {Jean}, {Marc,Pierre}, {Paul} *Zweiter Schritt:* Wir überarbeiten die Distanztabelle. Die neue **dis**

Die Tabelle ist unten aufgeführt:

	Jean	Marc und Pierre	Paul
Alain	1,22	1,5	2,35
Jean		1.8	2,65
Marc und Pierre			1.22

Alain und Jean werden dann mit Abstand gruppiert von 1,22, und wir erhalten die folgende Partition: {Alain, Jean}, {Marc, Pierre}, {Paul}

Dritter Schritt: Die neue Distanztabelle ist vorgegeben unter:

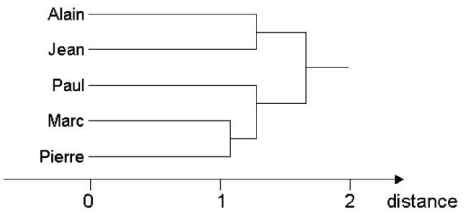
	Marc und Pierre	Paul
Alain und Jean	1.5	2,35
Marc und Pierre		1.22

Paul schließt sich dann Marc und Pierre an ein Abstand von 1,22, was die folgende Aufteilung ergibt: {Alain, Jean}, {Marc, Pierre, Paul}

Vierter Schritt: Die beiden Gruppen werden aggregiert im Abstand von 1,5.

Dies ergibt das folgende **Dendrogramm:**

Machine Translated by Google



Beachten Sie, dass Paul vorher positioniert werden muss der Gruppe von Marc und Pierre, um auszuweichen Verflechtung der Zweige des Diagramms.

WEITERLESEN

Einstufung

Clusteranalyse

Vollständige Verknüpfungsmethode

Distanz

Entfernungstabelle

VERWEISE

Mayr, E., Linsley, EG, Usinger, RL: Methoden und Prinzipien der systematischen Zoologie. McGraw-Hill, New York (1953)

Sneath, PHA, Sokal, RR: Numerische Steuer Onomie: Die Prinzipien und die Praxis von Numerische Klassifikation (Eine Reihe von Bücher zur Biologie). WH Freeman, San Francisco, Kalifornien (1973)

Sokal, RR, Sneath, PH: Prinzipien von Numerische Taxonomie. Freeman, San Francisco (1963)

Dichtefunktion

Die Dichtefunktion einer kontinuierlichen **Zufallsvariablen** ermöglicht es uns, die zu bestimmen **Wahrscheinlichkeit ,** die eine Zufallsvariable *X* annimmt Werte in einem bestimmten **Intervall.**

GESCHICHTE

Siehe **Wahrscheinlichkeit.**

MATHEMATISCHE ASPEKTE

Betrachten Sie P(a ÿ *X* ÿ b), die **Wahrscheinlichkeit** dass eine kontinuierliche **Zufallsvariable** *X* annimmt **einWertimIntervall[a,** b].Diese Wahrscheinlichkeit ist definiert durch:

$$P(a \ddot{y} X \ddot{y} b) = \int_A^B f(x)dx .$$

wobei f(x) die Dichtefunktion der Zufallsvariablen X ist.

Die Dichtefunktion wird grafisch auf einem Achsensystem dargestellt. Auf die verschiedenen **Werte** der **Zufallsvariablen** *X* werden gelegt die Abszisse und diejenigen, die von der Funktionf übernommen werden werden auf der **Ordinate platziert.** Der Graph der Funktion *f* erlaubt es uns nicht um die **Wahrscheinlichkeit** für einen bestimmten Punkt zu bestimmen, sondern um die Wahrscheinlichkeit für ein **Intervall** auf einer Oberfläche zu visualisieren. Die Gesamtfläche unter der Kurve korre reagiert auf einen Wert von 1:

$$\int f(x)dx = 1 \quad .$$

Die Variable *X* ist eine kontinuierliche Zufallsvariable möglich, wenn es eine nichtnegative Funktion *f* dafür gibt für reelle Zahlen definiert ist und für die die

Die folgende Eigenschaft gilt für jedes Intervall *[a,* b]:

$$P\,(a \ddot{y} X \ddot{y} b) = \int_A^B f\,(x)\,dx \quad ,$$

Dabei ist P *(a* ÿ *X* ÿ **b) die Wahrscheinlichkeitsfunktion. Daher ist die Wahrscheinlichkeit, dass** die kontinuierliche **Zufallsvariable** *X* einen **Wert** annimmt

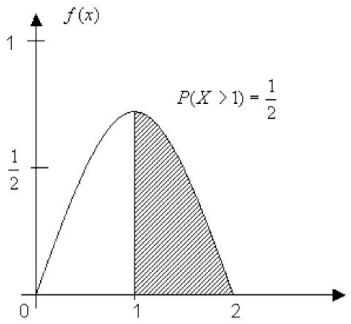
Das **Intervall** *[a,* b] kann durch Integration der Wahrscheinlichkeitsfunktion über *[a,* b] erhalten werden. Es gibt auch die folgende Bedingung für *f :*

$$\int_{\ddot{y}\ddot{y}}^{\ddot{y}} f\,(x)\,dx = P\,(\ddot{y} \ddot{y}\ X \ddot{y}\ \ddot{y}) = 1 \quad .$$

Wir können die Dichtefunktion grafisch auf dem Achsensystem darstellen. Der Unterschied

Machine Translated by Google

Auf der Abszisse sind Werte der Zufallsvariablen *X* und auf der Ordinate die Werte der Funktion *f* aufgetragen.



Die schattierte Oberfläche stellt den Bereich dar, in dem P(X > 1) ist.

DOMÄNEN UND EINSCHRÄNKUNGEN

Wenn *a* = *b* in P(a ÿ *X* ÿ b) = $\int_A^B f(x)dx$ also

$$P(X = a) = \int_A^A f(x)dx = 0 .$$

Dies bedeutet, dass die **Wahrscheinlichkeit ,** dass eine kontinuierliche **Zufallsvariable** einen isolierten Wert annimmt

Der feste **Wert** ist immer Null.
Es spielt daher keine Rolle, ob wir die Grenzen der einschließen oder ausschließen **Intervall** bei der Berechnung der mit einem Intervall verbundenen **Wahrscheinlichkeit .**

BEISPIELE

Betrachten Sie X, eine kontinuierliche **Zufallsvariable,** für die die Dichtefunktion gilt

$$f(x) = \begin{cases} \frac{3}{4}(\ddot{y}x2 + 2x) , & \text{wenn } 0 < x < 2 \\ 0 , & \text{wenn nicht} \end{cases} .$$

Wir können die **Wahrscheinlichkeit** berechnen , dass *X* größer als 1 ist. Wir erhalten:

$$P(X > 1) = \int_1^{\ddot{y}} f(x)dx ,$$

die unterteilt werden können in:

$$P(X > 1) = \int_1^2 \frac{3}{4}(\ddot{y}x2 + 2x)dx + \int_2^{\ddot{y}} 0\, dx$$

$$= \frac{3}{4}\left[\frac{\ddot{y}x3}{3} + x2\right]_1^2$$

$$= \frac{1}{2} .$$

Die Dichtefunktion und der Bereich, in dem P(X > 1) lässt sich grafisch darstellen:

WEITERLESEN

- **Kontinuierliche Verteilungsfunktion**
- **Kontinuierliche Wahrscheinlichkeitsverteilung**
- **Funktion der Gelenkdichte**
- **Wahrscheinlichkeit**

Abhängigkeit

Der Begriff der Abhängigkeit kann zwei Bedeutungen haben. Das erste betrifft die **Ereignisse** eines **Zufallsexperiments.** Man sagt, dass zwei Ereignisse voneinander abhängig sind, wenn das Eintreten des einen vom Eintreten des anderen abhängt.
Die zweite Bedeutung des Wortes betrifft die **Beziehung,** im Allgemeinen eine funktionale Beziehung, die zwischen **Zufallsvariablen bestehen kann.**

Diese Abhängigkeit kann gemessen werden, und in den meisten Fällen erfordern diese Messungen eine Zufallsvariable- **Kovarianz.**

Das am häufigsten verwendete Maß für die Abhängigkeit ist der **Korrelationskoeffizient.** Der **Nullhypothese,** die besagt, dass die beiden **Zufallsvariablen** *X* und *Y* unabhängig sind abhängig ist, kann anhand der **Alternativhypothese** der Abhängigkeit getestet werden. Zu den zu diesem Zweck verwendeten Tests gehören der **Chi-Quadrat-Unabhängigkeitstest** sowie Tests, die auf der **Spearman-Rangkorrelation** basieren.

Machine Translated by Google

Beziehungskoeffizient und der **Kendall-Rang Korrelationskoeffizient.**

WEITERLESEN

- **Chi-Quadrat-Test der Unabhängigkeit**
- **Korrelationskoeffizient**
- **Kovarianz**
- **Kendall-Rangkorrelationskoeffizient**
- **Rangkorrelationskoeffizient nach Spearman**
- **Test der Unabhängigkeit**

Abhängige Variable

Die abhängige Variable (oder Antwortvariable) in einem Regressionsmodell **ist** die **Variable ,** von der angenommen wird, dass sie abhängig von den anderen Variablen, die als unabhängige Variablen bezeichnet werden und in die Analyse einbezogen werden, variiert.

WEITERLESEN

- **Unabhängige Variable**
- **Multiple lineare Regression**
- **Regressionsanalyse**
- **Einfache lineare Regression**

Beschreibend Epidemiologie

Bei der deskriptiven Epidemiologie geht es um die Beschreibung der Häufigkeiten und Muster von Erkrankungen in der Bevölkerung.
Es basiert auf der Sammlung gesundheitsbezogener Daten Informationen (in Sterbetafeln, Morbiditätsregistern, meldepflichtigen Krankheiten ...) und über Informationen, die Auswirkungen auf die Gesundheit der Bevölkerung haben können (Daten zur Luftverschmutzung, Risikoverhalten ...), die zur Ermittlung von a statistisches Bild des allgemeinen Gesundheitszustands der Bevölkerung.

GESCHICHTE

Siehe **Epidemiologie.**

MATHEMATISCHE ASPEKTE

Siehe **Ursache und Wirkung in der Epidemiologie, Chancen und Chancenverhältnis, relatives Risiko, zuschreibbares Risiko, vermeidbares Risiko, Inzidenzrate, Prävalenzrate.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Auch wenn die deskriptive Epidemiologie nur „elementare“ Informationen liefert, ist sie dennoch von entscheidender Bedeutung: • Sie ermöglicht uns die Untersuchung des Ausmaßes und der Muster von Gesundheitsphänomenen (durch Untersuchung **ihrer Prävalenz und ihres Auftretens)** und sie erleichtert die epidemiologische Überwachung.

• Es hilft bei der Entscheidungsfindung im Zusammenhang mit der Planung und Verwaltung von Gesundheitseinrichtungen und -programmen, • Es kann zu Hypothesen über die Risikofaktoren einer Krankheit führen (Beispiel: die erhöhte Inzidenz von Hautkrebs im Süden von Frankreich, was vermutlich mit der im Süden Frankreichs im Vergleich zum Norden höheren Intensität der Sonnenstrahlen zusammenhängt.

Die beschreibende Epidemiologie kann nicht verwendet werden, um eine Ursache mit einer Wirkung in Beziehung zu setzen; sie ist kein Vorhersageinstrument.

WEITERFÜHRENDE LITERATUR

Siehe **Epidemiologie.**

VERWEISE

Siehe **Epidemiologie.**

Beschreibende Statistik

Die in einer Studie gesammelten Informationen können häufig unterschiedliche Formen annehmen, beispielsweise Häufigkeitsdaten

D

Machine Translated by Google

(zum Beispiel die Anzahl der für einen Kandidaten bei Wahlen abgegebenen Stimmen) und Skalendaten. Diese Daten sind häufig zunächst so angeordnet oder organisiert, dass sie schwer zu lesen und zu interpretieren sind.

Die deskriptive **Statistik** bietet uns einige Verfahren, die es uns ermöglichen, Daten in einer lesbaren und sinnvollen Form darzustellen.

Einige dieser Verfahren ermöglichen es uns, eine **grafische Darstellung** der Daten zu erhalten, beispielsweise in den folgenden Formen:

- **Histogramm,**
- **Balkendiagramm,**
- **Kreisdiagramm,**
- **Stamm- und Blattdiagramm,**
- **Boxplot** usw., während andere es uns ermöglichen, eine zu erhalten Parametersatz, der wichtige Eigenschaften der Basisdaten zusammenfasst:
- **Mittelwert,**
- **Standardabweichung,**
- **Korrelationskoeffizient,**
- **Indexzahl** usw.

Zur deskriptiven Statistik gehören auch Methoden der **Datenanalyse,** etwa die **Korrespondenzanalyse,** die darin bestehen, die Zusammenhänge zwischen den Zeilen und den Spalten einer **Kontingenztabelle grafisch darzustellen.**

GESCHICHTE

Die erste bekannte Form der deskriptiven Statistik war die **Volkszählung.** Volkszählungen wurden von den Herrschern antiker Zivilisationen angeordnet, die ihre Untertanen zählen und ihre Berufe und Güter überwachen wollten.

Siehe **Statistiken** und **grafische Darstellung.**

WEITERLESEN

Box-Plot

Volkszählung

Korrespondenzanalyse

Datenanalyse

Dendrogramm

Grafische Darstellung

Histogramm

Indexnummer

Liniendiagramm

Kuchendiagramm

Stamm-Blatt-Diagramm

Versuchsplanung

Das Entwerfen eines Experiments ist in gewisser Weise wie das Programmieren des **Experiments .** Jeder am Experiment beteiligte **Faktor** kann eine bestimmte Anzahl unterschiedlicher Werte annehmen (sogenannte Stufen), und der verwendete Versuchsaufbau gibt die Stufen eines oder mehrerer Faktoren (oder Kombinationen von Faktoren) an, die im Experiment verwendet werden.

GESCHICHTE

Experimentelle Entwürfe wurden erstmals in den 1920er Jahren eingesetzt, hauptsächlich im landwirtschaftlichen Bereich. Sir **Fisher, Ronald Aylmer** war der erste, der bei der Planung von Experimenten mathematische Statistiken verwendete. Im Jahr 1926 verfasste er einen Aufsatz, in dem er die Prinzipien des experimentellen Designs in nichtmathematischen Begriffen darlegte.

Federer, WT und Balaam, LN (1973) lieferten eine sehr detaillierte Bibliographie der Literatur zum experimentellen Design vor 1969 mit 8000 Referenzen.

DOMÄNEN UND EINSCHRÄNKUNGEN

Ziel des Versuchsdesigns ist es, das effizienteste und wirtschaftlichste Meth zu finden

Hinweise, die es uns ermöglichen, solide und angemessene Schlussfolgerungen aus den Ergebnissen der Erfahrung zu ziehen ment.

Machine Translated by Google

Das am häufigsten angewandte Experiment
Die Designs sind **vollständig zufällig**
Design, das **randomisierte Blockdesign** und
das **lateinische Quadratdesign.**
Jeder Entwurf impliziert eine andere mathematische
Analyse als die anderen
Designs, da die Designs wirklich übereinstimmen
zu verschiedenen mathematischen Modellen. Beispiele
Zu diesen Analysetypen gehört auch **die Varianz**
Analyse, Kovarianzanalyse und **Regressionsanalyse.**

WEITERLESEN

Varianzanalyse
Experiment
Modell
Optimales Design
Regressionsanalyse

VERWEISE

Federer, WT, Balaam, LN: Bibliographie
über Experiment und TreatmentDesign vor 1968.
Hafner, New York (1973)

Fisher, RA: Die Anordnung von Feldexperimenten. J.
Landwirtschaftsministerium. **33,** 503–513
(1926)

Bestimmend

Jeder quadratischen **Matrix** *A* der Ordnung *n* ist eine spezielle Zahl zugeordnet, die als bekannt ist Determinante der Matrix A.

MATHEMATISCHE ASPEKTE

Betrachten Sie *A = (aij),* eine quadratische **Matrix** der Ordnung n. Die Determinante von A, bezeichnet als det(A) oder |A|, ist wie folgt gegeben
Summe:

$$\det(A) = (\pm)a1i \cdot a2j \cdot \cdots \cdot anr ,$$

wobei die Summe über alle **Permutationen** des zweiten Index gebildet wird · Das Vorzeichen ist positiv, wenn die Anzahl **der Inversionen in** *(i, j, . . .,* r) gerade ist (eine gerade Permutation) und negativ, wenn die Zahl ungerade ist (eine ungerade Permutation).
Die Determinante kann auch definiert werden durch sich entlang einer Linie oder Spalte „entwickeln". Für Beispielsweise können wir die Matrix entlang der Matrix entwickeln erste Linie:

$$\det(A) = \sum_{j=1}^{N} (ÿ1)^{1+j} \cdot a1j \cdot A1j ,$$

wobei A1j die Determinante des Quadrats ist „Submatrix" der Ordnung *n* ÿ 1, erhalten aus *A*
durch Löschen der ersten Zeile und der j-ten Spalte.

A1j wird als Cofaktor des Elements *a1j bezeichnet.*
Da wir die Zeile oder beliebig wählen können
Spalte können wir schreiben:

$$\det(A) = \sum_{j=1}^{N} (ÿ1)^{i+j} \cdot aij \cdot Aij$$

für ein festes *i* (das sich entlang der i-ten Linie entwickelt),
oder

$$\det(A) = \sum_{i=1}^{N} (ÿ1)^{i+j} \cdot aij \cdot Aij$$

für ein festes *j* (entlang der j-ten Spalte entwickelnd).

Diese zweite Art, die Determinante zu definieren, ist rekursiv, weil die Determinante von
Eine quadratische Matrix der Ordnung *n* wird mit berechnet die Determinanten der Ordnungsmatrizen
n ÿ 1.

Eigenschaften der Determinante

- Die Determinante einer quadratischen **Matrix** *A* ist gleich der Determinante **seiner transponierten** Matrix *A :*

$$\det(A) = \det(A) .$$

Machine Translated by Google

• Werden in Matrix A zwei Zeilen (bzw. zwei Spalten) vertauscht, ändert sich das Vorzeichen der Determinante.
• Wenn alle Elemente einer Zeile (oder einer Spalte) Null sind, ist auch die Determinante Null. • Wenn ein Vielfaches einer Zeile aus Matrix *A* zu einer anderen Zeile aus A addiert wird, bleibt die Determinante dieselbe. Das bedeutet, wenn

Wenn zwei Linien von A identisch sind, ist es einfach, eine Linie zu erhalten, auf der sich alle Elemente befinden

Null, indem man einfach eine Linie von der anderen subtrahiert; Aufgrund der vorherigen Eigenschaft ist die Determinante von *A* Null. Die gleiche Situation tritt auf, wenn eine Zeile eine Multizeile ist

ple von einem anderen. • Die Determinante einer Matrix, die nur Nullen unter (oder über) der Diagonale hat, ist gleich dem Produkt der Elemente auf der Diagonale (eine solche Matrix wird als „dreieckig“ bezeichnet).
• Betrachten Sie *A* und B, zwei quadratische Matrizen der Ordnung n. Die Determinante des Produkts der beiden Matrizen ist gleich dem Produkt der Determinanten der beiden Matrizen.

$$\det(A \cdot B) = \det(A) \cdot \det(B) .$$

BEISPIELE

1) Betrachten Sie A, eine quadratische **Matrix** der Ordnung 2:

$$A = \begin{matrix} ab \\ CD \end{matrix}$$

$$\det(A) = a \cdot d \ddot{y} b \cdot c .$$

2) Betrachten Sie B, eine quadratische Matrix der Ordnung 3:

$$B = \begin{matrix} 3\ 21 \\ \ddot{y}\ \ 4\ \ddot{y}1\ 0 \\ \ddot{y}\ \ 2\ \ddot{y}2\ 0 \end{matrix} \qquad \ddot{y}\ \ddot{y} .$$

Durch die Entwicklung entlang der letzten Spalte:

$$\det(B) = 1 \cdot \begin{matrix} 4\ \ddot{y}1 \\ 2\ \ddot{y}2 \end{matrix} + 0 \cdot \begin{matrix} 3 & 2 \\ 2\ \ddot{y}2 \end{matrix}$$

$$+ 0 \cdot \begin{matrix} 3 & 2 \\ 4\ \ddot{y}1 \end{matrix}$$

$$= 1 \cdot \{4 \cdot (\ddot{y}2) \ddot{y} (\ddot{y}1) \cdot 2\}$$

$$= \ddot{y}8 + 2$$

$$= \ddot{y}6 .$$

WEITERLESEN

- **Umkehrung**
- **Matrix**
- **Permutation**

Abweichung

Der Begriff der Abweichung beschreibt den Unterschied zwischen einem beobachteten **Wert** und einem festen Wert

Wert aus der Menge möglicher Werte einer **quantitativen Variablen.**
Dieser feste **Wert** ist oft die **Arithmetik Mittelwert** der Wertemenge oder **Median.**

MATHEMATISCHE ASPEKTE

Betrachten Sie die Wertemenge **x1 ,** *x2,...,xn.* Das **arithmetische Mittel** dieser Werte wird angegeben durch $\bar{x}$. Die Abweichung eines gegebenen Wertes *xi* vom arithmetischen Mittel ist gleich:

$$\text{Abweichung} = (xi \ddot{y} \bar{x}) .$$

In ähnlicher Weise ist die Abweichung eines Werts *xi* in Bezug auf den **Median** gleich, wenn der Median dieser **Werte** mit m bezeichnet wird :

$$\text{Abweichung} = (xi \ddot{y} m) .$$

WEITERLESEN

- **Mittlere absolute Abweichung**
- **Standardabweichung**
- **Varianz**

Machine Translated by Google

Dichotome Variable

Eine **Variable** heißt dichotom, wenn sie dies kann

Nimm nur zwei Werte.

Das einfachste Beispiel ist die **qualitative kategoriale Variable** „Geschlecht“, die zwei Werte annehmen kann, „männlich“ und „weiblich“.

Beachten Sie, dass quantitative Variablen immer reduziert und dichotomisiert werden können. Die Variable „Umsatz“ kann beispielsweise auf zwei Kategorien reduziert werden: „geringer Umsatz“ und „hoher Umsatz“. enue“.

WEITERLESEN

- Binärdaten
- Kategorie
- Qualitative kategoriale Variable
- Variable

Dichotomie

Dichotomie ist die Aufteilung der Individuen einer **Population oder** einer **Stichprobe** in zwei Gruppen als Funktion vorgegebener Kriterien.
Eine **Variable** heißt **dichotom,** wenn sie

kann nur zwei **Werte annehmen.** Solche Daten werden **Binärdaten genannt.**

WEITERLESEN

- Binärdaten
- Dichotome Variable

Diskrete Verteilungsfunktion

Die Verteilungsfunktion eines diskreten **Laufs**
Die dom-Variable ist für alle reellen Zahlen definiert als **Wahrscheinlichkeit ,** dass die Zufallsvariable einen **Wert** annimmt , der kleiner oder gleich dieser reellen Zahl ist.

GESCHICHTE

Siehe **Wahrscheinlichkeit.**

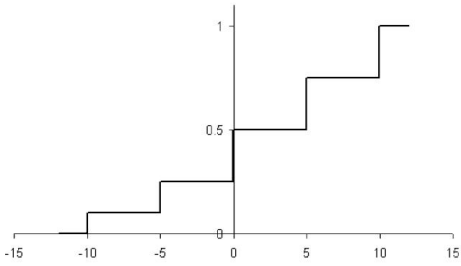
MATHEMATISCHE ASPEKTE

Die Funktion F, definiert durch

$$F(b) = P(X \leq b) .$$

heißt Verteilungsfunktion der dis
Kreta **Zufallsvariable** X.
Bei einer gegebenen reellen Zahl b entspricht die Verteilungsfunktion daher der Wahrscheinlichkeit , dass *X* kleiner oder gleich b ist.
Die Verteilungsfunktion lässt sich grafisch auf einem Achsensystem darstellen. Die verschiedenen **Werte** der diskreten **Zufallsvariable**

Auf der Abszisse werden die **möglichen** *X-Werte* und auf der Ordinate die kumulativen **Wahrscheinlichkeiten** angezeigt , die den verschiedenen Werten von X, *F(x),* entsprechen.



Für den Fall, dass die möglichen **Werte** der diskreten **Zufallsvariablen** b1, b2, sind . . . mit b1 < b2 < ist die diskrete **Verteilungsfunktion** eine Stufenfunktion.
F(bi) ist über das **Intervall** *[bi,* bi+1[konstant.

Eigenschaften des Diskreten Verteilungsfunktion

1. *F* ist eine nicht abnehmende Funktion; mit anderen Worten, wenn *a* < b, dann ist *F(a)* ≤ *F(b)*
2. *F* nimmt **Werte** aus dem **Intervall** [0, 1] an.

Machine Translated by Google

3. $\lim_{b \to -\infty} F(b) = 0$.
4. $\lim_{b \to \infty} F(b) = 1$.

BEISPIELE

Stellen Sie sich ein **Zufallsexperiment** vor , bei dem zwei Würfel gleichzeitig geworfen werden. Betrachten Sie die diskrete **Zufallsvariable** X, entsprechend der Gesamtpunktzahl der beiden Würfel.

Lassen Sie uns **nach der Wahrscheinlichkeit** des **Ereignisses** suchen $\{X \leq 7\}$, was per Definition der **Wert** von ist die **Verteilungsfunktion** für $x = 7$.

Die diskrete **Zufallsvariable** X nimmt ihren Wert an **Werte** aus der Menge $E = \{2, 3, ..., 12\}$.

Die mit jedem Wert von X verbundenen **Wahrscheinlichkeiten** sind in der folgenden Tabelle angegeben:

X	2	3	4	5	6
P(X)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$

X	7	8	9	10	11	12
P(X)	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

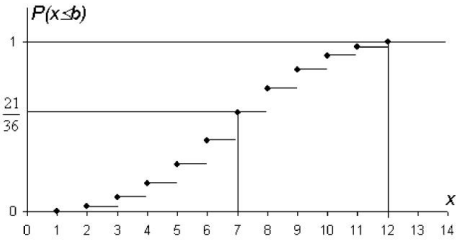
Um die **Verteilungsfunktion festzulegen,** haben wir müssen für jeden Wert b von X berechnen: die Summe der Wahrscheinlichkeiten für alle Werte kleiner oder gleich b. Deshalb schaffen wir eine neue Tabelle mit den kumulativen Wahrscheinlichkeiten:

B	2	3	4	5	6	7
P(X ≤ B)	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$

B	8	9	10	11	12	
P(X ≤ B)	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$	= 1

Die **Wahrscheinlichkeit** des **Ereignisses** ist $\{X \leq 7\}$ also gleich $\frac{21}{36}$.

Wir können die **Verteilungsfunktion** darstellen der diskreten **Zufallsvariablen** X wie folgt Tiefs:



WEITERLESEN

Wahrscheinlichkeit

Wahrscheinlichkeitsfunktion

Zufälliges Experiment

Wert

Diskrete Wahrscheinlichkeitsverteilung

Wenn jeder mögliche **Wert** ein diskreter **Zufall** ist **Da die Variable** mit einer bestimmten **Wahrscheinlichkeit** verknüpft ist , können wir die diskrete Wahrscheinlichkeit ermitteln Verteilung dieser Zufallsvariablen.

MATHEMATISCHE ASPEKTE

Die Wahrscheinlichkeitsverteilung einer diskreten **Die Zufallsvariable** X ist durch ihre **Wahrscheinlichkeitsfunktion** $P(x)$ oder ihre **Verteilungsfunktion** $F(x)$ gegeben .

Es kann allgemein durch seine charakterisiert werden **erwarteter Wert:**

$$E[X] = \sum_{D} x \cdot P(X = x) ,$$

und seine **Varianz:**

$$Var(X) = \sum_{D} (x - E[X])^2 \cdot P(X = x) ,$$

wobei D die Menge darstellt, aus der X kann nimm seine **Werte.**

Machine Translated by Google

BEISPIELE

Die diskreten Wahrscheinlichkeitsverteilungen, die Am häufigsten werden die **Bernoul-li-Verteilung,** die **Binomialverteilung und** die **negative Binomialverteilung,** die **geometrische Verteilung,** das **Multinomial Verteilung,** die **hypergeometrische Verteilung** und die **Poisson-Verteilung.**

WEITERLESEN

Bernoulli-Verteilung
Binomialverteilung
Kontinuierliche Wahrscheinlichkeitsverteilung
Diskrete Verteilungsfunktion
Erwarteter Wert
Geometrische Verteilung
Hypergeometrische Verteilung
Gemeinsame Wahrscheinlichkeitsverteilungsfunktion
Multinomiale Verteilung
Negative Binomialverteilung
Poisson-Verteilung
Wahrscheinlichkeit
Wahrscheinlichkeitsverteilung
Wahrscheinlichkeitsfunktion
Zufällige Variable
Varianz einer Zufallsvariablen

VERWEISE

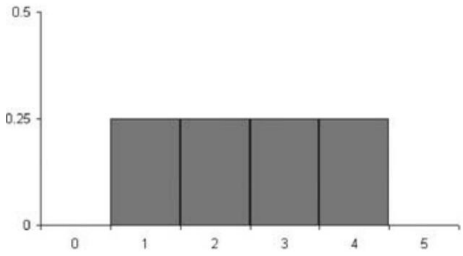
Johnson, NL, Kotz, S.: Verteilungen in Statistik: Diskrete Verteilungen. Wiley, New York (1969)

Diskrete Uniform Verteilung

Die diskrete Gleichverteilung ist eine **dis Kreta-Wahrscheinlichkeitsverteilung.** Die zugehörige **stetige Wahrscheinlichkeitsverteilung** ist die (stetige) **Gleichverteilung.**

Betrachten Sie *n* **Ereignisse,** von denen jedes das hat gleiche **Wahrscheinlichkeit** $P(X = x) = \frac{1}{N}$; der **rannte Die Domvariable** *X* folgt einer diskreten Uniform Verteilung und ihre **Wahrscheinlichkeitsfunktion** ist:

$$P(X = x) = \begin{cases} \frac{1}{N}, & \text{wenn } x = x1, x2,\ldots,xn \\ 0, & \text{wenn nicht} \end{cases}.$$



Diskrete Gleichverteilung, n = 4

MATHEMATISCHE ASPEKTE

Der **erwartete Wert** der diskreten Uniform Verteilung über die Menge vonfirstnnaturalnum bers ist per Definition gegeben durch:

$$E[X] = \sum_{x=1}^{N} x \cdot P(X = x)$$
$$= \frac{1}{N} \cdot \sum_{x=1}^{N} X$$
$$= \frac{1}{N} \cdot (1 + 2 + \cdots + n)$$
$$= \frac{1}{N} \cdot \frac{n2 + n}{2}$$
$$= \frac{n + 1}{2}.$$

Die **Varianz** dieser diskreten Gleichverteilung ist gleich:

$$\mathrm{Var}(X) = E\,X2\,\ddot{y}\,(E[X])^2.$$

Seit

$$E[X2] = \sum_{x=1}^{N} x2 \cdot P(X = x)$$
$$= \frac{1}{N} \cdot \sum_{x=1}^{N} x2$$

Machine Translated by Google

$$= \frac{1}{n} \cdot 12 + 22 + \cdots + n2$$

$$= \frac{1}{N} \cdot \frac{n(n+1)(2n+1)}{6}$$

$$= \frac{2n2 + 3n + 1}{6},$$

Und

$$(EX])^2 = \frac{(n+1)2}{22},$$

wir haben:

$$\text{Var}(X) = \frac{2n2 + 3n + 1}{} \quad \frac{(n+1)2}{22}$$

$$= \frac{6\ n2\ \ddot{y}\ 1}{12}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die diskrete Gleichverteilung ist häufig

Wird verwendet, um aus beliebigen **Zahlen Zufallszahlen** zu generieren

diskrete oder kontinuierliche **Wahrscheinlichkeitsverteilung.**

BEISPIELE

Betrachten Sie die **Zufallsvariable** X, den Score erhält man durch einen Würfelwurf. Wenn der Würfel ist nicht geladen, die **Wahrscheinlichkeit,** welche zu erhalten Die bestimmte Punktzahl ist gleich $\frac{1}{6}$. Daher haben wir haben:

$$P(X = 1) = \frac{1}{6}$$

$$P(X = 2) = \frac{1}{6}$$

...

$$P(X = 6) = \frac{1}{6}.$$

Die Anzahl möglicher **Ereignisse** beträgt $n = 6$. Wir habe daher:

$$P(X = x) = \begin{cases} \frac{1}{6}, & \text{für } x = 1, 2, \ldots, 6 \\ 0, & \text{wenn nicht} \end{cases}.$$

Mit anderen Worten folgt die Zufallsvariable *X* der diskreten Gleichverteilung.

WEITERLESEN

Diskrete Wahrscheinlichkeitsverteilung
Gleichmäßige Verteilung

Streuung

Siehe **Streuungsmaß.**

Distanz

Die Entfernung ist eine numerische Beschreibung der Abstand zwischen zwei Objekten. Ein Abstand entspricht also einer reellen Zahl:

- Null, wenn beide Objekte gleich sind;
- Streng positiv, wenn nicht.

MATHEMATISCHE ASPEKTE

Betrachten Sie drei Objekte X, *Y* und Z und die Abstand zwischen *X* und Y, *d(X,* Y).

Eine Distanz hat folgende Eigenschaften:

- Es ist positiv,

$$d(X, Y) > 0,$$

oder Null genau dann, wenn die Objekte die sind Dasselbe

$$d(X, Y) = 0 \ \ddot{y}\ X = Y.$$

- Es ist symmetrisch, das heißt

$$d(X, Y) = d(Y, X).$$

- Es überprüft die folgende Ungleichung, genannt die Dreiecksungleichung:

$$d(X, Z) \ \ddot{y}\ d(X, Y) + d(Y, Z).$$

Dies besagt, dass die Entfernung von einem Objekt zu einem anderen ist kleiner oder gleich der Entfernung, die man beim Durchqueren eines Dritten erhält Objekt.

Betrachten Sie *X* und Y, die zwei **Vektoren** sind mit *n* Komponenten,

$$X = (xi) \text{ und } Y = (yi),$$

für $i = 1, 2, \ldots, n.$

Machine Translated by Google

Eine Familie von Entfernungen, die häufig verwendet wird
Ein solcher Fall ist unten angegeben:

$$d(X, Y) = \sqrt[p]{\sum_{i=1}^{N} |x_i - y_i|^p} \text{ mit } p \geq 1 .$$

Diese Abstände werden Minkowski-Dis genannt
Beziehungen.

Der Abstand für $p = 1$ wird als absolut bezeichnet
Abstand oder die L1- Norm.
Der Abstand für $p = 2$ heißt
Euklidischer Abstand, und er ist definiert durch:

$$d(X, Y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} .$$

Der euklidische Abstand ist der am häufigsten
verwendete Abstand.
Bei der Art der verwendeten Entfernung ist dies nicht der Fall
spezifiziert, ist im Allgemeinen der euklidische Abstand
gemeint.
Es gibt auch die gewichteten Minkowski-Abstände,
definiert durch:

$$d(X, Y) = \sqrt[p]{\sum_{i=1}^{N} w_i \cdot |x_i - y_i|^p} \text{ mit } p \geq 1 ,$$

wobei wi,i = *1,...,n die* verschiedenen darstellen
Gewichte, deren Summe 1 ergibt.
Wenn $p = 1$ oder 2 ist, erhalten wir die Gewichtung
absolute bzw. euklidische Distanzen.

DOMÄNEN UND EINSCHRÄNKUNGEN

Für eine Menge von *r* verschiedenen Objekten X1, *X2,...,Xr,*
der Abstand zwischen Objekt *i* und Objekt *j* für
jedes Paar *(Xi, Xj)* kann berechnet werden. Diese
Distanzen werden verwendet, um eine **Distanztabelle** zu erstellen
oder allgemeiner eine Unähnlichkeitstabelle mit
Begriffen, die verallgemeinert werden können

$$d_{ij} = d(X_i, X_j) , \quad \text{für } i, j = 1, 2,...,r,$$

wie sie in Methoden der **Clusteranalyse verwendet werden.**
Beachten Sie, dass der Begriff **Maß für Unähnlichkeit ist**
ty wird verwendet, wenn die Dreiecksungleichung nicht vorliegt
befriedigt.

D

BEISPIELE

Betrachten Sie das Beispiel anhand der Noten
von fünf Studenten in ihrem erhalten
Prüfungen in Englisch, Französisch, Mathematik und
Physik.

Das sind Noten, die jeden Wert annehmen können
1 bis 6 sind in der folgenden Tabelle zusammengefasst:

	Englisch	Französisch	Mathematik	Physik
Alain	5.0	3.5	4,0	4.5
Jean	3.5	4,0	5,0	5.5
März	4.5	4.5	4,0	3.5
Paul	6,0	5.5	5.5	5,0
Pierre	4,0	4.5	2.5	3.5

Der euklidische Abstand zwischen Alain und
Jean, bezeichnet mit d2, ist gegeben durch:

$$\begin{aligned} &d2(\text{Alain, Jean}) \\ &= \sqrt{(5{,}0 - 3{,}5)^2 + \cdots + (4{,}5 - 5{,}5)^2} \\ &= \sqrt{2{,}25 + 0{,}25 + 1 + 1} \\ &= \sqrt{4{,}5} = 2{,}12 . \end{aligned}$$

Der absolute Abstand zwischen Alain und
Jean, bezeichnet mit d1, ist gegeben durch:

$$\begin{aligned} d1(\text{Alain, Jean}) &= |5{,}0 - 3{,}5| + \cdots \\ &\quad + |4{,}5 - 5{,}5| \\ &= 1{,}5 + 0{,}5 + 1 + 1 \\ &= 4{,}0 . \end{aligned}$$

Wenn wir die Berechnungen fortsetzen, erhalten wir:

$$d2(\text{Alain, Marc}) = \sqrt{2{,}25} = \sqrt{1{,}5}$$

$$d2(\text{Alain, Paul}) = \sqrt{7{,}5} = 2{,}74$$

Machine Translated by Google

d2(Alain, Pierre) = ÿ 5,25 = 2,29

d2(Jean, Marc) = ÿ 6,25 = *2,5*

d2(Jean, Paul) = ÿ 9 = 3

d2(Jean, Pierre) = ÿ *10,75* = 3,28

d2(Marc, Paul) = ÿ 7,75 = 2,78

d2(Marc, Pierre) = ÿ *2,5* = 1,58

d2(Paul, Pierre) = ÿ *16,25* = 4,03

d1(Alain, Marc) = *2,5*

d1(Alain, Paul) = *5,0*

d1(Alain, Pierre) = *4,5*

d1(Jean, Marc) = *4,5*

d1(Jean, Paul) = *5,0*

d1(Jean, Pierre) = *5,5*

d1(Marc, Paul) = *6,5*

d1(Marc, Pierre) = *2,0*

d1(Paul, Pierre) = *7,5* .

Die euklidische **Distanztabelle** wird erhalten durch Platzieren Sie diese Ergebnisse in der folgenden Tabelle:

	Alain	Jean	Marc	Paul	Pierre
Alain	0	2,12	1,5	2,74	2,29
Jean	2,12	0	2.5	3	3.28
März	1.5	2.5	0	2,78	1,58
Paulus	2,74	3	2,78	0	4.03
Pierre	2,29	3,28	1,58	4,03	0

Auf ähnliche Weise können wir die folgende **Tabelle** der absoluten Distanzen erhalten :

	Alain	Jean	Marc	Paul	Pierre
Alain	0	4	2.5	5	4.5
Jean	4	0	4.5	5	5.5
März	2.5	4,5	0	6,5	2
Paulus	5	5	6.5	0	7.5
Pierre	4.5	5,5	2	7,5	0

Beachten Sie, dass die Reihenfolge der Nähe der Personen je nach gewählter Person unterschiedlich sein kann.

Distanz. Wenn beispielsweise der euklidische Abstand verwendet wird, ist Alain am nächsten dran Marc (d2(Alain, Marc) = 1,5), wohingegen Alain belegt in der absoluten Kategorie den zweiten Platz Abstand verwendet wird, (d1(Alain, Marc) = 2,5), und es ist Pierre, der Marc am nächsten steht einen Abstand von 2 *(d1(Pierre.Marc)* = 2).

WEITERLESEN

- **Chi-Quadrat-Abstand**
- **Clusteranalyse**
- **Entfernungstabelle**
- **Mahalanobis-Entfernung**
- **Maß der Unähnlichkeit**

VERWEISE

Abdi, H.: Distanz. In: Salkind, NJ (Hrsg.) Enzyklopädie der Messung und Statistik. Salbei, tausend Eichen (2007)

Distanztabelle

Für eine Menge von *r* Individuen die Distanztabelle ist eine quadratische **Matrix** der Ordnung r, und der allgemeine Term *(i,* j) entspricht dem **Abstand** dazwischen das i-te und das j-te Individuum. Danke an die Eigenschaften der Distanz, die Distanztabelle ist symmetrisch und seine Diagonale ist Null.

MATHEMATISCHE ASPEKTE

Angenommen, wir haben *r* Individuen, bezeichnet X1, *X2,...,Xr;* Die aus diesen *r* Individuen erstellte Distanztabelle ist ein Quadrat **Matrix** *D* der Ordnung *r* und der allgemeine Term ist

$$dij = d\,Xi,\,Xj\,,$$

wobei *d Xi, Xj* der **Abstand** zwischen den ist das i-te und das j-te Individuum.

Die Eigenschaften der Entfernung erlauben uns eine Aussage Das:

Machine Translated by Google

- Die Diagonalelemente *dii* von *D* sind Null weil *d (Xi, Xi)* = 0 (für *i = 1, 2, ...,R),*
- Dis symmetrisch; das heißt, *dij = dji* für jeden *i* und für Werte von *j* von 1 bis r.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Verwendung einer Distanztabelle **bietet eine Art Clusteranalyse,** deren Ziel die Identifizierung ist die nächsten oder ähnlichsten Personen in um sie gruppieren zu können.
Die in die Distanztabelle eingegebenen Distanzen werden im Allgemeinen als euklidische **Abstände angenommen,** es sei denn, es handelt sich um einen anderen Abstandstyp angegeben wurde.

BEISPIELE

Nehmen wir das Beispiel der Noten von fünf Studenten in vier Prüfungen erworben: Englisch, Französisch, Mathematik und Physik.
Diese Noten können ein beliebiges Vielfaches davon annehmen 0,5 von 1 bis 6, sind in der folgenden Tabelle zusammengefasst:

	Englisch	Französisch	Mathematik	Phys
Alain	5.0	3.5	4,0	4.5
Jean	3.5	4,0	5,0	5.5
März	4.5	4.5	4,0	3.5
Paul	6.0	5.5	5.5	5,0
Pierre	4.0	4.5	2.5	3.5

Wir versuchen, die **Entfernungen** zu messen getrennte Schülerpaare. Um dies zu tun, haben wir Vergleichen Sie zwei Zeilen der Tabelle (entsprechend zwei Schülern) gleichzeitig: Um genau zu sein, wird jede Note in der zweiten Zeile von der entsprechenden Note in der ersten Zeile abgezogen Linie.

Wenn wir zum Beispiel die Entfernung berechnen von Alain bis Jean unter Verwendung des Euklidischen dis tanz, wir bekommen:

$$d(Alain.Jean) = ((5{,}0 \ddot{y} 3{,}5)^2 + (3{,}5 \ddot{y} 4{,}0)^2 + (4.0 \ddot{y} 5{,}0)^2 + (4{,}5 \ddot{y} 5{,}5)^2)^{\frac{1}{2}}$$
$$= \ddot{y} 2{,}25 + 0{,}25 + 1 + 1$$
$$= \ddot{y} \overline{4{,}5} = 2{,}12 .$$

Ebenso der Abstand zwischen Alain und Marc gleicht:

$$d(Alain, Marc) = ((5{,}0 \ddot{y} 4{,}5)^2 + (3{,}5 \ddot{y} 4{,}5)^2 + (4{,}0 \ddot{y} 4{,}0)^2 + (4{,}5 \ddot{y} 3{,}5)^2)^{\frac{1}{2}}$$
$$= \ddot{y} 0{,}25 + 1 + 0 + 1$$
$$= \ddot{y} \overline{2{,}25} = 1{,}5 .$$

Usw,

$$d(Alain, Paul) = \ddot{y} \overline{7{,}5} = \overline{2{,}74}$$
$$d(Alain, Pierre) = \ddot{y} 5{,}25 = \overline{2{,}29}$$
$$d(Jean, Marc) = \ddot{y} 6{,}25 = \overline{2{,}5}$$
$$d(Jean, Paul) = \ddot{y} 9 = \overline{3}$$
$$d(Jean, Pierre) = \ddot{y} \overline{10{,}75} = 3{,}28$$
$$d(Marc, Paul) = \ddot{y} 7{,}75 = \overline{2{,}78}$$
$$d(Marc, Pierre) = \ddot{y} 2{,}5 = \overline{1{,}58}$$
$$d(Paul, Pierre) = \ddot{y} \overline{16{,}25} = 4{,}03 .$$

Indem wir diese Ergebnisse in einer Tabelle organisieren, können wir Erhalten Sie die folgende Abstandstabelle:

	Alain	Jean	Marc	Paul	Pierre
Alain	0	2,12	1,5	2,74	2,29
Jean	2,12	0	2.5	3	3.28
März	1.5	2,5	0	2,78	1,58
Paulus	2,74	3	2,78	0	4.03
Pierre	2,29	3,28	1,58	4,03	0

Machine Translated by Google

Wir können sehen, dass die Werte entlang der Diagonale gleich Null sind und dass die **Matrix** symmetrisch ist. Dies bedeutet, dass die Entfernung von Jeder Schüler zum gleichen Schüler ist Null, und der Abstand, der den ersten Schüler vom zweiten trennt, ist identisch mit dem Abstand

das trennt das zweite vom ersten. Also *d (Xi, Xi)* = 0 und *d Xi, Xj = d Xj, Xi* .

WEITERLESEN

- **Clusteranalyse**
- **Dendrogramm**
- **Distanz**
- **Matrix**

Verteilungsfunktion

Die Verteilungsfunktion einer **Zufallsvariable** ist für jede reelle Zahl definiert als **Wahrscheinlichkeit ,** dass die Zufallsvariable einen **Wert** annimmt , der kleiner oder gleich dieser reellen Zahl ist.

Abhängig davon, ob die **Zufallsvariable** diskret oder stetig ist, erhalten wir eine **diskrete Verteilungsfunktion** oder eine **con kontinuierliche Verteilungsfunktion.**

GESCHICHTE

Siehe **Wahrscheinlichkeit.**

MATHEMATISCHE ASPEKTE

Wir definieren die Verteilungsfunktion eines **ran Domvariable** *X* beim Wert *b* (mit 0 < *b* < ÿ) durch:

$$F(b) = P(X \leq b) .$$

WEITERLESEN

- **Kontinuierliche Verteilungsfunktion**
- **Diskrete Verteilungsfunktion**
- **Wahrscheinlichkeit**
- **Zufällige Variable**
- **Wert**

Punktdiagramm

Das Punktdiagramm ist eine Art **Häufigkeitsdiagramm .** Wenn der **Datensatz** relativ klein ist (bis zu etwa 25 Datenpunkte), können die Daten und ihre **Häufigkeiten** in einem Punktdiagramm dargestellt werden. Diese Art der **grafischen Darstellung** eignet sich besonders zur Identifizierung **von Ausreißern.**

MATHEMATISCHE ASPEKTE

Das Punktdiagramm wird auf folgende Weise erstellt. Es wird eine Skala erstellt, die alle **Werte** im **Datensatz** enthält . Jedes Datum ist dann einzeln in der Grafik als Punkt über dem entsprechenden Wert markiert; Liegen mehrere Daten mit demselben Wert vor, werden die Punkte übereinander ausgerichtet.

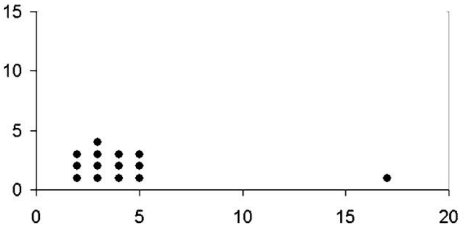
BEISPIELE

Betrachten Sie die folgenden **Daten:**

5, 2, 4, 5 , 3 , 2 , 4, 3, 5, 2, 3, 4, 3, 17 .

Um das Punktdiagramm zu erstellen, erstellen wir eine Skala von 0 bis 20.

Anschließend stellen wir die **Daten** mit Punkten über jedem **Wert dar.** Wir erhalten:



Beachten Sie, dass diese Reihe mit 17 einen **Ausreißer** enthält .

Machine Translated by Google

WEITERLESEN

Liniendiagramm

Häufigkeitsverteilung

Histogramm

Grafische Darstellung

Durbin-Watson-Test

Der Durbin-Watson-Test führt eine **Statistik** ein

Tic *d* , der zum Testen der Autokorrelation von verwendet wird die aus einem linearen Regressionsmodell erhaltenen Residuen. Dies ist ein Problem, das häufig bei der Anwendung eines linearen Modells auf eine **Zeitreihe auftritt,** wenn wir die Unabhängigkeit der auf diese Weise erhaltenen Residuen testen möchten.

GESCHICHTE

Durbin, J. und Watson, GS haben diesen Test 1950 erfunden.

MATHEMATISCHE ASPEKTE

Betrachten Sie den Fall eines **multiplen linearen Regressionsmodells** mit *p* ÿ 1 unabhängigen Variablen. Das Modell ist geschrieben:

$$Yt = ÿ0 + \sum_{j=1}^{pÿ1} ÿjXjt + ÿt , \qquad t = 1,...,T ,$$

Wo

Yt ist die **abhängige Variable,**

Xjt, mit *j = 1,...,p* ÿ 1 sind die unabhängigen Variablen

ÿj, mit *j = 1,...,pÿ1* sind die zu schätzenden Parameter,

ÿt mit *t = 1,...,T* ist eine Nichtbeobachtbare zufälliger Fehlerterm.

In der Matrixform wird das Modell wie folgt geschrieben:

$$\mathbf{Y} = \mathbf{Xÿ} + \mathbf{ÿ} \quad ,$$

Wo

Y ist der Vektor *(n* × 1) der Beobachtungen, die sich auf die abhängige Variable beziehen *(n* Beobachtungen),

ÿ ist ein Vektor *(p* × 1) der zu schätzenden Parameter,

ÿ ist ein Vektor *(n* × 1) der Fehler,

$$\text{undX} = \begin{matrix} ÿ \\ ÿ \\ ÿ \end{matrix} \begin{matrix} 1 & X11 & ... & X1(pÿ1) \\ \vdots & \vdots & & \vdots \\ 1 & Xn1 & ... & Xn(pÿ1) \end{matrix} \begin{matrix} ÿ \\ ÿ \\ ÿ \end{matrix}$$

ÿ ist das **Matrix** (n×p) des Designs, das den unabhängigen Variablen zugeordnet ist.

Die mit der Methode der **kleinsten Quadrate** erhaltenen Residuen sind gegeben durch:

$$\mathbf{e} = \mathbf{Y} ÿ \mathbf{Y} = I ÿ \mathbf{XXX} \overset{ÿ1}{} \mathbf{XY} \quad .$$

Die **Statistik** *d* ist wie folgt definiert:

$$d = \frac{\sum_{t=2}^{T} (et ÿ etÿ1)^2}{\sum_{t=1}^{T} et^2} .$$

Die Statistik *d* testet die Hypothese, dass die Fehler unabhängig sind, gegen die Alternative, dass sie einer Autoregression erster Ordnung folgen Verfahren:

$$ÿt = ÿÿtÿ1 + ut ,$$

wobei |ÿ| < 1 und sie sind unabhängig und normalverteilt mit einem **Mittelwert** von Null und einer **Varianz** von ÿ2. Wir nennen den Term ÿ die **Autokorrelation.** In Bezug auf ÿ wird die **Hypothesenprüfung** wie folgt geschrieben:

$$H0 : ÿ = 0$$

$$H1 : ÿ = 0 .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Konstruktion dieser Statistik bedeutet dies Es kann Werte zwischen 0 und 4 annehmen. Wir haben

Machine Translated by Google

$d = 2$, wenn ÿ = 0. Wir bezeichnen die Stichprobenschätzung der Beobachtung von ÿ mit ÿ. Um die Hypothese H0 zu testen, haben Durbin und Watson die kritischen Werte von d at tabellarisch aufgeführt ein Signifikanzniveau von 5 %; Diese kritischen Werte hängen von der Anzahl der Beobachtungen T und der Anzahl erklärender Variablen $(p$ ÿ 1) ab. Die Tabelle zum Testen der positiven Autokorrelation auf der ÿ-Ebene, der statistische Test d wird mit unteren und oberen kritischen Werten (dL, dU) verglichen und • lehnt H0 ab : $p = 0$, wenn $d < dL$. • lehnt H0 nicht ab , wenn $d > dU$ • Für $dL < d < dU$ ist der Test nicht schlüssig.

Ähnlich zum Testen der negativen Autokorrelation auf der ÿ-Ebene wird die Statistik 4-d mit unteren und oberen kritischen Werten $(dL,$ dU) verglichen und • H0 ablehnen : $p = 0$, wenn H4 ÿ $d <$ dL • H0 nicht ablehnen, wenn 4 ÿ $d >$ dU • Für $dL < 4$ ÿ $d < dU$ ist der Test nicht schlüssig.

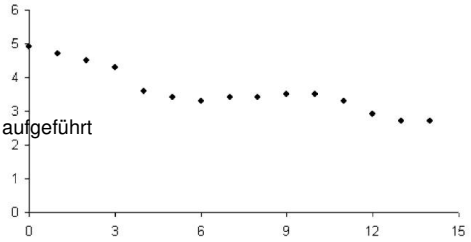
Beachten Sie, dass das Modell eine Konstante enthalten muss Term, da etablierte Durbin-Watson-Tabellen im Allgemeinen für Modelle mit konstantem Term gelten. Die zu erklärende Variable sollte nicht zu den erklärenden Variablen gehören.

BEISPIELE

Wir betrachten ein Beispiel für die Zahl der Straßenunfälle in den Vereinigten Staaten pro Million mit dem Auto zurückgelegter Meilen zwischen 1970 und 1984. Die Daten für diese Jahre sind unten aufgeführt:

4,9, 4,7, 4,5, 4,3 , 3,6 , 3,4, 3,3, 3,4, 3,4, 3,5, 3,5, 3,3, 2,9, 2,7, 2,7.

Die folgende Abbildung veranschaulicht, wie sich die Daten über die Zeit verteilen. Auf der Horizontalen Die Null entspricht dem Jahr 1970 und die 15 dem Jahr 1984.



$$\text{Anzahl der Unfälle} = a + b \cdot t \quad ,$$

mit $t = 0, 1,...,$ 14.

Die Gleichung für die einfache lineare Regressionslinie, die für diese Daten erhalten wurde, ist unten angegeben:

$$\text{Anzahl Unfälle} = 4{,}59 \text{ ÿ } 0{,}141 \cdot t \, .$$

Wir erhalten dann die folgenden Residuen:

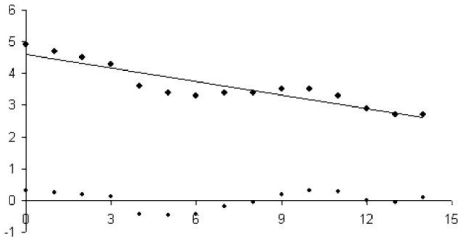
0,31, 0,25, 0,19, 0,13 , ÿ0,43 , ÿ0,49, ÿ0,45 , ÿ0,21, ÿ0,07, 0,17 , 0,32, 0,26, 0,00, ÿ0,06, 0,08.

Der Wert der Statistik d beträgt 0,53. Im Durbin-Watson-Tabelle für eine Ein-Variable Regression finden wir die Werte:

$$dL{,}0{,}05 = 1{,}08$$

$$dU{,}0{,}05 = 1{,}36 \, .$$

Daher haben wir den Fall, dass $0 < d < dL$ *ist,* also lehnen wir die H0- Hypothese zugunsten einer positiven Autokorrelation zwischen den Residuen des Modells ab. Die folgende Abbildung veranschaulicht die positive Korrelation zwischen den Residuen mit der Zeit:



Machine Translated by Google

WEITERLESEN

Autokorrelation

Multiple lineare Regression
Restwert

Einfache lineare Regression
Statistiken

Zeitfolgen

VERWEISE

Bourbonnais, R.: Econométrie, manuel et exercices corrigés, 2. Aufl. Dunod, Paris (1998)

Durbin, J.: Alternative Methode zum D-Test. Biometrie **56,** 1–15 (1969)

Durbin, J., Watson, GS: Testen auf serielle Korrelation in der Regression der kleinsten Quadrate, I. Biometrie **37,** 409–428 (1950)

Durbin, J., Watson, GS: Testen auf serielle Korrelation in der Regression der kleinsten Quadrate, II. Biometrie **38,** 159–177 (1951)

Harvey, AC: Die ökonometrische Analyse von Zeitreihen. Philip Allan, Oxford (Wiley, New York) (1981)

Machine Translated by Google

Ökonometrie

Ökonometrie befasst sich mit der Anwendung statistischer Methoden auf Wirtschaftsdaten . Ökonomen nutzen **Statistiken ,** um ihre Theorie zu testen riesen oder Prognosen zu erstellen.

Da Wirtschaftsdaten nicht experimenteller Natur sind und häufig eine gewisse Zufälligkeit enthalten, verwendet die Ökonometrie eher stochastische **Modelle** als deterministische Modelle.

GESCHICHTE

Die Vorteile der Anwendung von Mathematik und **Statistik** im Bereich der Wirtschaftswissenschaften wurden schnell erkannt. In der zweiten Hälfte des 17. Jahrhunderts veröffentlichte Petty, Sir William (1676) einen wichtigen Artikel, der die methodischen Grundlagen der Ökonometrie vorstellte.

Laut Jaffé,W. (1968) Walras, Léon, Professor an der Universität Lausanne, gilt als Begründer der allgemeinen Gleichgewichtsökonomietheorie, die die theoretische Grundlage für die moderne Wirtschaft liefert

Metriken.

Am 29. Dezember 1930 wurde in Cleveland

Ohio gründete eine Gruppe von Ökonomen, Statistikern und Mathematikern ein Econom

rics Gesellschaft zur Förderung der Erforschung mathematisch-statistischer Theorien

mit den Bereichen der Wirtschaftswissenschaften. Sie erstellten die zweimonatliche Zeitschrift „Econmetrica“. das erstmals im Januar 1933 erschien.

Die Entwicklung simultaner Gleichungsmodelle , die in **Zeitreihen** verwendet werden könnten

für Wirtschaftsprognosen **war** der richtige Zeitpunkt

Die Ökonometrie entwickelte sich zu einem eigenständigen Fachgebiet und ist auch heute noch ein wichtiger Teil der Ökonometrie.

WEITERLESEN

Multiple lineare Regression

Einfache lineare Regression

Zeitfolgen

VERWEISE

Jaffé, W.: Walras, Léon. In: Sills, DL (Hrsg.) Internationale Enzyklopädie der Sozialwissenschaften, Bd. 16. Macmillan und Free Press, New York, S. 447–453 (1968)

Petty, W.: Politische Arithmetik. In: William Petty, The Economic Writings. . . , Vol 1. Kelley, New York, S. 233–313 (1676)

Wonnacott, RJ, Wonnacott, TH: Econo-Kennzahlen. Wiley, New York (1970)

Edgeworth, Francis Y.

Edgeworth, Francis Ysidro wurde 1845 in Edgeworthstown (Irland) geboren und starb in

Machine Translated by Google

1926. Er trug zu verschiedenen Themen bei,
wie Moral, Wirtschaftswissenschaften,
Wahrscheinlichkeit und Statistik. Sein frühes und wichtiges
Arbeiten an Indizes und der Nutzentheorie waren
Später folgte ein statistisches Werk, das jetzt existiert
manchmal auch „bayesianisch“ genannt.
Die Begriffe „Modul“ und „Fluktuation“ gehen auf ihn
zurück.
Er wurde der erste Herausgeber des Economic
Zeitschrift im Jahr 1881.

Hauptwerk von Edgeworth, Francis Ysidro:

1881 Mathematische Physik: Ein Essay über die Anwendung der Mathematik auf die Moralwissenschaften. Kegan Paul, London.

Efron, Bradley

Efron, Bradley wurde im Mai 1938 in St. Paul,
Minnesota, als Sohn von Efron, Esther und geboren
Efron, Miles, jüdisch-russische Einwanderer.
1960 schloss er sein Mathematikstudium ab. In
diesem Jahr kam er nach Stanford, wo er
Er erhielt seinen Ph.D. unter der Leitung
von Miller, Rupert und Solomon, Herb in der
Statistikabteilung. Er hat mehrere übernommen
Positionen an der Universität: Lehrstuhl für Statistik,
stellvertretender Dekan für Naturwissenschaften, Vorsitzender
des Universitätsbeirats und Vorsitzender
des Fakultätssenats. Derzeit ist er Professor für
Statistik und Biostatistik an der Stanford University
Universität.
Er wurde mit zahlreichen Preisen ausgezeichnet,
darunter dem Ford-Preis, dem MacArthur-Preis und
die Wilks-Medaille für seine Forschungsarbeit
der Einsatz von Computeranwendungen in der
Statistik, insbesondere im Hinblick auf Bootstrap und die
Klappmessertechniken. Er ist Mitglied der
Nationale Akademie der Wissenschaften und die
amerikanische Akademie der Künste und Wissenschaften. Er hält
ein Stipendium am IMS und der ASA. Er auch
wurde Präsident des American Statistical
Verein im Jahr 2004.
Der Begriff „computerintensive Statistik“.
Methoden“ stammen von ihm und Diakonis, P. Seine
Forschungsinteressen umfassen ein breites Spektrum
Spektrum statistischer Themen.

Einige Hauptwerke und Artikel von Efron, Bradley:

1993 (mit Tibshirani, RJ) Eine Einführung in den Bootstrap. Chapman & Saal.

1983 Schätzung der Fehlerrate einer Vorhersageregel: Verbesserung der Kreuzvalidierung. Marmelade. Stat. Assoc. 78, 316-331.

1982 Jackknife, der Bootstrap und andere Resampling-Pläne. SIAM, Philadelphia, PA.

WEITERLESEN

- **Bootstrap**
- **Resampling**

Eigenwert

Sei Ab eine **quadratische Matrix** der Ordnung n. Wenn da ist
ein **Vektor x** = 0, der ein Vielfaches *(k* ·**x)** von ergibt
selbst, wenn man ihn mit A multipliziert, dann nennt
man diesen Vektor einen **Eigenvektor** und den
multiplikativen Faktor *k* einen Eigenwert von
die Matrix A.

MATHEMATISCHE ASPEKTE

Sei **A** eine quadratische **Matrix** der Ordnung *n* und **x**
ein **Eigenvektor ungleich Null.** *k* ist ein Eigenwert von
Ein Wenn

$$\mathbf{A} \cdot \mathbf{x} = k \cdot \mathbf{x} \quad .$$

Machine Translated by Google

das heißt, **(A** ÿ *k* · *In)* · **x** = 0, wobei *In* die **Identitätsmatrix** der Ordnung n ist.
Da **x** ungleich Null ist, können wir die möglichen Werte von *k* bestimmen , indem wir die Lösungen der Gleichung |**A** ÿ *k* · *In*| finden = 0.

DOMÄNEN UND EINSCHRÄNKUNGEN

Für eine quadratische **Matrix A** der Ordnung n ist die **Determinante** |**A** ÿ *k* · *In*| ist ein Polynom vom Grad *n* in *k ,* das höchstens *n* nicht unbedingt eindeutige Lösungen hat. Es wird also höchstens *n* Eigenwerte geben; einige davon können Nullen sein.

BEISPIELE

Betrachten Sie die folgende quadratische **Matrix** der Ordnung 3:

$$\mathbf{A} = \begin{matrix} \ddot{y} & 432 & \ddot{y} \\ \ddot{y} & 010 & \ddot{y} \\ \ddot{y} & \ddot{y}220 & \ddot{y} \end{matrix}$$

$$\mathbf{A}\ \ddot{y}\ k \cdot \mathbf{I3} = \begin{matrix} \ddot{y} & 432 & \ddot{y} \\ \ddot{y} & 010 & \ddot{y} \\ \ddot{y} & \ddot{y}220 & \ddot{y} \end{matrix}$$

$$\ddot{y}\ k \cdot \begin{matrix} \ddot{y} & 100 & \ddot{y} \\ \ddot{y} & 010 & \ddot{y} \\ \ddot{y} & 001 & \ddot{y} \end{matrix}$$

$$\mathbf{A}\ \ddot{y}\ k \cdot \mathbf{I3} = \begin{matrix} & 4\ \ddot{y}\ k & & 3\ 2 \\ \ddot{y} & & 0\ 1\ \ddot{y}\ k\ 0 & \\ \ddot{y} & \ddot{y}2 & & 2\ \ddot{y}k \end{matrix} \quad \ddot{y}\ \ddot{y}\ .$$

Die **Determinante** von **A** ÿ *k* · I3 ist gegeben durch:

$$\begin{aligned} |\mathbf{A}\ \ddot{y}\ k \cdot \mathbf{I3}| &= (1\ \ddot{y}\ k) \cdot [(4\ \ddot{y}\ k) \\ &\quad \cdot (\ddot{y}k)\ \ddot{y}\ 2 \cdot (\ddot{y}2)] \\ &= (1\ \ddot{y}\ k) \cdot (k2\ \ddot{y}\ 4 \cdot k + 4) \\ &= (1\ \ddot{y}\ k) \cdot (k\ \ddot{y}\ 2)^{2}\ . \end{aligned}$$

Die Eigenwerte erhält man, indem man die Lösungen der Gleichungen (1ÿk)·(kÿ2)2 = findet 0. Wir finden

$$k1 = 1 \text{ und}$$
$$k2 = k3 = 2\ .$$

Daher sind zwei der drei Eigenwerte verwechselt.

WEITERLESEN

Bestimmend
Eigenvektor
Matrix
Vektor

Eigenvektor

Sei **A** eine quadratische **Matrix** der Ordnung n. Ein „Eigenvektor" von **A** ist ein **Vektor x,** der *n* Komponenten (nicht alle Nullen) enthält , wobei das Matrixprodukt **von Aby x** ein Vielfaches von x ist, zum Beispiel *k* · x.
Wir sagen, dass dem Eigenvektor **x** ein **Eigenwert** k zugeordnet ist .

DOMÄNEN UND EINSCHRÄNKUNGEN

Sei **A** eine quadratische **Matrix** der Ordnung n; dann ist der **Vektor x** (nicht Null) ein Eigenvektor von **A ,** wenn es eine Zahl *k* gibt, so dass

$$\mathbf{A} \cdot \mathbf{x} = k \cdot \mathbf{x}\ .$$

Da **x** ungleich Null ist, bestimmen wir die möglichen Werte von *k ,* indem wir Lösungen für die Gleichung |A ÿ *k* · **In**| finden = 0, wobei In **die Identitätsmatrix** der Ordnung n ist.

Wir finden die entsprechenden Eigenvektoren, indem wir das Gleichungssystem auflösen, das definiert ist durch:

$$(\mathbf{A}\ \ddot{y}\ k \cdot \mathbf{In}) \cdot \mathbf{x} = 0\ .$$

Jedes von Null verschiedene Vielfache eines Eigenvektors kann als Eigenvektor betrachtet werden. Wenn *c*

Machine Translated by Google

ist dann eine Konstante ungleich Null

$$\mathbf{A} \cdot \mathbf{x} = k \cdot \mathbf{x}$$

$$c \cdot (\mathbf{A} \cdot \mathrm{x}) = c \cdot (k \cdot \mathrm{x})$$

$$\mathbf{A} \cdot (c \cdot \mathrm{x}) = k \cdot (c \cdot \mathrm{x}) .$$

Daher ist $c \cdot \mathbf{x}$ auch ein Eigenvektor, wenn c ist nicht Null.

Es ist möglich, einen Eigenvektor, den Vektor **der Norm1** oder den Einheitsvektor auf der Achse von **c·x,** zu nennen, und der Eigenvektor liegt auf der **Fakultätsachse.** Dieser Vektor ist (bis auf ein Vorzeichen) eindeutig.

BEISPIELE

Wir betrachten die folgende quadratische **Matrix** der Ordnung 3:

$$\mathbf{A} = \begin{matrix} \ddot{y} & 432 & \\ & 010 & \\ \ddot{y} & \ddot{y}220 & \ddot{y}\,\ddot{y} . \end{matrix}$$

Die Eigenwerte dieser Matrix A werden erhalten, indem die **Determinante** von **(A** ÿ $k \cdot$ **I3)** gleich Null gemacht wird

$$\mathbf{A} \ddot{y} k \cdot \mathbf{I3} = \begin{matrix} & 432 & \\ \ddot{y} & 010 & \ddot{y} \\ \ddot{y} & \ddot{y}220 & \ddot{y} \end{matrix}$$

$$\ddot{y}\, k \cdot \begin{matrix} & 100 & \\ \ddot{y} & 010 & \ddot{y} \\ & 001 & \ddot{y} \end{matrix}$$

$$\mathbf{A} \ddot{y} k \cdot \mathbf{I3} = \begin{matrix} & 4 \ddot{y} k & 3\ 2 \\ \ddot{y} & 0\ 1 \ddot{y} k\ 0 & \\ \ddot{y} & \ddot{y}2 & 2\ \ddot{y}k \end{matrix} \quad \ddot{y}\,\ddot{y} .$$

Die Determinante von **A** ÿ $k \cdot$ **I3** ist gegeben durch:

$$|\mathbf{A} \ddot{y} k \cdot \mathbf{I3}| = (1 \ddot{y} \mathrm{k}) \cdot [(4 \ddot{y} \mathrm{k}) \cdot (\ddot{y}k) \ddot{y}\ 2 \cdot (\ddot{y}2)]$$

$$= (1 \ddot{y} \mathrm{k}) \cdot \mathrm{k}2 \ddot{y}\ 4 \cdot k + 4$$

$$= (1 \ddot{y} \mathrm{k}) \cdot (k \ddot{y}\ 2)^{\ 2} .$$

Daher sind die Eigenwerte gleich:

$$\mathrm{k}1 = 1 \text{ und}$$

$$\mathrm{k}2 = \mathrm{k}3 = 2 \ .$$

Wir berechnen nun die Eigenvektoren, die dem Eigenwert k1 zugeordnet sind:

$$\mathrm{k}1 = 1 \ ,$$

$$\mathbf{A} \ddot{y} \mathrm{k}1 \cdot \mathbf{I3} = \begin{matrix} & 4 \ddot{y} 1\ 32 & \\ \ddot{y} & 0\ 1 \ddot{y} 1\ 0 & \ddot{y} \\ \ddot{y} & \ddot{y}2 \quad 2 \ddot{y}1 & \ddot{y} \end{matrix}$$

$$= \begin{matrix} & 33\ 2 & \\ \ddot{y} & 00\ 0 & \\ \ddot{y} & \ddot{y}2\ 2\ \ddot{y}1 & \ddot{y}\,\ddot{y} . \end{matrix}$$

Bezeichnet die Komponenten des Eigenvec tor x as

$$\mathbf{x} = \begin{matrix} \ddot{y} & \mathrm{x}1 & \\ & \mathrm{x}2 & \\ \ddot{y} & \mathrm{x}3 & \ddot{y}\,\ddot{y} . \end{matrix}$$

Wir müssen dann das folgende Gleichungssystem auflösen:

$$(\mathbf{A} \ddot{y} \mathrm{k}1 \cdot \mathbf{I3}) \cdot \mathbf{x} = 0$$

$$\begin{matrix} & 33\ 2 & \\ \ddot{y} & 00\ 0 & \\ \ddot{y} & \ddot{y}2\ 2\ \ddot{y}1 & \end{matrix} \quad \begin{matrix} & \mathrm{x}1 & \\ \ddot{y} & \cdot\ \ddot{y}\ \mathrm{x}2 & \ddot{y} \\ \ddot{y} & \ddot{y}\ \mathrm{x}3 & \ddot{y} \end{matrix} = \begin{matrix} & 0 & \\ \ddot{y} & 0 & \\ \ddot{y} & 0 & \ddot{y}\,\ddot{y} . \end{matrix}$$

das ist

$$3x1 + 3x2 + 2x3 = 0$$

$$\ddot{y}2x1 + 2x2\ \ddot{y}\ \mathrm{x}3 = 0 .$$

wobei die zweite Gleichung eliminiert wurde, weil sie trivial ist.

Die Lösung dieses Gleichungssystems in Abhängigkeit von x2 ist gegeben durch:

$$\mathbf{x} = \begin{matrix} \ddot{y} & 7 \cdot \mathrm{x}2 & \\ & \mathrm{x}2 & \\ \ddot{y} & \ddot{y}12 \cdot \mathrm{x}2 & \ddot{y}\,\ddot{y} . \end{matrix}$$

Der Vektor x ist ein Eigenvektor der Matrix A für jeden **Wert von** x2 ungleich Null. Um die Einheit zu finden

Machine Translated by Google

Eigenvektor, wir lassen:

$$1 = \mathbf{x}\quad 2 = (7 \cdot x2)^2 + (x2)^2 + (\ddot{y}12 \cdot x2)^2$$

$$= (49 + 1 + 144) \cdot x2^2 = 194 \cdot x2^2.$$

von wo $x2^2 = \frac{1}{194,}$

also $x2 = = 0{,}07\overline{18.}$

$\ddot{y}194$

Der Einheitseigenvektor der Matrix **A** zum Eigenwert k1 = 1 ist dann gleich:

$$\mathbf{x} = \begin{matrix} \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} 0{,}50 \\ 0{,}07 \\ \ddot{y}0{,}86 \end{matrix} \begin{matrix} \ddot{y} \\ \ddot{y}. \end{matrix}$$

WEITERLESEN

Bestimmend

Eigenwert

Eigenvektor

Faktorielle Achse

Matrix

VERWEISE

Dodge, Y.: Basismathematik für
Wirtschaftswissenschaftler. Springer, Berlin Heidelberg
New York (2002)

Meyer, CD: Matrixanalyse und Angewandte
Lineare Algebra, SIAM (2000)

Epidemiologie

Epidemiologie (das Wort leitet sich von ab
Die griechischen Wörter „epidemos“ und „logos“, was „Lehre von Epidemien“ bedeutet, sind das Studium der Häufigkeit und Verteilung von Krankheiten und die Parameter und die Risikofaktoren, die das definieren Gesundheitszustand einer Bevölkerung, bezogen auf Zeit, Raum und Gruppen von Individuen.

GESCHICHTE

Die Epidemiologie entstand im 17. Jahrhundert Jahrhundert, als Petty, William (1623–1687), Ein englischer Arzt, Ökonom und Wissenschaftler sammelte Informationen über die Bevölkerung, die er in dem Werk „ Political *Arithmetic“ beschrieb.* Graunt, John (1620–1694) schlug 1662 die erste strenge Analyse der Todesursachen vor. **Quetelet, Adolphe** (1796–1874), belgischer Astronom und Mathematiker, gilt als Gründer der modernen Bevölkerungsstatistik, die Mutter Disziplin der Epidemiologie, Statistik, Wirtschaftsmetrik und anderer quantitativer Disziplinen Beschreibung sozialer und biologischer Merkmale des menschlichen Lebens. Mit der Arbeit von Farr, William (1807–1883), Epidemiologe war als von der Statistik getrenntes Fachgebiet anerkannt, da er die Todesursachen untersuchte und die Art und Weise, wie es je nach Alter, Geschlecht, Jahreszeit variierte, Wohnort und Beruf. Der schottische Arzt Lind, James (1716–1794) bewies dass der Verzehr von Zitrusfrüchten Skorbut stoppte, was markierte einen wichtigen Schritt in der Geschichte von die Epidemiologie. Schnee, John (1813–1858), ein Arzt in London, zeigte Theaterepidemien durch Ansteckungen verbreitet, was durch eine In-situ-Studie untermauert wurde, die dies bewies Trinken von durch Müll verschmutztem Themsewasser aus der Kanalisation trägt zur Verbreitung bei von Cholera.

MATHEMATISCHE ASPEKTE

Siehe **Ursache und Wirkung in der Epidemiologie, Quoten und Quotenverhältnis, relatives Risiko, angreifbares Risiko, vermeidbares Risiko, Inzidenz Rate, Prävalenzrate.**

BEISPIELE

Nehmen wir als Beispiel die Sterblichkeitsraten in zwei fiktiven Regionen. Vermuten

Machine Translated by Google

dass wir die jährliche Sterblichkeit in zwei jeweils besiedelten Regionen A und B analysieren wollen mit 10000 Einwohnern. Die Sterblichkeitsdaten sind in der folgenden Tabelle angegeben:

	Region A	Region B
Todesfälle/Gesamtbevölkerung	500/10000	1000/10000
Gesamtsterblichkeit pro Jahr	5 %	10 %
Jährliche Sterblichkeitsrate unter denen < 60 Jahre alt	5 %	5 %
Jährliche Sterblichkeitsrate darunter > 60 Jahre alt	15%	15%
% der Bevölkerung < 60 Jahre alt	100%	50 %

Anhand dieser Daten können wir feststellen, dass die Bewohner der Region B einem Anstieg ausgesetzt sind Sterbewahrscheinlichkeit als die Einwohner von Region A? Mit anderen Worten: Würden wir den Menschen empfehlen, Region A zu verlassen, um in B zu leben? Wenn wir beide Bevölkerungsgruppen nach Alter teilen, erhalten wir Beachten Sie, dass die unterschiedlichen Gesamtsterblichkeitsraten der Regionen A und B dies jedoch nicht zu tun scheinen lassen sich durch die unterschiedliche Altersstruktur erklären von A und B.

Das vorangegangene Beispiel verdeutlicht die Vielzahl von Faktoren, die im Laufe der Zeit berücksichtigt werden sollten Datenanalyse. Die Hauptfaktoren dafür Wir sollten bei einer epidemiologischen Studie Folgendes berücksichtigen: Alter (als absolutes Alter und als die Generation, der das Individuum angehört zu), Geschlecht und ethnische Zugehörigkeit, unter anderem.

WEITERLESEN

- Zurechenbares Risiko
- Vermeidbares Risiko
- Bio-Statistiken
- Ursache und Wirkung in der Epidemiologie
- Inzidenzrate
- Quoten und Quotenverhältnis
- Prävalenzrate
- Relatives Risiko

VERWEISE

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Kleiner Brauner, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Rothmann, JK: Epidemiologie. Eine Einleitung. Oxford University Press (2002)

Fehler

Der Fehler ist die Differenz zwischen den Schätzungen Vergleichswert und der wahre **Wert** (oder Referenz). Wert) der betreffenden Menge.

GESCHICHTE

Die Fehlerverteilungen sind die **Wahrscheinlichkeitsverteilungen ,** die den Fehler beschreiben das bei wiederholten Messungen von auftritt eine gleiche Menge unter den gleichen Bedingungen. Sie wurden in der zweiten Hälfte eingeführt des 18. Jahrhunderts, um zu veranschaulichen, wie Zur Ermittlung kann das **arithmetische Mittel** verwendet werden eine gute Annäherung an den **Referenzwert** der untersuchten Größe.

In einem Brief an den Präsidenten der Royal Society schlug Simpson, T. (1756) vor, dass die Die Wahrscheinlichkeit, bei **einer Beobachtung** einen bestimmten Fehler zu erhalten, kann durch **eine diskrete Wahrscheinlichkeitsverteilung** beschrieben werden . Er schlug die ersten beiden Fehlerverteilungen vor. Andere diskrete Verteilungen wurden vorgeschlagen und untersucht von Lagrange, JL (1776).

Machine Translated by Google

Simpson, T. (1757) schlugen ebenfalls kontinuierliche Fehlerverteilungen vor, ebenso wie **Laplace, PS** (1774 und 1781). Die wichtigste Fehlerverteilung, die **Normalverteilung ,** wurde jedoch **von Gauss,** CF (1809) vorgeschlagen .

MATHEMATISCHE ASPEKTE

Im Bereich der Messtechnik können folgende Fehlerarten unterschieden werden: • Der absolute Fehler, also der absolute Lautenwert der Differenz zwischen den beobachteter **Wert** und der Referenzwert von eine bestimmte Menge; • Der relative Fehler, der das Verhältnis zwischen dem absoluten Fehler und dem Wert darstellt der Menge selbst. Es charakterisiert die Genauigkeit einer physikalischen Messung; • Der zufällige Fehler, der nicht mit dem Messgerät, sondern mit dem Experimentator selbst zusammenhängt; • Der experimentelle Fehler, also der Fehler aufgrund unkontrollierter **Variablen;** • Der Zufallsfehler, also der zufällige Fehler, der sich aus einer **Kombination** von Fehlern aufgrund des Instruments und/oder des Geräts ergibt Benutzer. Die statistischen Eigenschaften zufälliger Fehler und ihre **Schätzung** werden untersucht Verwendung von **Wahrscheinlichkeiten;** • Der Restfehler, der der Differenz zwischen den geschätzten entspricht **Wert** und dem beobachteten Wert. Der Begriff Für diesen Fehler wird manchmal **„Residuum“** verwendet. • Der systematische Fehler, der Fehler, der auf konsistente Ursachen zurückzuführen ist (zum Beispiel die unsachgemäße Kalibrierung eines Messgeräts) und immer in der gleichen Richtung auftritt. Die **Voreingenommenheit,** die in verwendet wird Statistiken sind hierfür ein besonderes Beispiel. • Der gerundete Fehler, also der Fehler, der bei einem numerischen **Wert** entsteht wird durch einen gekürzten Wert in der Nähe ersetzt. Bei Berechnungen mit Bei Zahlen mit *n* Dezimalstellen wird der Wert des gerundeten Fehlers ermittelt zwischen

$$\ddot{y}1 \perp 10\ddot{y}n \text{ und } 2 \quad {}_{\overline{T}2} \cdot 10\ddot{y}n .$$

Bei langen Rechenvorgängen können sich Rundungsfehler häufen und zu sehr ungenauen Ergebnissen führen. Diese Art von Fehler tritt häufig bei der Verwendung eines Computers auf.

Jeder Fehler, der bei einem **statistischen** Problem auftritt, ist eine Kombination der verschiedenen oben aufgeführten Fehlertypen.

In **der Statistik** und insbesondere beim **Testen von Hypothesen** gibt es außerdem: • **Fehler vom Typ I,** entsprechend dem Fehler, der begangen wird, wenn die **Nullhypothese** abgelehnt wird, wenn sie wahr ist; • **Fehler vom Typ II,** der Fehler, der begangen wird, wenn die Nullhypothese nicht abgelehnt wird, wenn sie falsch ist.

WEITERLESEN

Analyse von Residuen

Voreingenommenheit

Einschätzung

Hypothesentest

Restwert

Fehler vom Typ I

Fehler vom Typ II

VERWEISE

Gauss, CF: Theoria Motus Corporum Coelestium. Werke, 7 (1809)

Lagrange, JL: Erinnerung an die Verwendung der Methode zur Untersuchung der Umgebung zwischen Ergebnissen zusätzlicher Beobachtungen; Untersuchen Sie daher die Vorteile dieser Methode zur Wahrscheinlichkeitsberechnung. Und es gibt verschiedene Probleme im Zusammenhang damit.

E

Machine Translated by Google

ifs à cette matière. Sonstiges Taurinensia 5, 167–232 (1776)

Laplace, PS de: Erinnerung an die wahrscheinliche Ursache der Ereignisse. Mem. Acad. Roy. Wissenschaft. (vorgetragen von verschiedenen Wissenschaftlern) 6, 621–656 (1774) (oder Laplace, PS de (1891) Œuvres complètes, Bd. 8. Gauthier-Villars, Paris, S. 27–65)

Laplace, PS de: Mémoire sur les probabil ités.Mem. Acad. Roy. Wissenschaft. Paris, 227–332 (1781) (oder Laplace, PS de (1891) Œuvres vollständig, vol. 9. Gauthier-Villars, Paris, S. 385–485.)

Simpson, T.: Ein Brief an den rechten Ehrenwerten George Earl of Macclesfield, Präsident von Die Royal Society nutzte den Vorteil, den Mittelwert einer Reihe von Beobachtungen zu ermitteln in der praktischen Astronomie. Philos.Trans.Roy. Soc. London. **49,** 82–93 (1755)

Simpson, T.: Verschiedene Traktate zu einigen Kuriose und sehr interessante Themen in Mechanik, Physikalischer Astronomie und Spekulative Mathematik. Nourse, London (1757)

Einschätzung

Schätzung ist das Verfahren, das verwendet wird Bestimmen Sie den **Wert** eines bestimmten **Parameters, der** einer **Population zugeordnet ist.** Um den Parameter abzuschätzen, wird eine **Stichprobe** entnommen die Bevölkerung und der Wert des **Schätzers** für den unbekannten Parameter berechnet. Die Schätzung ist in zwei große **Kategorien** unterteilt : **Punktschätzung** und **Intervallschätzung.**

GESCHICHTE

Das Konzept der Schätzung geht auf zurück die ersten Arbeiten zur mathematischen **Statistik,** insbesondere von **Bernoulli, Jacques** (1713), **Laplace, PS** (1774) **und Bernoulli, Daniel** (1778).

Der größte Fortschritt in der Schätztheorie nach der Einführung des **Geringsten**

Die Methode **der Quadrate** war wahrscheinlich die Formel der **Momentenmethode von Pearson, K.** (1894, 1898). Allerdings sind die Grundlagen von

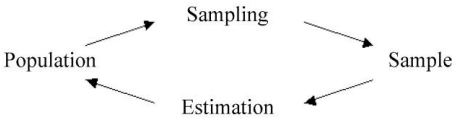
Die Schätztheorie geht auf **Fisher, RA zurück.** In seinem ersten Werk von 1912 führte er die **Maximum-Likelihood** -Methode ein.

Im Jahr 1922 verfasste er eine grundlegende Arbeit klar beschrieben, was Schätzung wirklich ist zum ersten Mal.

Fisher, RA (1925) führte ebenfalls eine Menge ein von Definitionen, die zur Beschreibung übernommen wurden **Schätzer.** Begriffe wie „voreingenommene“, „effiziente“ und „ausreichende“ Schätzer wurden eingeführt von ihm in seiner Schätztheorie abgeleitet.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die folgende Grafik zeigt den Zusammenhang zwischen Probenahme und Schätzung. „Probenahme“ ist der Vorgang der Entnahme einer Probe aus einer Population, während „Schätzung“ die ist umgekehrter Prozess: von der Stichprobe zur Grundgesamtheit.



WEITERLESEN

- **Konfidenzintervall**
- **Schätzer**
- **L1 -Schätzung**
- **Punktschätzung**
- **Robuste Schätzung**
- **Probe**
- **Probenahme**

Machine Translated by Google

VERWEISE

Armatte, M.: La construction des notions d'estimation et de vraisemblance chez Ronald A. Fisher. In: Mairesse, J. (Hrsg.) Schätzung und Sondagen. Economica, Paris (1988)

Bernoulli, D.: Die wahrscheinlichste Wahl zwischen mehreren widersprüchlichen Beobachtungen, Übersetzung und Kommentare (1778). In: Pearson, ES und Kendall, MG (Hrsg.) Studium der Geschichte der Statistik und Wahrscheinlichkeit. Griffin, London (1970)

Bernoulli, J.: ArsConjectandi, Opus Posthu mum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Basel (1713)

Fisher, RA: Über ein absolutes Kriterium für passende Frequenzkurven. Durcheinander. Mathematik. **41,** 155–160 (1912)

Fisher, RA: Über die mathematischen Grundlagen der theoretischen Statistik. Philos. Trans. Roy. Soc. London. Ser. A **222,** 309–368 (1922)

Fisher, RA: Theorie der statistischen Schätzung. Proz. Camb. Philos. Soc. **22,** 700–725 (1925)

Gauss, CF: Theoria Motus Corporum Coelestium. Werke, 7 (1809)

Laplace, PS de: Erinnerung an die wahrscheinliche Ursache der Ereignisse. Mem. Acad. Roy. Wissenschaft. (vorgetragen von verschiedenen Wissenschaftlern) 6, 621–656 (1774) (oder Laplace, PS de (1891) Œuvres complètes, Bd. 8. Gauthier-Villars, Paris, S. 27–65)

Laplace, PS de: Théorie analytique des Wahrscheinlichkeiten. Courcier, Paris (1812)

Legendre, AM: Neue Methoden für Die Bestimmung der Umlaufbahnen der Kometen. Courcier, Paris (1805)

Pearson, K.: Beiträge zur mathematischen Evolutionstheorie. I. In: Karl Pearsons frühe statistische Arbeiten. Cambridge University Press, Cambridge, S. 1–40 (1948). Erstveröffentlichung 1894 as: Zur Zerlegung asymmetrischer Frequenzkurven. Philos. Trans. Roy. Soc. London. Ser. A **185,** 71–110

Pearson, K., Filon, LNG: Mathematisch Beiträge zur Evolutionstheorie. IV: Zu den wahrscheinlichen Häufigkeitsfehlern Konstanten und über den Einfluss des Zufalls Auswahl auf Variation und Korrelation.Phi los. Trans. Roy. Soc. London. Ser. A **191,** 229–311 (1898)

Schätzer

Jede statistische Funktion einer **Stichprobe** verwendet Die Schätzung eines unbekannten **Parameters** der **Bevölkerung** wird als Schätzer bezeichnet.

Jeder von diesem Schätzer erhaltene **Wert** ist ein **Schätzung** des **Parameters** der **Bevölkerung.**

GESCHICHTE

Siehe **Schätzung.**

MATHEMATISCHE ASPEKTE

Betrachten Sie ÿ, einen unbekannten definierten **Parameter** für eine **Population** und (X1, *X2,...,Xn),* eine **Stichprobe** aus dieser Population. Eine statistische Funktion (oder einfach eine **Statistik)** davon Die Probe *g(X1,...,Xn)* wird zur Schätzung von ÿ verwendet. Um die **Statistik** *g(X1,...,Xn) zu unterscheiden,* Das ist eine **Zufallsvariable** aus dem **Wert** dass es im Einzelfall dauert, ist üblich Man nennt die Statistik *g(X1,...,Xn)* den Schätzer, und der **Wert ,** den sie in einem bestimmten Fall annimmt, wird als **Schätzung bezeichnet.**

Machine Translated by Google

Ein Schätzer sollte nicht mit verwechselt werden
eine Schätzmethode .

Ein **Parameter** einer **Population** ist normalerweise
wird mit einem griechischen Buchstaben bezeichnet, und sein **Schätzer**
wird mit demselben Buchstaben bezeichnet, jedoch mit einem Zirkumflex
Akzent (ˆ) oder ein Hut, um es vom zu unterscheiden
Parameter der Bevölkerung. Manchmal die
Stattdessen wird der entsprechende römische Buchstabe verwendet.

DOMÄNEN UND EINSCHRÄNKUNGEN

Bei der Berechnung eines Schätzers, wie z
Im arithmetischen Mittel wird erwartet, dass sich
der **Wert** dieses Schätzers dem Wert von annähert
Der wahre Parameter (in diesem Fall der Mittelwert
der Bevölkerung), wenn **die Stichprobengröße** zunimmt
es.

Es ist daher klar, dass ein guter Schätzer erforderlich ist
nahe am wahren **Parameter liegen.** Je näher
Je besser sich die Schätzer um die wahren Parameter
gruppieren, desto besser sind die Schätzer.
In diesem Sinne sind gute Schätzer stark
im Zusammenhang mit statistischen **Streuungsmaßen**
wie zum Beispiel die **Varianz.**

Im Allgemeinen sollten Schätzer über bestimmte Werte verfügen
Eigenschaften, wie sie im Folgenden beschrieben werden.

1. *Schätzer ohne Voreingenommenheit:*

   Ein Schätzer ÿˆ eines unbekannten **Parameters**
   eter ÿ heißt unbefangen, wenn es so ist

   Der erwartete Wert ist gleich ÿ,

   $$E[\ddot{y}\hat{}] = \ddot{y} .$$

   Der Mittelwert x¯ und die **Varianz** S2 einer
   Stichprobe sind unverzerrte Schätzer des
   Mittelwerts ÿ bzw. der Varianz
   ÿ2 der **Bevölkerung,** mit

   $$x^{-} = \frac{1}{N} \cdot \sum_{i=1}^{N} xi \text{ und}$$

   $$S2 = \frac{1}{n \ddot{y} 1} \cdot \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^{2} .$$

2. *Effizienter Schätzer:*

   Man spricht von einem Schätzer ÿˆ eines **Parameters** ÿ
   effizient sein, wenn die **Varianz** von ÿˆ kleiner ist
   als die Varianz jedes anderen Schätzers
   von ÿ.

   Daher für zwei Schätzer ohne Voreingenommenheit
   von ÿ wird eins effizienter sein als das
   anders, wenn seine **Varianz** kleiner ist.

3. *Konsistenter Schätzer:*

   Man spricht von einem Schätzer ÿˆ eines **Parameters** ÿ
   konsistent sein, wenn die **Wahrscheinlichkeit ,**
   dass es von ÿ abweicht, mit der **Stichprobengröße** abnimmt
   erhöht sich.

   Daher ist ein Schätzer konsistent, wenn

   $$\lim_{n\ddot{y}\ddot{y}} P(|\ddot{y}\hat{} \ddot{y} \ddot{y}| < \ddot{y})\hat{} = 1 \quad ,$$

   wie klein die Zahl ÿ >ˆ 0 auch sein mag.

BEISPIELE

Betrachten Sie (x1, *x2,...,xn),* eine gezogene **Stichprobe**
aus einer **Grundgesamtheit** mit einem **Mittelwert** von ÿ und
eine **Varianz** von ÿ2.

Der Schätzer x¯, der aus der **Stichprobe** berechnet
und zur Schätzung des **Parameters ÿof** verwendet wird
Die **Bevölkerung** ist ein Schätzer ohne Voreingenommenheit.
Dann:

$$\begin{aligned} E[\bar{x}] &= E\left[\frac{1}{N}\sum_{i=1}^{N} xi\right] \\ &= \frac{1}{N} E\left[\sum_{i=1}^{N} xi\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E[xi] \\ &= \frac{1}{N}\sum_{i=1}^{N} \ddot{y} \\ &= \frac{1}{N} n \cdot \ddot{y} \\ &= \ddot{y} . \end{aligned}$$

Machine Translated by Google

Um die **Varianz** ÿ2 der **Population abzuschätzen,** ist es verlockend, den aus der **Stichprobe** (x1, *x2,...,xn)* berechneten Schätzer S2 zu verwenden , der hier definiert ist durch:

$$S2 = \frac{1}{n}\sum_{i=1}^{N} 2\,(xi \ddot{y}\ \bar{x})\ .$$

Dieser Schätzer ist ein voreingenommener Schätzer ÿ2, da:

$$\sum_{i=1}^{N} (xi \ddot{y}\ \bar{x})^2 = \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ n(\bar{x} \ddot{y}\ \ddot{y})2$$

auf die folgende Weise:

$$\sum_{i=1}^{N} (xi \ddot{y}\ \bar{x})^2$$

$$= \sum_{i=1}^{N} [(xi \ddot{y}\ \ddot{y}) \ddot{y}\ (\bar{x} \ddot{y}\ \ddot{y})]^2$$

$$= \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ 2(xi \ddot{y}\ \ddot{y})(\bar{x} \ddot{y}\ \ddot{y}) + (\bar{x} \ddot{y}\ \ddot{y})2$$

$$= \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ 2(\bar{x} \ddot{y}\ \ddot{y}) \cdot \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y}) + n(\bar{x} \ddot{y}\ \ddot{y})2$$

$$= \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ 2(\bar{x} \ddot{y}\ \ddot{y})n \sum_{i=1}^{N} xi\ddot{y}n \cdot \ddot{y} + n(\bar{x} \ddot{y}\ \ddot{y})2$$

$$= \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ 2(\bar{x} \ddot{y}\ \ddot{y})(n \cdot \bar{x} \ddot{y}\ n \cdot \ddot{y}) + n(\bar{x} \ddot{y}\ \ddot{y})2$$

$$= \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ 2n(\bar{x} \ddot{y}\ \ddot{y})(\bar{x} \ddot{y}\ \ddot{y}) + n(\bar{x} \ddot{y}\ \ddot{y})2$$

$$= \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ 2n(\bar{x} \ddot{y}\ \ddot{y})2 + n(\bar{x} \ddot{y}\ \ddot{y})2$$

$$= \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ n(\bar{x} \ddot{y}\ \ddot{y})2\ .$$

Somit:

$$E\,S2 = E\ \frac{1}{N}\sum_{i=1}^{N} (xi \ddot{y}\ \bar{x})^2$$

$$= E\ \frac{1}{N}\sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y}\ (\bar{x} \ddot{y}\ \ddot{y})2$$

$$= \frac{1}{N} E \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})2 \ddot{y} E\,(\bar{x} \ddot{y} \ddot{y})2$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\,(xi \ddot{y}\ \ddot{y})2 \ddot{y} E\,(\bar{x} \ddot{y}\ \ddot{y})2$$

$$= \mathrm{Var}(xi) \ddot{y}\ \mathrm{Var}(\bar{x})$$

$$= \ddot{y}2 \ddot{y}\ \frac{\ddot{y}2}{N}$$

$$= \frac{n \ddot{y}\ 1}{N} \ddot{y}2 = \ddot{y}2\ .$$

Um einen unverzerrten Schätzer für die **Varianz** ÿ2 zu erhalten, vermeiden wir daher die Verwendung von S2 und verwenden stattdessen denselben Wert $\frac{n\ddot{y}1}{N}$ dividiert durch:

$$S^2 = \frac{N}{S2\ n \ddot{y}\ 1}$$

$$= \frac{N}{n \ddot{y}\ 1} \cdot \frac{1}{N}\sum_{i=1}^{N} (xi \ddot{y}\ \bar{x})^2$$

$$= \frac{1}{n \ddot{y}\ 1}\sum_{i=1}^{N} (xi \ddot{y}\ \bar{x})^2\ .$$

WEITERLESEN

Voreingenommenheit

Einschätzung

Probe

Probenahme

VERWEISE

Hogg, RV, Craig, AT: Einführung in die Mathematische Statistik, 2. Aufl. Macmillan, New York (1965)

E

Machine Translated by Google

Lehmann, EL: Theory of Point Estimation, 2. Aufl.
Wiley, New York (1983)

Fall

In einem **Zufallsexperiment** ist ein Ereignis eine Teilmenge der Menge des **Stichprobenraums.** Mit anderen Worten: Ein Ereignis ist eine Reihe möglicher Ergebnisse in einem Zufallsexperiment.

MATHEMATISCHE ASPEKTE

In Betracht ziehen , der **Probenraum** eines **Zufallsexperiments.** Jede Teilmenge von heißt an Fall.

Verschiedene Arten von Ereignissen

• *Unmögliche Ereignisse*

Ein unmögliches Ereignis entspricht einer leeren Menge. Die **Wahrscheinlichkeit** eines unmöglichen Ereignisses ist gleich 0:

$$\text{wenn } B = \ddot{y} \text{ , } P(B) = 0 \text{ .}$$

• *Sicheres Ereignis*

Ein sicheres Ereignis ist ein Ereignis, das dem **Beispielraum entspricht.** Die **Wahrscheinlichkeit** eines sicheren Ereignisses ist gleich 1:

$$P() = 1 \quad .$$

• *Einfaches Ereignis*

Ein einfaches Ereignis ist ein Ereignis, das nur ein Ergebnis enthält.

Operationen auf Ereignissen

• *Negation eines Ereignisses*

Betrachten Sie das EreignisE. Die Negation dieses Ereignisses ist ein anderes Ereignis, das sogenannte Gegenereignis von E, das realisiert wird, wenn *E* nicht realisiert wird und umgekehrt. Das entgegengesetzte Ereignis *von E* wird durch die Menge E¯, das Komplement von E, bezeichnet:

$$\overline{E} = \backslash E = \{\ddot{y}\ \ddot{y}\ ;\ \ddot{y}\ /\ddot{y}\ \mathrm{E}\}\ .$$

• *Schnittmenge von Ereignissen*

Betrachten Sie die beiden Ereignisse *E* und F. Die Der Schnittpunkt zwischen diesen beiden Ereignissen ist ein weiteres Ereignis, das realisiert wird, wenn beides geschieht Ereignisse werden gleichzeitig realisiert. Es wird als „*E* andF“ -Ereignis bezeichnet und mit *E* ÿ F bezeichnet.

$$E \ddot{y} F = \{\ddot{y}\ \ddot{y}\ ;\ \ddot{y}\ \ddot{y}\ E \text{ und } \ddot{y}\ \ddot{y}\ \mathrm{F}\}\ .$$

Allgemeiner gesagt: Wenn E1, *E2,...,Er r*-Ereignisse sind , dann

$$\bigcap_{i=1}^{R} Ei = \{\ddot{y}\ \ddot{y}\ ;\ \ddot{y}\ \ddot{y}\ Ei,\ \ddot{y}i = 1,...,n\}$$

ist definiert als der Schnittpunkt der *r* Ereignisse, die realisiert werden, wenn alle *r* Ereignisse gleichzeitig realisiert werden. •

Vereinigung von Ereignissen Betrachten Sie die beiden gleichen Ereignisse *E* und Die Vereinigung dieser beiden Ereignisse ist eine andere Es handelt sich um ein Ereignis, das eintritt, wenn mindestens eines der beiden Ereignisse *E* oder *F* eintritt. Es wird manchmal als „*E*- oder F“-Ereignis bezeichnet und mit *E* ÿ F bezeichnet :

$$E \ddot{y} F = \{\ddot{y}\ \ddot{y}\ ;\ \ddot{y}\ \ddot{y}\ E \text{ oder } \ddot{y}\ \ddot{y}\ \mathrm{F}\}\ .$$

Eigenschaften von Operationen auf den Ereignissen

Betrachten Sie die drei Ereignisse E, *F* und G. Der Schnittpunkt und die Vereinigung dieser Ereignisse sind

Überprüfen Sie die folgenden Eigenschaften: • *Kommutativität:*

$$E \ddot{y} F = F \ddot{y} E \quad ,$$

$$E \ddot{y} F = F \ddot{y} E \quad .$$

• *Assoziativität:*

$$E \ddot{y} (F \ddot{y} \mathrm{G}) = (E \ddot{y} \mathrm{F}) \ddot{y} GE \ddot{y} (F \ddot{y} \quad ,$$

$$\mathrm{G}) = (E \ddot{y} \mathrm{F}) \ddot{y} G\ .$$

Machine Translated by Google

• *Idempotenz:*

$$E \ddot{y} E = E \quad ,$$
$$E \ddot{y} E = E \quad .$$

Die Distributivität des Schnittpunktes bezüglich der Vereinigung ist:

$$E \ddot{y} (F \ddot{y} G) = (E \ddot{y} F) \ddot{y} (E \ddot{y} G) .$$

Die Distributivität der Vereinigung in Bezug auf den Schnittpunkt ist:

$$E \ddot{y} (F \ddot{y} G) = (E \ddot{y} F) \ddot{y} (E \ddot{y} G) .$$

Die wichtigsten Negationsbeziehungen sind:

$$\overline{E \ddot{y} F} = \bar{E} \ddot{y} \bar{F} \quad ,$$
$$\overline{E \ddot{y} F} = \bar{E} \ddot{y} \bar{F} \quad .$$

BEISPIELE

Stellen Sie sich ein **Zufallsexperiment** vor , bei dem gleichzeitig ein gelber und ein blauer Würfel geworfen werden.

Der **Probenraum** dieses **Experiments** wird aus der Menge der Paare möglicher Ergebnisse der beiden Würfel gebildet:

$$= \{(1, 1),(1, 2),(1, 3),...,(2, 1) ,$$
$$(2, 2),...,(6, 5),(6, 6)\}.$$

wobei zum Beispiel:(1, 2) = „1 auf dem gelben Würfel und „2“ auf dem blauen Würfel. Die Zahl

Die Anzahl der Paare (oder einfachen Ereignisse), die gebildet werden können, ergibt sich aus der Anzahl der **Anordnungen** mit Wiederholung von zwei Objekten unter sechs, also 36. Für diesen **Beispielraum** können beispielsweise die folgenden Ereignisse beschrieben werden: 1. Die Punktesumme beträgt sechs:

$$A = \{(1, 5),(2, 4),(3, 3),(4, 2),(5, 1)\}.$$

2. Die Punktesumme ist gerade:

$$B = \{(1, 1),(1, 3),(1, 5),$$
$$(2, 2),(2, 4),...,(6, 4),(6, 6)\}.$$

3. Die Punktesumme beträgt weniger als 5:

$$C = \{(1, 1),(1, 2),(1, 3),$$
$$(2, 1),(2, 2),(3, 1)\}.$$

4. Die Punktesumme ist ungerade:

$$D = \{(1, 2),(1, 4),(1, 6), (2, 1),$$
$$(2, 3),...,(6, 5)\} = \bar{B} .$$

5. Die Punktesumme ist gerade und kleiner als 5:

$$E = \{(1, 1),(1, 3),(2, 2),(3, 1)\}$$
$$= B \ddot{y} C .$$

6. Der blaue Würfel landet auf „2“:

$$F = \{(1, 2),(2, 2),(3, 2), (4, 2),$$
$$(5, 2),(6, 2)\}.$$

7. Der blaue Würfel landet auf „2“ oder der Gesamtpunktzahl ist kleiner als 5:

$$G = \{(1, 2),(2, 2),(3, 2),(4, 2),(5, 2),$$
$$(6, 2),(1, 1),(1, 3),(2, 1),(3, 1)\}$$
$$= F \ddot{y} C .$$

8. *A* und *D* sind nicht kompatibel.

9. *A* impliziert B.

WEITERLESEN

Kompatibilität

Bedingte Wahrscheinlichkeit

Unabhängigkeit

Wahrscheinlichkeit

Zufälliges Experiment

Probenraum

Machine Translated by Google

Erwarteter Wert

Der erwartete Wert einer **Zufallsvariablen** ist der gewichtete **Mittelwert** der **Werte ,** die der Zufallsvariable kann das Gewicht annehmen ist die **Wahrscheinlichkeit ,** dass ein bestimmter Wert vorliegt von der Zufallsvariablen übernommen.

GESCHICHTE

Das mathematische Prinzip des Erwartungswertes erschien erstmals in dem Werk mit dem Titel *De Ratiociniis in Aleae Ludo,* veröffentlicht in 1657 vom niederländischen Wissenschaftler Huygens, Chris tiaan (1629–1695). Seine Gedanken zu diesem Thema scheinen die späteren Arbeiten von Pascal und Fermat zur **Wahrscheinlichkeit stark beeinflusst zu haben.**

MATHEMATISCHE ASPEKTE

Je nachdem, ob die **Zufallsvariable** diskret oder stetig ist, sprechen wir von entweder der Erwartungswert einer diskreten Zufallsvariablen oder der Erwartungswert einer kontinuierlichen Zufallsvariablen.

Betrachten Sie eine diskrete **Zufallsvariable** *X* das eine **Wahrscheinlichkeitsfunktion** p(X) hat . Der erwarteter Wert von X, bezeichnet mit *E[X]* oder ÿ, ist definiert durch

$$ÿ = E[X] = \sum_{i=1}^{N} xip(xi)$$

wenn *X n* Werte annehmen kann .

Wenn die **Zufallsvariable** *X* stetig ist, ist die Erwartungswert wird

$$ÿ = E[X] = \int_{D} xf(x)dx ,$$

wenn *X* **Werte** über das **Intervall** D annimmt , wobei *f(x)* ist die **Dichtefunktion** von X.

Eigenschaften des Erwartungswerts

Betrachten Sie die beiden Konstanten *a* und b sowie die **Zufallsvariable** X. Dann gilt:

1. *E[aX* + b] = aE[X] + b;
2. E[X + Y] = *E[X]* + *E[Y];*
3. E[X ÿ Y] = *E [X]* ÿ *E* [Y];
4. Wenn *X* und *Y* unabhängig sind, dann:

$$E[X \cdot Y] = E[X] \cdot E[Y] .$$

BEISPIELE

Wir werden zwei Beispiele betrachten, eines für eine diskrete **Zufallsvariable, das** andere für eine kontinuierliche Zufallsvariable.

Stellen Sie sich ein Glücksspiel vor, bei dem es um einen Würfel geht geworfen und der Punktestand notiert. Angenommen, Sie Gewinnen Sie einen Euro, wenn er auf einer geraden Zahl, nämlich zwei, landet Euro, wenn er auf einer „1“ oder einer „3“ landet, und verlieren drei Euro, wenn es auf einer 5 landet.

Die hier betrachtete **Zufallsvariable** *X* ist die Anzahl der Euro, die gewonnen oder verloren werden. Der

Die folgende Tabelle stellt die verschiedenen **Werte** von *X* und ihre jeweiligen Wahrscheinlichkeiten dar:

X	ÿ3 1	2
P (X)	1/6 3/6 2/6	

Der erwartete Wert ist daher gleich:

$$E[X] = \sum_{i=1}^{3} xip(xi)$$
$$= ÿ3 \cdot \frac{1}{6} + 1 \cdot \frac{3}{6} + 2 \cdot \frac{2}{6}$$
$$= \frac{2}{3} .$$

Mit anderen Worten: Der Spieler gewinnt im Durchschnitt $\frac{2}{3}$ von einem Euro pro Wurf.

Betrachten Sie eine kontinuierliche **Zufallsvariable** *X* mit folgender **Dichtefunktion:**

$$f(x) = \begin{cases} 1 \text{ für } 0 < x < 1 \\ 0 \text{ anderswo.} \end{cases}$$

Machine Translated by Google

Der erwartete Wert dieser **Zufallsvariablen** X ist gleich:

$$E[X] = \int_0^1 x \cdot 1 \, dx = \frac{x2}{2}\Big|_0^1 = \frac{1}{2}.$$

WEITERLESEN

- **Dichtefunktion**
- **Wahrscheinlichkeitsfunktion**
- **Gewichtetes arithmetisches Mittel**

Experiment

Ein Experiment ist ein Vorgang, der unter kontrollierten Bedingungen durchgeführt wird, um dies zu verhindern einen bisher unbekannten Effekt abdecken, eine **Hypothese** testen oder aufstellen oder ein bekanntes Gesetz beweisen.

Das Ziel des Experimentierens besteht darin, den Zusammenhang zwischen den kontrollierbaren Bedingungen zu klären und das Ergebnis des Experiments.
Die experimentelle Analyse erfolgt an **Beobachtungen, die** nicht nur durch kontrollierbare Bedingungen, sondern auch durch unkontrollierte Bedingungen und Messfehler beeinflusst werden .

DOMÄNEN UND EINSCHRÄNKUNGEN

Wenn die möglichen Ergebnisse des Experiments beschrieben werden können und es möglich ist, jedem möglichen Ergebnis eine Realisierungswahrscheinlichkeit zuzuordnen, wird das Experiment als **Zufallsexperiment bezeichnet.** Die Menge aller möglichen **Ergebnisse** eines Experiments wird als **Probenraum bezeichnet.**
Bei einem **faktoriellen Experiment** organisiert der Experimentator ein Experiment mit zwei oder mehr **Faktoren.** Die Faktoren sind die Kontrolle stabile Bedingungen des Experiments. Erfahrung Geistesfehler kommen von unkontrollierten Zuständen tionen und aus **Messfehlern** resultieren in jeder Art von Experiment vorhanden.
Abhängig vom Ziel des Experimentators oder den möglichen verfügbaren Ressourcen wird ein Versuchsplan erstellt.

BEISPIELE

Siehe **Zufallsexperiment** und **Fakultätsexperiment.**

WEITERLESEN

- **Versuchsplanung**
- **Faktorielles Experiment**
- **Hypothesentest**
- **Zufälliges Experiment**

VERWEISE

Box, GEP, Hunter, WG, Hunter, JS: Statistiken für Experimentatoren. Eine Einführung in Design, Datenanalyse und Modellbildung. Wiley, New York (1978)

Rüegg, A.: Probabilités et statistique. Press es Polytechniques Romandes, Lausanne, Schweiz (1985)

Experimentelle Einheit

Eine Versuchseinheit ist der kleinste Teil des Versuchsmaterials, auf das wir uns beziehen **Behandlung.**

Der Versuchsplan legt die Anzahl der Versuchseinheiten fest, die den verschiedenen Behandlungen unterzogen werden.

WEITERLESEN

- **Versuchsplanung**
- **Experiment**
- **Faktor**
- **Behandlung**

Machine Translated by Google

Explorativ Datenanalyse

Die explorative Datenanalyse ist ein Ansatz zur Datenanalyse, bei der die Merkmale und Eigenschaften der Daten überprüft werden ein „offener Geist“; mit anderen Worten, ohne der Versuch, ein bestimmtes Modell darauf anzuwenden die Daten. Es wird oft beim ersten Kontakt verwendet mit den Daten, bevor irgendwelche Modelle erstellt wurden für die strukturelle oder stochastische Kom gewählt Komponenten, und es wird auch verwendet, um nach Abweichungen von gängigen Modellen zu suchen.

GESCHICHTE

Bei der explorativen Datenanalyse handelt es sich um eine Reihe von Techniken, die hauptsächlich entwickelt wurden **von Tukey, John Wilders** seit 1970. Die Philosophie hinter diesem Ansatz besteht darin, zu untersuchen die Daten, bevor ein spezifisches Wahrscheinlichkeitsmodell angewendet wird. Laut Tukey, JW, Die explorative Datenanalyse ähnelt der Detektivarbeit. Bei der explorativen Datenanalyse sind dies Hinweise können numerisch sein und (sehr oft) grafisch. Tatsächlich führte Tukey mehrere neue semigrafische Datendarstellungen ein Tools zur Unterstützung der explorativen Datenanalyse, einschließlich des „Box-and-Whisker-Plots“ (auch bekannt als **Boxplot) im** Jahr 1972 und der **Stamm und Blattdiagramm** im Jahr 1977. Dieses Diagramm ist ähnlich dem **Histogramm,** das aus stammt das achtzehnte Jahrhundert.

MATHEMATISCHE ASPEKTE

Eine gute Möglichkeit, die wesentlichen Informationen in einer Datenanalyse mit explorativer Datenanalyse zusammenzufassen, bietet die *fünfstellige Zusammenfassung.* die in Form einer Tabelle dargestellt wird als folgt:

Median	
Erstes Quartil	Drittes Quartil
Minimum	Maximal

Sei *n* die Anzahl der Beobachtungen; Wir Ordnen Sie die Daten in aufsteigender Reihenfolge an. Wir definieren:

$$\text{mittlerer Rang} = \frac{n+1}{2}$$

$$\text{Quartilrang} = \frac{\text{mittlerer Rang} + 1}{2},$$

Dabei ist *x* der nach unten gekürzte Wert von *x* auf die nächstkleinere ganze Zahl.

In seinem Buch **Tukey** nennt JW das erste und die vierten Quartile „Scharniere“.

DOMÄNEN UND EINSCHRÄNKUNGEN

In der explorativen Datenanalyse definiert durch Tukey, JW, es gibt vier Hauptthemen. Dies sind **grafische Darstellungen,** Re-Expression (was einfach die **Transformation** von Variablen ist), Residuen und Widerstand (was synonym für das Konzept ist). der Robustheit). Der Widerstand ist ein Maß der Sensibilität der Analyse oder Zusammenfassung gegenüber „schlechte Daten“. Die Notwendigkeit, den Widerstand zu studieren spiegelt die Tatsache wider, dass selbst „gute Daten selten.“ Enthält weniger als 5 % Fehler und ist daher wichtig, um die Analyse davor schützen zu können die nachteiligen Auswirkungen von Fehlern“. *Die widerstandsfähige Linie* von Tukey sorgt für eine robuste Passform an einer Reihe von Punkten. Das bedeutet, dass diese Linie nicht übermäßig empfindlich auf eine bestimmte Beobachtung reagiert. Der Median ist sehr resistent, der Mittelwert jedoch nicht. Grafische Darstellungen werden verwendet, um das Verhalten der Daten, Datenanpassungen, Diagnosemaßnahmen und Residuen zu analysieren. Auf diese Weise wir kann unerwartete Merkmale erkennen und ihre Regelmäßigkeit in den Daten erkennbar. Die Entwicklung der explorativen Daten Die Analyse ist eng mit der Betonung der visuellen Darstellung und der Verwendung von verbunden eine Vielzahl relativ neuer grafischer Techniken, wie etwa das **Stamm- und** Blattdiagramm und der **Boxplot.**

Machine Translated by Google

BEISPIELE

Einkommensindizes der Schweizer Kantone im Jahr 1993

Kanton	Index Kanton	Index
Zürich	125,7 Schaffhouse	99,2
Bern	86.2 Appenzell Rh.-Ext.	84,3
Luzern	87,9 Appenzell Rh.-Int.	72,6
Uri	88,2 St. Gallen	89,3
Schwytz	94,5 Graubünden	92,4
Obwald	80,3 Argovie	98,0
Nidwald	108,9 Thurgau	87,4
Glarus	101,4 Tessin	87,4
Zug	170,2 Waadt	97,4
Freiburg	90,9 Wallis	80,5
Soleure	88,3 Neuchâtel	87,3
Basel-Stadt 124.2 Genf		116,0
Basel-Landschaft 105,1 Jura		75.1

Die Tabelle enthält die Einkommensindizes von Schweizer Kantone pro Einwohner (Schweiz = 100) im Jahr 1993. Wir können die zusammenfassende Statistik mit fünf Zahlen berechnen:

$$\text{mittlerer Rang} = \frac{n+1}{2} = \frac{26+1}{2} = 13{,}5 .$$

Daher ist der Median Md der Mittelwert von dreizehnte und vierzehnte Beobachtung:

$$\text{Md} = = \frac{89{,}3 + 90{,}9}{2} = 90{,}1 .$$

Dann berechnen wir:

$$\text{Quartilrang} = \frac{13+1}{2} = 7 .$$

Das erste Quartil ist die siebte Beobachtung ab Beginn der Daten (sofern angeordnet).

in aufsteigender Reihenfolge) und das dritte Quartil wird sei die siebte Beobachtung vom Ende von die Daten:

$$\text{erstes Quartil} = 87{,}3$$

$$\text{drittes Quartil} = 101{,}4 .$$

Das Minimum und das Maximum betragen jeweils 72,6 und 170,2.

Dies ergibt die folgende fünfstellige Summe Maria:

90.1
87,3 101,4
72,6 170,2

Widerstand (oder Robustheit) des Medians in Bezug auf den Durchschnitt

Wir betrachten die folgenden Zahlen: 3, 3, 7, 7, 11, 11.

Wir berechnen zunächst den Mittelwert dieser Zahlen:

$$\bar{x} = \frac{\sum_{i=1}^{N} xi}{N} = \frac{3+3+7+7+11+11}{6} = 7 .$$

Dann berechnen wir den Median Md:

$$\text{mittlerer Rang} = \frac{n+1}{2} = \frac{6+1}{2} = 3{,}5$$

$$\text{Md} = = \frac{7+7}{2} .$$

Wir stellen fest, dass in diesem Fall der Mittelwert und der Der Median beträgt beide gleich 7.

Nehmen wir nun an, wir fügen eine weitere Zahl hinzu: ÿ1000. Wir berechnen daher den Mittelwert neu und der Median der folgenden Zahlen: ÿ1000, 3, 3, 3, 7, 11, 11. Der Mittelwert ist:

$$\bar{x} = \frac{\sum_{i=1}^{N} xi}{N} = \frac{\text{ÿ}1000+3+3+7+7+11+11}{7} = \text{ÿ}136{,}86 .$$

Es ist klar, dass die Anwesenheit auch nur eines „schlechten“ ist „Datum“ kann den Mittelwert stark beeinflussen.

Der Median beträgt:

$$\text{mittlerer Rang} = \frac{n+1}{2} = \frac{7+1}{2} = 4$$

$$\text{Md} = 7 .$$

Machine Translated by Google

Im Gegensatz zum Mittelwert ist dies beim Median nicht der Fall Änderung, die uns sagt, dass der Median ist resistenter gegen Extremwerte („schlechte Daten“) als der Mittelwert.

WEITERLESEN

- **Box-Plot**
- **Grafische Darstellung**
- **Restwert**
- **Stamm-Blatt-Diagramm**
- **Transformation**

VERWEISE

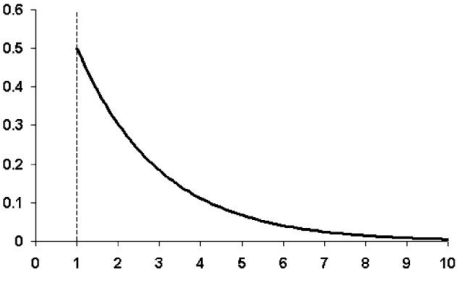
Tukey, JW: Einige grafische und semigrafische Darstellungen. In: Bancroft, TA (Hrsg.) Sta tistical Papers in Honor of George W. Snedecor. Iowa State University Press, Ames, IA, S. 293–316 (1972)

Tukey, JW: Explorative Datenanalyse. Addison-Wesley, Reading, MA (1977)

Exponentiell Verteilung

Eine **Zufallsvariable** *X* folgt einer Exponentialverteilung mit **dem Parameter** ÿ, wenn ihre **Dichtefunktion** gegeben ist durch:

$$f(x) = \begin{cases} \ddot{y} \cdot e^{\ddot{y}\ddot{y}x}, \text{ wenn } x \ddot{y} 0; \ddot{y} > 0 \\ 0 \quad \text{wenn nicht} \end{cases} .$$



Exponentielle Verteilung, ÿ = 1, ÿ = 2

Die Exponentialverteilung wird auch Exponentialverteilung genannt die negative Exponentialfunktion.

Die Exponentialverteilung ist eine **stetige Wahrscheinlichkeitsverteilung.**

MATHEMATISCHE ASPEKTE

Die **Verteilungsfunktion** eines **Zufalls**

Die Variable X , die einer Exponentialverteilung mit **dem Parameter** ÿ folgt, ist wie folgt:

$$F(x) = \int_0^x \ddot{y} \cdot e^{\ddot{y}\ddot{y}t}\, dt$$

$$= 1 \ddot{y} e^{\ddot{y}\ddot{y}x} .$$

Der **Erwartungswert** ergibt sich aus:

$$E[X] = \ddot{y} \frac{1}{\ddot{y}} .$$

Seit wir ... Haben

$$E\,X2 = \ddot{y}\, 2 \frac{2}{\ddot{y}} .$$

die **Varianz** ist gleich:

$$\mathrm{Var}(X) = E\,X2 \ddot{y} (E[X])^2$$

$$= \frac{2}{\ddot{y}\, 2} \frac{1}{\ddot{y}\, 2} = \frac{1}{\ddot{y}\, 2} .$$

Die Exponentialverteilung ist die **kontinuierliche Wahrscheinlichkeitsverteilung** analog zur **geometrische Verteilung.**

Es handelt sich tatsächlich um einen Sonderfall des **Gammas Verteilung,** wobei der **Parameter** ÿ des Die Gammaverteilung ist gleich 1.

Es entsteht eine **Chi-Quadrat-Verteilung** mit zwei **Freiheitsgrade ,** wenn der **Parameter** ÿ der Exponentialverteilung gleich ist bis $\frac{1}{2}$.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Exponentialverteilung wird verwendet Beschreiben Sie zufällige **Ereignisse** in der Zeit. Für die Prüfung Beispielsweise ist die Lebensdauer ein Merkmal, das häufig durch eine exponentielle **Zufallsvariable dargestellt wird.**

Machine Translated by Google

WEITERLESEN

Chi-Quadrat-Verteilung
Kontinuierliche Wahrscheinlichkeitsverteilung
Gammaverteilung
Geometrische Verteilung

Extrapolation

Statistisch gesehen ist eine Extrapolation eine **Schätzung** der **abhängigen Variablen** für Werte einer unabhängigen Variablen, die außerhalb der Menge der **Beobachtungen liegen.**

Extrapolation wird häufig in **Zeitreihen** verwendet , wenn ein **Modell** ermittelt wird

Die zu den Zeitpunkten t1 , *t2,...,tn* beobachteten Werte u1, *u2,...,un* (mit t1 < t2 < < *tn)* werden verwendet, um den Wert der Variablen zum Zeitpunkt tn+1 vorherzusagen.

Im Allgemeinen kann eine **einfache lineare Regressionsfunktion** basierend auf den **Beobachtungen** *(Xi,* Yi), *i = 1, 2,...,n* verwendet werden, um **Werte** von *Y* für Werte von *X* zu schätzen , die außer

die Menge X1, *X2,...,Xn.* Es ist auch möglich, Extrapolationen aus **einem multilinearen Regressionsmodell** zu erhalten .

E

DOMÄNEN UND EINSCHRÄNKUNGEN

Bei der Verwendung von Extrapolationen ist Vorsicht geboten, da diese auf der **Hypothese** beruhen , dass sich das **Modell** für die **Werte** der für die Extrapolation verwendeten unabhängigen Variablen nicht ändert .

WEITERLESEN

Prognose
Multiple lineare Regression
Einfache lineare Regression
Zeitfolgen

Machine Translated by Google

Faktor

Als Faktor bezeichnet man jeden kontrollierbaren Zustand in einem **Experiment.** Jeder Faktor kann eine vorgegebene Anzahl von **Werten annehmen,** die als **Stufen** des Faktors bezeichnet werden. **Experimente,** die Faktoren verwenden, werden **faktische Experimente** genannt .

BEISPIELE

Ein Soziologe möchte die Meinung der Einwohner einer Stadt zur Drogenproblematik herausfinden. Ihn interessiert der Einfluss von Geschlecht und Einkommen auf die Einstellung zu diesem Thema.

Die **Stufen** des Faktors „Geschlecht“ sind: männlich (M) und weiblich (F); Die Stufen des Faktors „Einkommen“ entsprechen niedrig (L), Durchschnittsalter (A) und hoch (H), wobei die Aufteilung zwischen diesen Stufen durch die Erfahrung festgelegt wird Mentor.

Der faktorielle Raum G dieses **Experiments** setzt sich aus den folgenden sechs **Kombinationen** der **Niveaus** der beiden Faktoren zusammen:

$$G = \{(M,L);(M,A);(M,H);(F,L);$$
$$(F,A);(F,H)\} .$$

In diesem Beispiel ist der Faktor „Geschlecht“ qualitativ und nimmt Nominalwerte an ; wohingegen der Faktor „Einkommen“ quantitativ ist und Ordnungswerte annimmt.

WEITERLESEN

- Varianzanalyse
- Versuchsplanung
- Experiment
- Faktorielles Experiment
- Behandlung

VERWEISE

Raktoe, BL Hedayat, A., Federer, WT: Faktorielle Designs. Wiley, New York (1981)

Faktorebene

Die Stufen eines **Faktors** sind die unterschiedlichen Werte dass es dauern kann. Der **Versuchsaufbau** legt fest, welche Niveaus aus allen möglichen Kombinationen solcher Niveaus im **Experiment verwendet werden.**

WEITERLESEN

- Versuchsplanung
- Experiment
- Faktor
- Behandlung

Faktorielle Achse

Betrachten Sie S, eine quadratische **Matrix** der Ordnung n. Wenn *Xi* der **Eigenvektor** von S ist, der dem **Eigenwert** ki zugeordnet ist , sagen wir, dass *Xi* die i-te Fak-

Machine Translated by Google

Torialachse; genauer gesagt ist es die **Achse ,** die den Vektor Xi trägt .

Wir ordnen Eigenwerte im Allgemeinen in absteigender Reihenfolge an, was die Notwendigkeit erklärt, die Fakultätsachsen zu nummerieren.

GESCHICHTE

Siehe **Korrespondenzanalyse.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Konzept einer Faktorachse wird in der Korrespondenzanalyse verwendet.

WEITERLESEN

Korrespondenzanalyse
Eigenwert
Eigenvektor
Trägheit

VERWEISE

Siehe **Korrespondenzanalyse.**

Faktorielles Experiment

Ein faktorielles Experiment ist ein **Experiment ,** bei dem alle möglichen **Behandlungen ,** die sich aus zwei oder mehr **Faktoren** ableiten lassen , wobei jeder Faktor zwei oder mehr **Ebenen hat,** so untersucht werden, dass die **Haupteffekte** und **Wechselwirkungen** untersucht werden können.

Der Begriff „Faktorenexperiment“ beschreibt ein Experiment, bei dem alle verschiedenen Faktoren auf alle möglichen Arten kombiniert werden, spezifiziert jedoch nicht den Versuchsaufbau, der zur Durchführung eines solchen Experiments verwendet wird.

MATHEMATISCHE ASPEKTE

In einem faktoriellen Experiment kann eine **Modellgleichung** verwendet werden , um die **abhängige Variable** mit der **unabhängigen Variablen in Beziehung zu setzen.**

Wenn also das faktorielle Experiment zwei Faktoren impliziert , sieht das zugehörige **Modell** wie folgt aus:

$$Yijk = ÿ + ÿi + ÿj + (ÿÿ)ij + ÿijk ,$$

$i = 1, 2,...,a$ **(Stufen** des **Faktors** A) ,

$j = 1, 2,...,b$ (Stufen des Faktors B) ,

$k = 1, 2,...,nij$ ÿ Anzahl der **Experimentale Einheiten,** die ÿ die **Behandlung** erhalten *ij*

Wo:

ÿ ist der allgemeine **Mittelwert** aller Behandlungen, ist die

ÿi Wirkung der i-ten Fac-Ebene für A,

ÿj ist die Auswirkung der j-ten Ebene von Faktor B, **(ÿÿ)ij** ist die Auswirkung der Wechselwirkung zwischen der i-ten Ebene von A, der j-ten Ebene von B und

ÿijk ist der experimentelle **Fehler** in der Beobachtung *Yijk.*

Im Allgemeinen ist ÿ eine Konstante und *ijk* eine **Zufallsvariable,** die gemäß der **Normalverteilung** mit einem **Mittelwert** von Null und einer **Varianz** von ÿ2 verteilt ist.

Wenn der Interaktionsterm (ÿÿ)ij unterdrückt wird, erhält man ein additives **Modell ,** was bedeutet, dass die Wirkung jedes **Faktors** eine Konstante ist addiert sich zum allgemeinen **Mittelwertÿ,** und dieser hängt nicht von der **Höhe** des anderen Faktors ab. Dies wird als Modell ohne **Interaktion** bezeichnet **tionen.**

DOMÄNEN UND EINSCHRÄNKUNGEN

In einem faktoriellen Experiment, bei dem alle **Faktoren** die gleiche Anzahl an **Stufen haben,** ergibt sich die Anzahl der im Experiment verwendeten Behandlungen normalerweise aus der Anzahl der Stufen, die mit der Zahl potenziert werden

Machine Translated by Google

von Faktoren. Zum Beispiel für ein Experiment
Das Experiment verwendet drei Faktoren mit
jeweils zwei Ebenen und wird als 23- Faktoren-
Experiment bezeichnet, das acht **Behandlungen verwendet.**

Die am häufigsten verwendete Notation für
Kombinationen von Behandlungen , die tatsächlich verwendet werden
Das experimentelle Experiment kann wie folgt beschrieben werden.
Verwenden Sie einen Großbuchstaben, um den **Faktor** zu bezeichnen
und einen numerischen Index zur Angabe seines **Niveaus.**
Betrachten Sie zum Beispiel ein faktorielles Experiment
mit drei Faktoren: einem Faktor A mit zwei Stufen,
ein Faktor B mit zwei Ebenen und ein Faktor C mit
drei Ebenen. Die 12 Kombinationen sind:

A1B1C1 , A1B1C2 , A1B1C3 ,
A1B2C1 , A1B2C2 , A1B2C3 ,
A2B1C1 , A2B1C2 , A2B1C3 ,
A2B2C1 , A2B2C2 , A2B2C3 .

Manchmal sind nur die Indizes (geschrieben im
Reihenfolge wie die Faktoren):

111 , 112 , 113 ,
121 , 122 , 123 ,
211 , 212 , 213 ,
221 , 222 , 223 .

Doppelfaktorielle Experimente oder allgemeiner
mehrfaktorielle Experimente sind
aus folgenden Gründen wichtig:

- Ein doppelfaktorielles Experiment verwendet
  Ressourcen effizienter als zwei **Experimente,**
  die jeweils einen einzelnen **Faktor verwenden.**
  Der erste nimmt weniger Zeit in Anspruch und erfordert weniger
  weniger **Versuchseinheiten** zu erreichen
  ein bestimmtes **Maß** an Präzision.
- Ein doppelfaktorielles Experiment ermöglicht die
  Einfluss eines zu untersuchenden Faktors
  jede Ebene des anderen Faktors, weil die
  Die Niveaus beider Faktoren sind unterschiedlich. Dies führt dazu
  zu Schlussfolgerungen, die über einen größeren Zeitraum hinweg gültig sind
  Bereich der experimentellen Bedingungen als wenn
  Es wird eine Reihe von Ein-Faktor-Designs verwendet.
- Schließlich, gleichzeitige Forschung an zwei
  Faktoren sind bei **Interaktionen** notwendig
  zwischen den Faktoren vorhanden sind, Bedeutung
  dass die Wirkung eines Faktors von der abhängt
  Höhe des anderen Faktors.

WEITERLESEN

Versuchsplanung
Experiment

VERWEISE

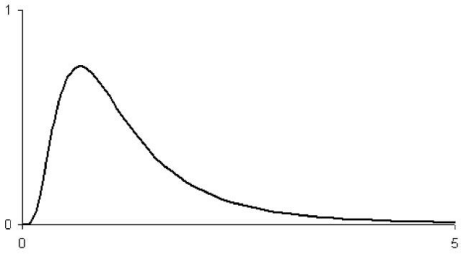
Dodge, Y.: Wichtigste Erfahrungspläne.
In: Aeschlimann, J., Bonjour, C., Stock er, E.
(Hrsg.) Méthodologies et Techniques
Erlebnispläne: 28. Kurs
perfectionnement de l'Association vau doise
des chercheurs en physique, Saas Fee, 2.–8.
März 1986 (1986)

Fisher-Verteilung

Die **Zufallsvariable** *X* folgt einem Fisher
Verteilung, wenn ihre **Dichtefunktion** die annimmt
form:

$$f(x) = \frac{\frac{m+n}{2}}{\frac{M}{2}\,\frac{N}{2}} \left(\frac{M}{N}\right)^{\frac{M}{2}} \cdot \frac{X^{\frac{m\ddot{y}2}{2}}}{1+\frac{M}{N}X^{\frac{m+n}{2}}},$$

Dabei ist die Gammafunktion (siehe
Gammaverteilung) und *m, n* die **Grade**
der Freiheit *(m, n = 1, 2,...).*



Fisher-Verteilung, m = 12, n = 8

F

Machine Translated by Google

Die Fisher-Verteilung ist **kontinuierlich**

Wahrscheinlichkeitsverteilung.

GESCHICHTE

Die Fisher-Verteilung wurde von entdeckt
Fisher, RA im Jahr 1925. Das Symbol F wird verwendet
zur Bezeichnung der Fisher-Verteilung wurde 1934
von Snedecor zu Ehren von Fisher, RA, eingeführt.

MATHEMATISCHE ASPEKTE

Wenn *U* und *V* zwei unabhängige **Zufallszahlen** sind
Variablen, die jeweils einem **Chi-Quadrat** folgen
Verteilung, jeweils mit *m* und *n*
Freiheitsgrade, dann ist die Zufallsvariable:

$$F = \frac{\text{Äh}}{\text{V/n}}$$

folgt einer Fisher-Verteilung mit *m* und *n*

Freiheitsgrade.

Der erwartete Wert einer **Zufallsvariablen F**
die einer Fisher-Verteilung folgt, ist gegeben durch:

$$E[F] = n \frac{N}{\text{ÿ } 2} \quad \text{für } n > 2 \quad ,$$

und die **Varianz** ist gleich:

$$\text{Var(F)} = 2 \frac{2n2\ (m + n \text{ ÿ } 2)}{m\ (n \text{ ÿ } 2)} \quad n > 4 \quad .$$

Die Fisher-Verteilung mit 1 und *v* **Grad**
der Freiheit ist identisch mit dem Quadrat der
Studentenverteilung mit *v* Freiheitsgraden.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Bedeutung der Fisher-Verteilung in
Die statistische Theorie hängt mit ihrer Anwendung zusammen
zur Verteilung des Verhältnisses der Unabhängigen
Schätzer der **Varianz.** Derzeit ist dies
Die Verteilung wird am häufigsten in der verwendet
Standardtests im Zusammenhang mit **der Analyse von**
Varianz- und **Regressionsanalyse .**

WEITERLESEN

- **Varianzanalyse**
- **Chi-Quadrat-Verteilung**
- **Kontinuierliche Wahrscheinlichkeitsverteilung**
- **Fisher-Tisch**
- **Fisher-Test**
- **Studentenverteilung**

VERWEISE

Fisher, RA: Statistische Methoden für
Forschungsmitarbeiter. Oliver & Boyd, Edinburgh
(1925)

Snedecor, GW: Berechnung und Interpretation der
Varianz- und Kovarianzanalyse. Collegiate,
Ames, IA (1934)

Fisher-Index

Der Fisher-Index **ist eine zusammengesetzte**
Indexzahl , die es uns ermöglicht, den Anstieg zu untersuchen
die Lebenshaltungskosten (Inflation). Es ist das
geometrische Mittel zweier Indexzahlen:

- Der **Laspeyres-Index** und
- Der **Paasche-Index.**

$$\text{Fisher-Index} = \sqrt{\text{(Laspeyres-Index)} \times \text{(Paasche-Index)}} \quad .$$

GESCHICHTE

1922 wurde der Ökonom und Mathematiker
Fisher und Irving erstellten ein Modell, das darauf basiert
den Fisher-Index, um einige zu umgehen
Probleme im Zusammenhang mit der Verwendung der **Indexnummer**
Mitglieder von Laspeyres und Paasche:
Da der **Laspeyres-Index** immer das verwendet
Menge der zunächst als Referenz verkauften Waren,
Es kann dazu führen, dass die Kostensteigerung überschätzt wird
des Lebens (weil der Konsum von mehr
teure Waren können mit der Zeit fallen).
Umgekehrt gilt seit **dem Paasche-Index** immer
Verwendet die aktuell verkaufte Warenmenge

Machine Translated by Google

Zeitraum als Referenz, es kann unterschätzt werden
Erhöhung der Lebenshaltungskosten.
Denn der Fisher-Index ist der **geometrische**
Der Mittelwert dieser beiden Indizes kann als
idealer Kompromiss angesehen werden
ihnen.

MATHEMATISCHE ASPEKTE

Unter Verwendung der Definitionen **des Laspeyres-Index**
undderPaasche-Index,der FisherindexFn/0
wird wie folgt berechnet:

$$Fn/0 = 100 \cdot \sqrt{\frac{PnQ0 \cdot PnQn}{P0Q0 \cdot P0Qn}},$$

wobei Pn und Qn den Warenpreis darstellen
und die Menge davon, die derzeit verkauft wird
Periode n, während P0 und Q0 der Preis und sind
Verkaufsmenge der gleichen Ware im
Referenzzeitraum 0. Es ergeben sich unterschiedliche Werte
für verschiedene Waren.
Beachten Sie, dass der Fisher-Index reversibel ist.
bedeutet, dass:

$$Fn/0 = \frac{1}{F0/n}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Auch wenn der Fisher- **Index** eine Art ist
der „idealen Kompromissindexzahl“ ist es
werden in der Praxis selten verwendet.

BEISPIELE

Betrachten Sie die folgende fiktive Tabelle mit den
Preisen und den jeweiligen Mengen
Verkauf von drei Konsumgütern zum Referenzwert
Jahr 0 und im aktuellen Jahr n.

	Mengen (Tausende)		Preis (Euro)	
Waren	1970	1988	1970	1988
	(Q0)	*(Qn)*	(P0)	*(Pn)*
Milch	50,5	85,5	0,20	1,20
Brot	42,8	50,5	0,15	1.10
Butter	15.5	40,5	0,50	2,00

Aus der folgenden Tabelle haben wir:

$$PnQn = 239,15,$$

$$P0Qn = 44,925,$$

$$PnQ0 = 138,68 \text{ und}$$

$$P0Q0 = 24,27.$$

Waren	$PnQn$	P_0Qn	PnQ_0	P_0Q_0
Milch	102,60	17,100	60,60	10,10
Brot	55,55	7,575	47,08	6.42
Butter	81,00	20,250	31,00	7,75
Gesamt	239,15	44,925	138,68	24,27

Wir können dann den **Paasche-Index finden:**

$$In/0 = \frac{Pn \cdot Qn}{P0 \cdot Qn} \cdot 100$$

$$= \frac{239.15}{44,925} \cdot 100 = 532,3,$$

und der **Laspeyres-Index:**

$$In/0 = \frac{Pn \cdot Q0}{P0 \cdot Q0} \cdot 100$$

$$= \frac{138,68}{24.27} \cdot 100 = 571,4.$$

Der Fisher-Index ist die Quadratwurzel von
Produkt der Indexzahlen von Paasche und
von Laspeyres (oder das **geometrische Mittel** der
zwei **Indexnummern):**

$$\text{Fisher-Index} = \sqrt{532,3 \times 571,4} = 551,5.$$

Laut Fisher-Index beträgt der Preis von
Die betrachtete Ware ist um 451,5 % gestiegen
(551,5ÿ100) zwischen dem Referenzjahr und
das laufende Jahr.

WEITERLESEN

Zusammengesetzte Indexnummer
Indexnummer

Laspeyres-Index
Paasche-Index

Einfache Indexnummer

F

Machine Translated by Google

VERWEISE

Fisher, I.: Die Entstehung von Indexzahlen.
Houghton Mifflin, Boston (1922)

Fisher, Irving

Fisher, Irving wurde am 27. Februar 1867 in New York geboren. Er promovierte 1892 an der Yale University. Sein mathematischer Ansatz zur Theorie der Werte und Preise, der erstmals in seiner Dissertation beschrieben wurde, führte dazu, dass er als einer der ersten Amerikaner bekannt wurde mathematische Ökonomen. Er gründete

1930 gründete er mit Frisch, Ragnar und Roos, Charles F. die Econometrische Gesellschaft und wurde ihr erster Präsident. Fisher, I. blieb während seiner gesamten Karriere an der Yale University. Er unterrichtete zunächst Mathematik, dann Wirtschaftswissenschaften. Nach einem Aufenthalt in Europa wurde er 1898 zum Professor für Sozialwissenschaften ernannt und starb 1947 in New York.

Fisher war ein produktiver Autor: Seine Publikationsliste umfasst mehr als 2000 Titel. Er interessierte sich sowohl für Wirtschaftstheorie als auch für wissenschaftliche Forschung. 1920 schlug er ökonometrische und statistische Methoden für vor Berechnen von Indizes. Sein „Idealindex“, den wir jetzt Fisher-Index nennen , ist das geometrische Mittel des **Laspeyres-Index** und des **Paasche-Index.**

Einige Hauptwerke und Artikel von Fisher, Irving:

1892 Mathematische Untersuchungen zur Wert- und Preistheorie. Connecticut Academy of Arts and Science, New Haven, CT.

1906 Die Natur von Kapital und Einkommen. Macmillan, New York.

1907 Der Zinssatz: Seine Natur, Bestimmung und Bezug zu Öko nomische Phänomene. Macmillan, New York.

1910 Einführung in die Wirtschaftswissenschaften. Macmillan, New York.

1911 Die Kaufkraft des Geldes. Macmillan, New York.

1912 Grundprinzipien der Wirtschaftswissenschaften. Macmillan, New York.

1921 Die beste Form der Indexnummer. Bin. Stat. Assoc. Quart., 17, 533–537.

1922 Die Herstellung von Indexzahlen. Houghton Mifflin, Boston, MA.

1926 Eine statistische Beziehung zwischen unem Einsatz- und Preisänderungen. Int. Labour Rev., 13, 785–792.

1927 Eine statistische Methode zur Messung des „Grenznutzens“ und zur Prüfung der Gerechtigkeit einer progressiven Einkommensteuer. In: Hollander, JH (Hrsg.) Economic Essays Contributed in Honor of John Bates Clark. Macmillan, New York.

1930 Die Theorie des Interesses, bestimmt durch die Ungeduld, Einkommen auszugeben, und die Möglichkeit, es zu investieren Macmillan, New York.

1932 Booms und Depressionen. Adelphi, New York.

1933 Die Schulden-Deflation-Theorie der Weltwirtschaftskrise. Econmetrica, 1, 337–357.

1935 100 % Geld. Adelphi, New York.

WEITERLESEN

Fisher-Index

Indexnummer

Machine Translated by Google

Fisher, Ronald Aylmer

Der 1890 in East Finchley bei London geborene Fisher, Ronald Aylmer studierte Mathematik in Harrow und am Gonville and Caius College in Cambridge. Er begann seine statistische Karriere im Jahr 1919, als er seine Arbeit am Institut aufnahm

Lehrstuhl für Agrarforschung in Rothamsted („Rothamsted Experimental Station"). Die Zeit, die er dort bis 1933 verbrachte, war sehr produktiv. 1933 wechselte er als Professor für Eugenik an das University College London und übernahm von 1943 bis 1957 den Balfour-Lehrstuhl für Genetik

Cambridge. Auch nach seiner Pensionierung führte Fisher einige Forschungsarbeiten für die Mathematical Statistics Division des Common durch Reichtum Wissenschaftliche und industrielle Forschung Organisation in Adelaide, Australien, wo er 1962 starb.

Fisher gilt als einer der Begründer der modernen **Statistik.** Sehr bekannt

Auf dem Gebiet der Genetik leistete er außerdem wichtige Beiträge zur allgemeinen Theorie **der Maximum-Likelihood-Schätzung.** Seine Arbeit an

Erwähnenswert ist auch das experimentelle Design, das für die Untersuchung landwirtschaftlicher **Versuchsdaten** erforderlich ist, sowie seine Entwicklung der Technik zur **Analyse von Vari ance.**

Fisher entwickelte die Prinzipien der **Randomisierung,** der randomisierten **Blöcke, der lateinischen Quadrate** und der faktoriellen **Anordnungen.**

Einige der Hauptwerke und Artikel von Fisher, RA:

1912 Über ein absolutes Kriterium zur Anpassung von Frequenzkurven.Mess.Math., 41, 155– 160.

1918 Die Korrelation zwischen Verwandten auf die Annahme von Mendelian inher itance. Trans. Roy. Soc. Edinburgh, 52, 399–433.

1920 Eine mathematische Untersuchung von Methoden zur Bestimmung der Genauigkeit einer Beobachtung anhand des mittleren Fehlers und des mittleren quadratischen Fehlers. Mo. Nicht. R. Astron. Soc., 80, 758–770.

1921 Über den „wahrscheinlichen Fehler" eines daraus abgeleiteten Korrelationskoeffizienten eine kleine Kostprobe. Metron, 1(4), 1–32.

1922 Über die mathematischen Grundlagen von Theoretische Statistik. Phil. Trans. A, 222, 309–368.

1925 Theorie der statistischen Schätzung. Proz. Camb. Philos. Soc., 22, 700–725.

1925 Statistische Methoden für die Forschung Arbeitskräfte. Oliver & Boyd, London.

1934 Wahrscheinlichkeit, Wahrscheinlichkeit und Menge von Informationen in der Logik der unsicheren Inferenz.Proc.Roy.Soc.A, 146, 1–8.

1935 Die Versuchsplanung. Oliver & Boyd, Edinburgh.

1971–1974 Bennett, JH (Hrsg.) Collected Papers of RA Fisher. Univ. Adelaye, Adelaide, Australien

WEITERLESEN

Fisher-Verteilung

Fisher-Tisch

Fisher-Test

VERWEISE

Fisher-Box, J.: RA Fisher, das Leben eines Wissenschaftlers. Wiley, New York (1978)

F

Machine Translated by Google

Fisher-Tisch

Die Fisher-Tabelle gibt die Werte der **Verteilungsfunktion** einer **Zufallsvariablen** an **in der Lage ,** die einer Fisher-Verteilung folgt .

GESCHICHTE

Siehe **Fisher-Verteilung.**

MATHEMATISCHE ASPEKTE

Die **Zufallsvariable** *X* folge einem **Fisher Verteilung** mit *m* und *n* Freiheitsgraden. Die **Dichtefunktion** der Zufallsvariablen *X* ist gegeben durch:

$$f(t) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{M}{2}\right)\Gamma\left(\frac{N}{2}\right)} \left(\frac{M}{N}\right)^{\frac{M}{2}} \cdot \frac{t^{\frac{m}{2}-1}}{\left(1+\frac{M}{N}t\right)^{\frac{m+n}{2}}}, \quad t > 0 \quad ,$$

wo stellt die Gammafunktion dar (siehe **Gammaverteilung).**

Angesichts der Freiheitsgrade *m* und n ermöglicht uns die Fisher-Tabelle, den **Wert** *x* zu bestimmen , für den die **Wahrscheinlichkeit** P(X ÿ x) einem bestimmten Wert entspricht.

Die am häufigsten verwendeten Fisher-Tabellen geben die Werte von *x* für an

$$P(X \geq x) = 95\ \% \text{ und } (X \geq x) = 97{,}5\ \% \ .$$

Im Allgemeinen verwenden wir $(x =) F_{m,n,\alpha}$, um den Wert der ZufallsvariablenX zu symbolisieren, für die

$$P\left(X \geq F_{m,n,\alpha}\right) = 1 - \alpha \ .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Fisher-Tabelle wird beim Testen von Hypothesen verwendet, wenn es um verteilte **Statistiken** geht gemäß einer **Fisher-Verteilung,** und insbesondere während **der Varianzanalyse** und in der **einfachen** und **multiplen linearen Regressionsanalyse.**

Merrington und Thompson (1943) erstellten Fisher-Tabellen für Werte mit einer Länge von bis zu fünf Dezimalstellen, und Fisher und Yates (1938) erstellten Tabellen mit einer Länge von bis zu zwei Dezimalstellen.

BEISPIELE

Wenn *X* einer **Fisher-Verteilung** mit *m* = folgt 10 und *n* = 15 Freiheitsgrade beträgt der **Wert** von *x* , der **einer Wahrscheinlichkeit** von 0,95 entspricht , 2,54:

$$P(X \geq 2{,}54) = 0{,}95 \ .$$

Ebenso beträgt der Wert von *x* , der einer Wahrscheinlichkeit von 0,975 entspricht, 3,06:

$$P(X \geq 3{,}06) = 0{,}975 \ .$$

Ein Beispiel für die Verwendung der Fisher-Tabelle finden Sie unter **Fisher-Test.**

WEITERLESEN

Varianzanalyse
Fisher-Verteilung
Multiple lineare Regression
Einfache lineare Regression
Statistische Tabelle

VERWEISE

Fisher, RA, Yates, F.: Statistische Tabellen für die biologische, landwirtschaftliche und medizinische Forschung. Oliver und Boyd, Edinburgh und London (1963)

Merrington, M., Thompson, CM: Tabellen mit Prozentpunkten der invertierten Beta- Verteilung *(F)* . Biometrie **33,** 73–88 (1943)

Machine Translated by Google

Fisher-Test

Ein Fisher-Test ist ein Hypothesentest zum Beobachteter **Wert** einer **Statistik** der Form:

$$F = \frac{U}{M} \cdot \frac{N}{V},$$

Dabei sind U und V unabhängige Zufallsvariablen, die jeweils einer **Chi-Quadrat-Verteilung** mit *m* bzw. *n* Grad folgen der Freiheit.

GESCHICHTE

Siehe **Fisher-Verteilung.**

MATHEMATISCHE ASPEKTE

Wir betrachten ein lineares **Modell** mit einem **Faktor** angewendet auf *t*- Ebenen:

$$Yij = \ÿ + \ÿi + \ÿij ,$$

$$i = 1, 2,...,t ; j = 1, 2,...,ni ,$$

Wo

Yij stellt die j-te **Beobachtung** dar, die empfangen wird die **Behandlung** ich,

ÿ ist der allgemeine **Mittelwert** aller Behandlungen,

ÿi ist die tatsächliche Auswirkung der Behandlung *i* auf die Beobachtung und

ÿij ist der experimentelle **Fehler** in der Beobachtung *Yij.*

Die Zufallsvariablen sind unabhängig und normalverteilt mit Mittelwert 0 und **Varianz** ÿ2 : N(0, ÿ2).
In diesem Fall gilt die **Nullhypothese, die** bestätigt dass es einen signifikanten Unterschied gibt
Die *t*- Behandlungen sind wie folgt:

$$H0 : \ÿ1 = \ÿ2 = \cdots = \ÿt .$$

Die **Alternativhypothese** lautet wie folgt:

H1 : die Werte von ÿi(i = *1, 2,...,t)* sind nicht alle identisch.

Um die Variabilität innerhalb der Behandlung zu vergleichen, werden auch Fehler oder Residuen genannt

Um die Variabilität zwischen den Behandlungen zu ermitteln, konstruieren wir ein Verhältnis, bei dem der Zähler ein **Wert** ist

mation der **Varianz** zwischen Behandlungen und der Nenner eine Schätzung des

Abweichung innerhalb der Behandlung oder Fehler. Der Fisch

Eine Teststatistik, auch F-Verhältnis genannt, ist dann gegeben von:

$$F = \frac{\sum_{i=1}^{\tau} ni \cdot \frac{\bar{Y}_{i.} \ÿ \bar{Y}_{..}2}{t \ÿ 1}}{\sum_{i=1}^{\tau} \sum_{j=1}^{ni} \frac{Yij \ÿ \bar{Y}_{i.}2}{N \ÿ t}},$$

Wo

ni ist die Anzahl der Beobachtungeninithtreat ment *(i = 1,...,t)*

$\boldsymbol{N} = \sum_{i=1}^{\tau} \boldsymbol{ni}$ ist die Gesamtzahl der Beobachtungen tionen,

$\bar{\boldsymbol{Y}}_{\boldsymbol{i.}} = \sum_{j=1}^{\boldsymbol{ni}} \frac{\boldsymbol{Yij}}{\boldsymbol{ni}}$ ist der Mittelwert der i-ten Behandlung, für *i = 1,...,t,* und

$\bar{\boldsymbol{Y}}_{\boldsymbol{..}} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{i=1}^{\tau} \sum_{j=1}^{\boldsymbol{ni}} \boldsymbol{Yij}$ ist der globale Mittelwert.

Nach Auswahl eines **Signifikanzniveaus** ÿ
Für den Test vergleichen wir den berechneten Wert *F* mit dem entsprechenden **kritischen Wert** in der **Fisher-Tabelle** für *t* ÿ 1

und *N* ÿ *t* Freiheitsgrade, bezeichnet

Ftÿ1,Nÿt,ÿ.
Wenn F ÿ Ftÿ1,Nÿt,ÿ, lehnen wir die Nullhypothese ab, was bedeutet, dass sich mindestens eine der Behandlungen von den anderen unterscheidet.

Wenn *F < Ftÿ1,Nÿt,ÿ,* lehnen wir die Null nicht ab Hypothese; mit anderen Worten, die Behandlungen Es gibt keinen signifikanten Unterschied, was auch der Fall ist bedeutet, dass die Proben davon abgeleitet sind **Bevölkerung.**

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Fisher-Test wird am häufigsten bei der **Varianzanalyse** in der **Kovarianz** verwendet **Analyse** und in der **Regressionsanalyse.**

BEISPIELE

Siehe **einseitige Varianzanalyse** und **zweiseitige Varianzanalyse.**

WEITERLESEN

Varianzanalyse
Fisher-Verteilung
Fisher-Tisch
Hypothesentest
Regressionsanalyse

Prognose

Prognose ist die Nutzung statistischer Daten Daten, Wirtschaftstheorie und nichtökonomische Bedingungen, um eine angemessene zu bilden Meinung über zukünftige Ereignisse.
Zur Prognose kommen zahlreiche Methoden zum Einsatz Zukunft. Sie basieren alle auf einer begrenzten Anzahl grundlegender Hypothesen.
Eine solche Hypothese bestätigt, dass die beobachteten signifikanten Tendenzen seine eigenen sind torisch wird auch in Zukunft Bestand haben.
Ein anderer besagt, dass messbare Fluktuationen in einer **Variablen** werden bei reproduziert regelmäßigen Abständen, so dass diese Variationen auftreten vorhersehbar werden.
Die in einem Prognoseprozess durchgeführte Zeitreihenanalyse basiert auf zwei Hypothesen:

1. *Langfristige Prognose:*
   Wenn wir langfristige Prognosen erstellen (mehr als fünf Jahre in die Zukunft), Analyse und die Extrapolation der **säkularen Tendenz** werden wichtig.
   Langfristige Prognosen basierend auf dem Säkularen Die Tendenz nimmt normalerweise keine zyklische Fluktuation an Berücksichtigung. Darüber hinaus haben saisonale Schwankungen keinen Einfluss auf die Jahresdaten und werden langfristig nicht berücksichtigt gießen.
2. *Mittelfristprognose:*
   Um die wahrscheinliche Wirkung zu erfassen zyklische Schwankungen berücksichtigen Mechanismus zur Ableitung eines Mittelwerts Prognose (ein bis fünf Jahre im Voraus), multiplizieren wir den **Wert** der Projektion der säkularen Tendenz mit der zyklischen Fluktuation, was den prognostizierten Wert ergibt. Dies, von Natürlich wird davon ausgegangen, dass dieselben Datenstrukturen vorliegen, die die zyklischen Variationen erzeugen wird in Zukunft gelten, und sodas beobachtete Variationen werden weiterhin regelmäßig auftreten.
   In Wirklichkeit haben nachfolgende Zyklen die Möglichkeit, ihre Perioden enorm zu variieren Periodizität, ihre Amplitude und die Modelle, denen sie folgen.
3. *Kurzfristige Prognose:*
   Der **Saisonindex** muss berücksichtigt werden Berücksichtigung in einer kurzfristigen Prognose (einige Monate im Voraus), zusammen mit den prognostizierten Werte für die säkulare Tendenz und zyklisch Schwankungen.
   Im Allgemeinen versuchen wir nicht, unregelmäßige Schwankungen vorherzusagen.

MATHEMATISCHE ASPEKTE

Sei Y eine **Zeitreihe.** Wir können es beschreiben unter Verwendung folgender Komponenten:

- Die **säkulare Tendenz** T;
- Die **saisonalen Schwankungen** S;
- Die **zyklischen Schwankungen** C;
- Die **unregelmäßigen Variationen** I.

Wir unterscheiden zwischen:

- Das multiplikative **Modell** $Y = T \cdot S \cdot C \cdot I$;
- Das additive Modell $Y = T + S + C + I$.

Für einen Wert von t, der größer als die Zeit ist genommen, um eine **Beobachtung durchzuführen,** spezifizieren wir

Machine Translated by Google

Das:

Yt ist eine Prognose des **Wertes** *Y* zum Zeitpunkt t;
Tt ist eine Projektion der säkularen Tendenz zum Zeitpunkt
t; ***Ct*** ist die zyklische Fluktuation zum Zeitpunkt
t; ***St*** ist die Vorhersage für die saisonale Variation zum Zeitpunkt t.

In diesem Fall:

1. *Für eine langfristige Prognose* kann der Wert der Zeitreihe geschätzt werden mit:

$$Yt = Tt .$$

2. *Für eine mittelfristige Prognose* kann der Wert der Zeitreihe geschätzt werden mit:

$Yt = Tt \cdot Ct$ (multiplikatives Modell)

oder $Yt = Tt + Ct$ (additives Modell) .

3. *Für eine kurzfristige Prognose* kann der Wert der Zeitreihe geschätzt werden mit:

$$Yt = Tt \cdot Ct \cdot St$$

(für das multiplikative Modell)

oder $Yt = Tt + Ct + St$

(für das additive Modell).

BEREICHE UND EINSCHRÄNKUNGEN

Planung und Entscheidungsfindung sind zwei Aktivitäten, bei denen es bis zu einem gewissen Grad darum geht, die Zukunft vorherzusagen. Mit anderen Worten: Die Administratoren, die diese Aufgaben wahrnehmen, müssen Prognosen erstellen.

Bemerkenswert ist, dass sich die Gewinnung von Prognosen mittels Zeitreihenanalyse lohnt, auch wenn sich die Prognose meist als ungenau herausstellt. Selbst wenn sie ungenau ist, wird die auf diese Weise gewonnene Prognose wahrscheinlich genauer sein als eine andere Prognose, die nur auf Intuition basiert. Daher ist die Zeitreihenanalyse trotz ihrer Einschränkungen und Probleme ein nützliches Werkzeug im Prognoseprozess.

BEISPIELE

Das Konzept einer „Prognose“ hängt vom Untersuchungsgegenstand ab. Zum Beispiel: • In der Meteorologie:

– Eine kurzfristige Prognose sieht Stunden aus voraus,

– Langfristprognoseblickt Tage in die Zukunft. • In den Wirtschaftswissenschaften:

– Eine kurzfristige Prognose sieht Monate aus voraus;

– Eine mittelfristige Prognose sieht eins zu fünf aus Jahre voraus; –

Eine langfristige Prognose sieht mehr als fünf Jahre voraus.

WEITERLESEN

Zyklische Schwankung
Unregelmäßige Variation
Saisonindex
Saisonale Unterschiede
Säkularen Trend
Zeitfolgen

Bruchfaktorielles Design

Ein fraktionierter faktorieller Versuchsplan ist ein **faktorieller Versuch ,** bei dem nur ein Bruchteil der Kombinationen der **Faktorstufen** berücksichtigt wird Mögliches wird realisiert.
Diese Art von Design wird verwendet, wenn ein **Experiment** eine Reihe von **Faktoren** enthält

Es wird davon ausgegangen, dass ein Faktor wichtiger ist als die anderen, und/oder es gibt eine große Anzahl von Faktorstufen . Das Design ermöglicht eine Reduzierung der Anzahl der benötigten **Versuchseinheiten .**

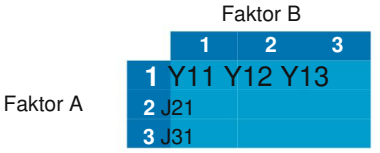
BEISPIELE

Für ein **faktorielles Experiment** mit zwei **Faktoren** mit jeweils drei **Stufen** gibt es neun mögliche Kombinationen. Wenn vier dieser **Beobachtungen** unterdrückt werden, ist ein Bruchteil von

F

Machine Translated by Google

Es bleiben die Beobachtungen übrig, die Schema sein können auf folgende Weise tisiert:



Dabei stellt *Yij* die **Beobachtung** dar , die für die i-te Ebene von Faktor A und die j-te Ebene von Faktor B vorgenommen wurde.

WEITERLESEN

- **Versuchsplanung**
- **Experiment**

REFERENZEN

Yates, F.: Das Design und die Analyse faktorieller Experimente, Techn. Komm. 35, Imperial Bureau of SoilScience, Harpenden (1937)

Frequenz

Die Häufigkeit oder absolute Häufigkeit entspricht der Häufigkeit des Auftretens einer bestimmten **Beobachtung** oder eines Ergebnisses in einem **Experiment.**

Die absolute Häufigkeit kann von der relativen Häufigkeit unterschieden werden. Die relative Häufigkeit einer **Beobachtung** ist definiert als das Verhältnis der Anzahl der Vorkommnisse zur Gesamtzahl der Beobachtungen.

Für bestimmte Analysen ist es wünschenswert, die Anzahl der **Beobachtungen** zu kennen, für die die

Der Wert ist beispielsweise „kleiner oder gleich“ oder „höher als“ ein bestimmter Wert. Die Anzahl der Werte, die kleiner oder gleich diesem angegebenen Wert sind, wird als kumulative Häufigkeit bezeichnet.

Der **Prozentsatz** der **Beobachtungen ,** bei denen der **Wert** „kleiner oder gleich“ oder „mehr als“ eines bestimmten Werts ist, kann ebenfalls gewünscht sein. Der Anteil der Werte, die kleiner oder gleich diesem Wert sind

Der gegebene Wert wird als kumulative relative Häufigkeit bezeichnet.

BEISPIELE

Betrachten Sie zum Beispiel die folgenden 16 **Beobachtungen,** die die Körpergröße (in cm) einer Gruppe von Individuen darstellen:

174 169 172 174
171 179 174 176
177 161 174 172
177 168 171 172

Hier beträgt die Häufigkeit der Beobachtung von 171 cm 2; wohingegen die Häufigkeit der Beobachtung von 174 cm 4 beträgt.

In diesem Beispiel beträgt die relative Häufigkeit von $174\frac{4}{16}$ 0,25 (oder 25 % aller **Beobachtungen).** Die absolute Häufigkeit beträgt 4 und die Gesamtzahl der Beobachtungen beträgt 16.

WEITERLESEN

- **Frequenzkurve**
- **Häufigkeitsverteilung**
- **Häufigkeitstabelle**

Frequenzkurve

Wenn ein **Frequenzpolygon** geglättet wird, erhält man eine Kurve, die sogenannte Frequenzkurve. Diese Glättung kann durchgeführt werden, wenn die Anzahl der **Beobachtungen** in der **Häufigkeitsverteilung** unendlich groß und die **Breiten** der Klassen unendlich klein werden.

Die Frequenzkurve entspricht der Grenzform eines **Frequenzpolygons.**

GESCHICHTE

Gegen Ende des neunzehnten Jahrhunderts sind die allgemeine Tendenz war, alles zu berücksichtigen

Machine Translated by Google

Verteilungen als **Normalverteilungen.**

Histogramme, die mehrere **Modi** darstellten , wurden angepasst, um einen Satz normaler Frequenzkurven zu bilden, und diejenigen, die Asymmetrie aufwiesen, wurden analysiert, indem die Daten so transformiert wurden, dass das aus **dieser Transformation** resultierende **Histogramm** mit einer normalen Kurve verglichen werden konnte.

Im Jahr 1895 schlug **Pearson, K.** vor, einen Satz verschiedener theoretischer Frequenzkurven zu erstellen, um bessere Annäherungen an die Histogramme erhalten zu können.

WEITERLESEN

Kontinuierliche Wahrscheinlichkeitsverteilung
Häufigkeitsverteilung
Frequenzpolygon
Normalverteilung

REFERENZEN

Pearson, K.: Beiträge zur mathematischen Evolutionstheorie. II: Skew-Variation in homogenem Material. In: Karl Pearsons frühe statistische Arbeiten.
Cambridge University Press, Cambridge, S. 41–112 (1948). Erstveröffentlichung 1895 in Philos. Trans. Roy. Soc. London.
Ser. A **186,** 343–414

Häufigkeitsverteilung

Die Verteilung einer **Bevölkerung** kann relativ (z. B. in **Prozent),** in Bruchzahlen oder absolut ausgedrückt werden. In all diesen Fällen wird das Ergebnis als **Häufigkeitsverteilung** bezeichnet .

Außerdem erhalten wir entweder eine diskrete oder eine kontinuierliche Häufigkeitsverteilung, je nachdem, ob die untersuchte **Variable** diskret oder kontinuierlich ist.

Die Anzahl des Auftretens eines bestimmten **Werts** oder die Anzahl der Beobachtungen für einen bestimmten Wert Eine bestimmte Klasse wird als **Häufigkeit** dieses Werts oder dieser Klasse bezeichnet.

Die so erhaltene Häufigkeitsverteilung kann entweder als **Häufigkeitstabelle** oder grafisch dargestellt werden, beispielsweise als **Histogramm ,** Häufigkeitspolygon oder **Liniendiagramm.**

Jede Häufigkeitsverteilung hat eine entsprechende relative **Häufigkeitsverteilung ,** die die Verteilung jedes **Wertes** bzw. Wertes ist

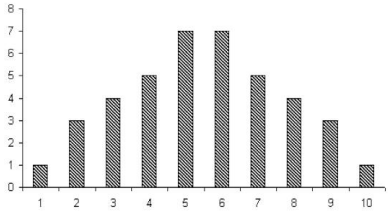
jede Klasse in Bezug auf die Gesamtzahl der **Beobachtungen.**

GESCHICHTE Siehe **grafische Darstellung.**

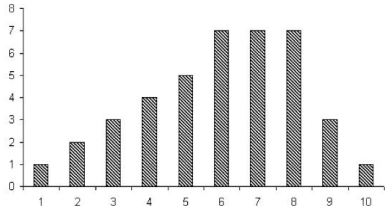
DOMÄNEN UND EINSCHRÄNKUNGEN

In der Praxis können die Häufigkeitsverteilungen erheblich variieren. Zum Beispiel:

- Einige sind symmetrisch; andere sind einfach zu treffen ric:



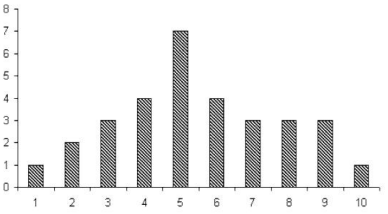
Symmetrische Verteilung



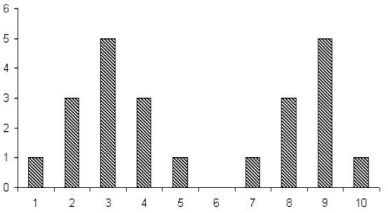
Asymmetrische Verteilung

- Einige haben nur einen **Modus** (unimodale Verteilungen); andere haben mehrere (plurimodale Verteilungen):

Machine Translated by Google



Unimodale Verteilung



Bimodale Verteilung

Dies bedeutet, dass Häufigkeitsverteilungen dies können in vier Haupttypen eingeteilt werden:

1. Unimodale symmetrische Verteilungen;
2. Unimodale asymmetrische Verteilungen;
3. J-förmige Verteilungen;
4. U-förmige Verteilungen.

Bei statistischen Studien ist es oft notwendig, zwei Verteilungen zu vergleichen. Wenn beides Wenn Verteilungen vom gleichen Typ sind, ist es möglich, sie zu vergleichen, indem man ihre Hauptverteilungen untersucht Merkmale, wie das **Maß der zentralen Tendenz,** das **Maß der Streuung** oder das **Maß der Form.**

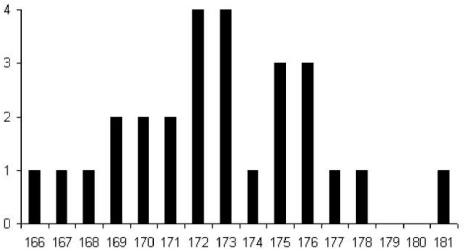
BEISPIELE

Die folgenden **Daten** stellen die Höhen dar, gemessen in Zentimetern, beobachtet für **eine Population** von 27 Schülern einer Mittelschulklasse:

169 177 178 181 173
172 175 171 175 172
173 174 176 170 172
173 172 171 176 176
175 168 167 166 170
173 169

Indem wir diese **Beobachtungen** nach Wert ordnen und neu gruppieren , erhalten wir die folgende **Häufigkeitsverteilung :**

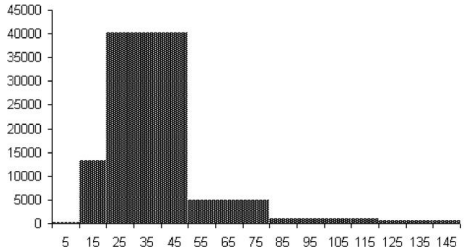
Höhenfrequenz		Höhenfrequenz	
166	1	174	1
167	1	175	3
168	1	176	3
169	2	177	1
170	2	178	1
171	2	179	0
172	4	180	0
173	4	181	1



Hier ist ein weiteres Beispiel, das Anlass zur Sorge gibt die **Häufigkeitsverteilung** einer kontinuierlichen Variable. Nach dem Abrufen und Klassifizieren der persönliche Einkommen der Bevölkerung von Neuchâtel (ein Staat in der Schweiz), für die Zeitraum 1975–1976 erhalten wir Folgendes **Häufigkeitstabelle:**

Nettoumsatz (in Tausend Schweizer Franken)	Frei quenzen	Relativ Frequenz
0–10	283	0,0047
10–20	13175	0,2176
20–50	40316	0,6661
50–80	5055	0,0835
80–120	1029	0,0170
120+	670	0,0111

Machine Translated by Google



WEITERLESEN

- Frequenz
- Frequenzkurve
- Frequenzpolygon
- Häufigkeitstabelle
- Histogramm
- Intervall
- Liniendiagramm

Frequenzpolygon

Ein Häufigkeitspolygon ist eine **grafische Darstellung** einer **Häufigkeitsverteilung,** was zum **Histogramm** der Häufigkeit passt Verteilung. Es handelt sich um eine Art Häufigkeitsdiagramm. Es hat die Form einer segmentierten Linie, die sich verbindet der Mittelpunkt der Oberseite jedes Rechtecks in ein Histogramm.

MATHEMATISCHE ASPEKTE

Wir konstruieren ein Häufigkeitspolygon, indem wir für jede Klasse der **Häufigkeit** einen Punkt zeichnen **Verteilung.** Jeder Punkt entspricht dem Mittelpunkt der Klasse entlang der Abszisse und die **Häufigkeit** der Klasse entlang der Ordinate. Dann verbinden wir jeden Punkt durch Linien mit seinen Nachbarpunkten. Die Linien verbinden daher effektiv den Mittelpunkt der Oberseite jedes einzelnen Rechteck im Histogramm der Häufigkeit Verteilung.

Normalerweise schließen wir das Frequenzpolygon bei die beiden Enden der Verteilung durch Auswahl entsprechende Punkte auf der horizontalen Achse und connectthemtothe Punkte für die erste und letzte Unterrichtsstunden.

DOMÄNEN UND EINSCHRÄNKUNGEN

Ein Häufigkeitspolygon wird normalerweise aus gruppierten Daten erstellt Intervalle. In diesem Fall ist es vorzuziehen Verwenden Sie Intervalle gleicher Breite. Wenn wir passen eine glatte Kurve zu einem Frequenzpolygon, wir Erhalten Sie eine **Frequenzkurve.** Bei diesem Glättungsprozess werden die Daten so neu gruppiert die Klassenbreite ist so gering wie möglich und Anschließend wird die Verteilung neu dargestellt. Natürlich können wir nur diese Art von durchführen Glättung, wenn die Anzahl der Einheiten in der Bevölkerung groß genug ist, um einen signifikanten Wert zu ergeben Anzahl der Beobachtungen in jeder Klasse danach Umgruppierung.

BEISPIELE

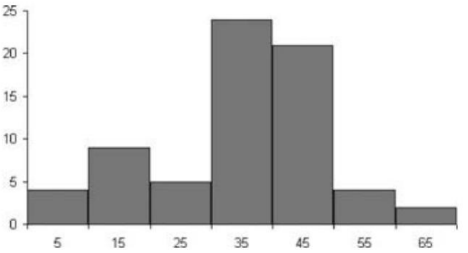
Die folgende **Häufigkeitstabelle** gibt die jährliche Häufigkeit an mittlere Niederschläge für 69 Städte in den USA:

Jährlicher mittlerer Niederschlag (in Zoll)	Frequenz
0–10	4
10–20	9
20–30	5
30–40	24
40–50	21
50–60	4
60–70	2
Gesamt	69

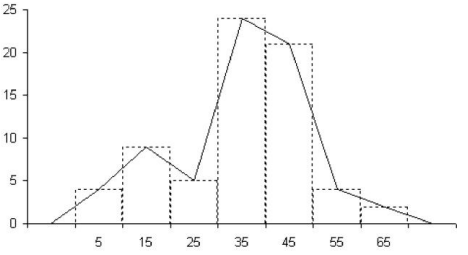
Das Frequenzpolygon wird über konstruiert **Histogramm** der Häufigkeitsverteilung:

F

Machine Translated by Google



Um das Häufigkeitspolygon zu konstruieren, zeichnen wir es auf einen Punkt für jede Klasse. Die Position dazu Der Punkt entlang der Abszisse ist durch den Mittelpunkt der Klasse und seine Position entlang der Abszisse gegeben Die Ordinate ist durch die Häufigkeit gegeben Klasse. Mit anderen Worten, jeder Punkt tritt bei auf der Mittelpunkt der Oberseite des Rechtecks für die Klasse im Histogramm. Jeder Punkt ist dann durch Linien mit seinen nächsten Nachbarn verbunden. In Um das Polygon zu schließen, wird ein Punkt platziert auf der Abszisse eine Klassenbreite vor dem Mittelpunkt der ersten Klasse und ein weiterer Punkt wird eine Klassenbreite auf der Abszisse platziert nach der Mitte der letzten Unterrichtsstunde. Diese Punkte werden dann durch Linien mit den Punkten verbunden jeweils für die erste und letzte Klasse. Der Das Frequenzpolygon, das wir aus diesem Verfahren erhalten, ist wie folgt:



WEITERLESEN

- **Häufigkeitsverteilung**
- **Häufigkeitstabelle**
- **Grafische Darstellung**
- **Histogramm**
- **Ogive**

Häufigkeitstabelle

Die Häufigkeitstabelle ist ein Hilfsmittel zur Darstellung einer **Häufigkeitsverteilung.** Es bietet die Fähigkeit, statistische Beobachtungen darzustellen.

Die **Häufigkeit** eines **Wertes** einer **Variablen** ist die Häufigkeit, mit der dieser Wert vorkommt eine Bevölkerung. Eine **Häufigkeitsverteilung** kann dann als Liste der Frequenzen definiert werden erhalten für die verschiedenen Werte, die von der angenommen wurden **Variable.**

MATHEMATISCHE ASPEKTE

Betrachten Sie eine Menge von *n* Einheiten, die durch eine **Variable** *X* beschrieben werden und *k* Werte x1, *x2,...,xk* annehmen können . Letni sei die Anzahl der Einheiten mit dem Wert *xi; ni* ist dann die **Häufigkeit** des Wertes *xi.* Die relative Häufigkeit von *xi* beträgt $fi = \frac{ni}{N}$. Da die Werte unterschiedlich und erschöpfend sind, die Summe aller Frequenzen *ni* ist gleich Gesamtzahl *n* der Einheiten in der Menge oder die Summe der relativen Häufigkeitenfi gleich Eins; In andere Worte:

$$\sum_{i=1}^{k} ni = n \text{ und } \sum_{i=1}^{k} fi = 1 \cdot$$

Die entsprechende Häufigkeitstabelle dazu Szenario ist wie folgt:

Werte der Variable X	Frequenzen	relativ Frequenzen
x1	n1	f1
x2	n2	f2
...	...	...
xk	*nk*	*fk*
Gesamt	*N*	1

Manchmal ist das Konzept der kumulativen Häufigkeit wichtig (siehe zweites Beispiel).

Machine Translated by Google

zum Beispiel). Dies ist die Summe aller Häufigkeiten

Städte bis einschließlich der Frequenz *fi : fj*.

j=1

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Erstellung einer Häufigkeitstabelle ist der einfachste und am häufigsten verwendete Ansatz zur Darstellung von Daten. Es gibt jedoch welche einige Regeln, die man beachten sollte, wenn So eine Tabelle erstellen:

- Die Tabelle muss umfassend sein und prägnanter Titel, der die Einheiten nennt beschäftigt;
- Die Namen der Zeilen und Spalten müssen angegeben werden seien Sie präzise und kurz;
- Die Spaltensummen sollten angegeben werden;
- Die Quelle der in der Tabelle angegebenen Daten muss angegeben werden.

Generell muss die Tabelle so verständlich sein, dass man sie ohne Notwendigkeit verstehen kann Lesen Sie den Begleittext.

Die Anzahl der zur Darstellung ausgewählten Klassen Die Daten hängen von der Größe der untersuchten Menge ab. Eine große Anzahl von Klassen für eine kleine verwenden Einstellung führt zu unregelmäßigen Frequenzen die geringe Anzahl von Einheiten pro Klasse. Auf der Andererseits wird eine kleine Anzahl von Klassen verwendet Eine große Menge führt zum Verlust von Informationen über die Struktur der Daten.

Es gibt jedoch keine allgemeine Regel zur Bestimmung der Anzahl der Klassen, die es geben sollte Wird zur Erstellung einer Häufigkeitstabelle verwendet. Für relativ kleine Mengen *(n* ÿ 200) zwischen 7 und Empfohlen werden jedoch 15 Unterrichtsstunden Regel ist nicht absolut.

Der Einfachheit halber werden häufig Klassen von verwendet die gleiche Breite und die Klassenintervalle sollten nicht überlappen. In bestimmten Fällen, zum Beispiel wenn Wir haben eine Variable, die einen großen Bereich annehmen kann von Werten, sondern ein bestimmtes Werteintervall voraussichtlich weitaus häufiger als der Rest, Es kann nützlich sein, eine offene Klasse zu verwenden, z „1000 oder mehr.“

BEISPIELE

Häufigkeitsverteilung australischer Einwohner (in Tausend) nach ihrem Familienstand am *30.* Juni 1981

in der folgenden Tabelle angegeben.

Ehe Status	Frequenz (in Tausenden)	Relativ Frequenz
Einzel	6587	0,452
Verheiratet	6837	0,469
Geschieden	403	0,028
Verwitwet	749	0,051
Gesamt	14576	1.000

Quelle: ABS (1984) Pocket Year Book Australia. Australisches Statistikamt, Canberra, P. 11.

Die Daten werden in Form einer Häufigkeitstabelle dargestellt. Die Variable „Ehestand“ kann vier verschiedene Werte annehmen: ledig, verheiratet, geschieden oder verwitwet. Die **Frequenz,** Der Wert wird in Tausend ausgedrückt und die relative Häufigkeit wird für jeden Wert angegeben.

Im folgenden Beispiel betrachten wir die Fall, in dem Beobachtungen gruppiert werden Klassen. Diese Tabelle stellt die Klassifizierung von 3014 Personen basierend auf ihrem Einkommen dar im Jahr 1955.

Einkommen	Frequenz ni	Relativ	Relativ Kumulativ
Weniger als 1000	261	0,0866	0,0866
1000–1999	331	0,1098	0,1964
2000–2999	359	0,1191	0,3155
3000–3999	384	0,1274	0,4429
4000–4999	407	0,1350	0,5780

Machine Translated by Google

Einkommen	Frequenz		
	ni	Relativ	Relativ Kumulativ
5000–7499	703	0,2332	0,8112
7500–9999	277	0,0919	0,9031
10000 oder mehr	292	0,0969	1,0000
Gesamt	3014	1,0000	

Aus dieser Tabelle können wir ersehen, dass 57,8 % dieser 3014 Personen ein Einkommen haben, das unter 5000 Dollar liegt.

WEITERLESEN

- Frequenz
- Häufigkeitsverteilung

Machine Translated by Google

Galton, Francis

Galton, Francis wurde 1822 in der Nähe von Birmingham, England, als Sohn einer Familie von Intellektuellen geboren: Darwin, Charles war sein Cousin, und seine Familie Großvater war Mitglied der Royal Society.

Sein Interesse an der Wissenschaft manifestierte sich erstmals in den Bereichen Geographie und Meteorologie. Als Mitglied der Royal Society gewählt im Jahr 1860 begann er sich dafür zu interessieren Genetik und statistische Methoden im Jahr 1864. Galton war eng mit Pearson, K. befreundet . Tatsächlich kam ihm Pearson finanziell zu Hilfe als Galton die Zeitschrift „Biometri ka“ gründete. Galtons „Eugenics Record Office“ fusionierte mit Pearsons **Biometrielabor** am University College London und wurde bekannt als „Galton-Labor“. Er starb 1911 und hinterließ mehr als 300 Veröffentlichungen, darunter vor allem 17 Bücher zu statistischen Methoden im Zusammenhang mit **der Regression Analyse** und das Konzept der **Korrelation,** beides wird ihm zugeschrieben.

Einige der Hauptwerke und Artikel von Galton, F.:

1869 Erbliches Genie: Eine Untersuchung über seine Gesetze und Konsequenzen.Macmillan, London.

1877 Typische Gesetze der Vererbung. Natur, 15, 492–495, 512–514, 532–533.

1889 Natürliche Vererbung. Macmillan, London.

1907 Wahrscheinlichkeit, die Grundlage der Eugenik. Henry Froude, London.

1908 Erinnerungen an mein Leben. Methuen, London.

1914–1930 Pearson, K. (Hrsg.) The Life, Letters and Labors of Francis Galton. Cambridge University Press, Cam Bridge

WEITERLESEN

Korrelationskoeffizient

Regressionsanalyse

Gammaverteilung

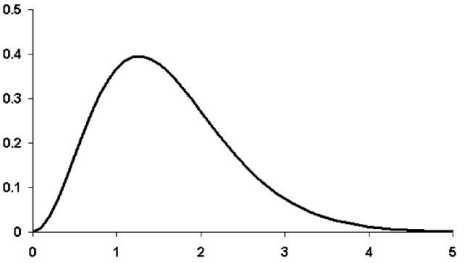
Eine **Zufallsvariable** *X* folgt einem Gamma Verteilung mit **Parameter** ÿ wenn seine **Dichte Funktion** ist gegeben durch

$$f(x) = \frac{ÿÿ}{} \text{xÿÿ1} \cdot \text{eÿÿx} \, (ÿ) ,$$

wenn $x > 0$, $ÿ > 0$, $ÿ > 0$,

wobei (ÿ) die Gammafunktion ist.

Machine Translated by Google



Gammaverteilung, ÿ = 2, ÿ = 1

Die Standardform der Gammaverteilung erhält man durch Setzen von ÿ = 1, was ergibt

$$f(x) = \frac{1}{(ÿ)} xÿÿ1 \cdot eÿx \quad ,$$

wenn $x > 0$, $ÿ > 0$.

Die Gammafunktion () erscheint häufig in der statistischen Theorie. Es ist definiert durch:

$$(ÿ) = \int_0^{ÿ} rÿÿ1 eÿt \, dt .$$

In **der Statistik** interessieren uns nur **Werte** von ÿ > 0.

Durch partielle Integration erhalten wir:

$$(ÿ + 1) = ÿ(ÿ) .$$

Seit

$$(1) = 1 \quad ,$$

wir haben

(ÿ+1) = ÿ! für jede positive ganze Zahl *ÿ*.

GESCHICHTE

Laut Lancaster (1966) wurde diese **kontinuierliche Wahrscheinlichkeitsverteilung** von **Laplace, PS** (1836) entwickelt .

MATHEMATISCHE ASPEKTE

Der **Erwartungswert** der Gammaverteilung ergibt sich aus:

$$E[X] = ÿ \quad \frac{1}{—} \cdot \frac{(ÿ + 1)}{(ÿ)} = \frac{ÿ}{ÿ} .$$

Seit

$$E\,X2 = ÿ2 \quad \frac{1}{—} \cdot \frac{(ÿ + 2)}{(ÿ)} = \frac{ÿ(ÿ + 1)}{ÿ2} ,$$

die **Varianz** ist gleich:

$$\begin{aligned} \mathrm{Var}(X) &= E\,X2 ÿ (E[X])^2 \\ &= \frac{ÿ(ÿ + 1)\, ÿ\, ÿ2}{ÿ2} \\ &= \frac{ÿ}{ÿ2} . \end{aligned}$$

Eine Besonderheit ist die **Chi-Quadrat-Verteilung** Fall der Gammaverteilung mit ÿ = und ÿ = $\frac{N}{2}$ $\frac{1}{2}$, und *n* ist die **Gradzahl der Freiheit** der Chi-Quadrat-Verteilung.

Die Gammaverteilung mit **Parameter** ÿ = 1 ergibt die **Exponentialverteilung:**

$$f(x) = ÿ \cdot eÿÿx \quad .$$

WEITERLESEN

- **Chi-Quadrat-Verteilung**
- **Kontinuierliche Wahrscheinlichkeitsverteilung**
- **Exponentialverteilung**

VERWEISE

Lancaster, HO: Vorläufer des Pear-Son-Chi-Quadrats. Aust. J. Stat. 8, 117–126 (1966)

Laplace, PS de: Théorie analytique des Wahrscheinlichkeiten, suppl. bis 3. Aufl. Kurier, Paris (1836)

Gauß, Carl Friedrich

Gauß, Carl Friedrich wurde 1777 geboren Braunschweig in Deutschland. Er wurde schnell ein renommierter Astronom und Mathematiker und gilt immer noch als gleichwertig mit Archimedes und Newton als einer der größten Mathematiker aller Zeiten.

Machine Translated by Google

Er promovierte 1799 und arbeitete anschließend an der Universität Helmsted. 1807 zog er nach Göttingen, wo er Laboratoriumsdirektor wurde. Den Rest seines Lebens verbrachte er in Göttingen, wo er 1855 starb.

Seine Beiträge zur Wissenschaft, insbesondere zur Physik, sind von großer Bedeutung. In **der Statistik** befassten sich seine Arbeiten mit der Schätztheorie , und die Methode der **kleinsten Quadrate** sowie die Anwendung der **Normalverteilung** auf Probleme im Zusammenhang mit Messfehlern **stammen** beide von ihm.

Einige der Hauptwerke und Artikel von Gauss, CF:

1803–1809 Disquisitiones de elementis ellipticis pallidis. Werke, 6, 1–24.

1809 Theoria motus corporum coelestium. Werke, 7. (1963 englische Übersetzung von Davis, CH, veröffentlicht von Dover, New York).

1816 Bestimmung der Genauigkeit der Beobachtungen. Werke, 4, 109–117. **1821, 1823 und 1826** Theoria combinatio nis Observationum Erroribus Minimis obnoxiae, Teile 1, 2 und suppl.Werke, 4, 1–108.

1823 Anwendungen der Wahrschein lichkeitsrechnung auf eine Aufgabe der praktischen Geometrie. Werke, 9, 231–237.

1855 Méthode des Moindres Carrés: Erinnerungen an die Kombination des Beobachtungen. Französische Übersetzung von die Arbeit von Gauss, CF von Bertrand, J. (autorisiert von Gauss, CF selbst). Mallet-Bachelier, Paris.

1957 Gauß' Werk (1803–1826). Zur Theorie der kleinsten Quadrate. Englische Übersetzung von Trotter, HF Technical Bericht Nr. 5, Statistical Techniques Research Group, Princeton, NJ.

WEITERLESEN

De Moivre, Abraham
Gauß-Markov-Theorem
Normalverteilung

Gauß-Markov-Theorem

Das Gauß-Markov-Theorem postuliert, dass, wenn die Fehlerwahrscheinlichkeitsverteilung **in** einem linearen **Modell unbekannt ist,** unter allen linearen erwartungstreuen Schätzern für die Parameter des linearen **Modells** ist der **Schätzer,** der mit der Methode der **kleinsten Quadrate** ermittelt wird , derjenige, der die **Varianz minimiert.** Es wird davon ausgegangen, dass der **mathematische Erwartungswert** jedes Fehlers Null ist und alle dieselbe (unbekannte) Varianz aufweisen.

GESCHICHTE

Einen Beweis dieses Theorems lieferte **Carl Friedrich** im ersten Teil seines Werkes „Theoria combinationis Beobachtung“. um erroribus minimis obnoxiae“ (1821). Markov, Andrei Andreyevich entdeckte diesen Satz im Jahr 1900 wieder. Eine Version des Gauß-Markov-Theorems in moderner Notation geschrieben, wurde 1976 von Graybill zur Verfügung gestellt.

MATHEMATISCHE ASPEKTE

Betrachten Sie das lineare **Modell**

$$Y = X \cdot \ddot{y} + \ddot{y} \quad ,$$

Wo

Y ist der $n \times 1$ **-Vektor** der Beobachtungen,
X ist die $n \times p$ **-Matrix** der unabhängigen Variablen, die als fest gelten,

Machine Translated by Google

ÿ ist der p×1-Vektor der unbekannten Parameter und **ÿ** ist der *n* ×1-Vektor der Zufallsfehler.

Wenn die **Fehlerwahrscheinlichkeitsverteilung** unbekannt ist, aber die folgenden Bedingungen erfüllt sind:

1. Der **mathematische Erwartungswert** E[ÿ] = 0,
2. Die **Varianz** Var(ÿ) = ÿ2 ·**In,** wobei **In** ist die **Identitätsmatrix,** 3.

Die Matrix *X* hat einen vollen Rang, dann ist der **Schätzer** von ÿ,

$$\hat{\ddot{y}} = XX \quad \ddot{y}1 \quad XY \quad ,$$

Der über die Methode der **kleinsten Quadrate** abgeleitete Wert ist der lineare Schätzer ohne **Bias** von ÿ, der die geringste Varianz aufweist.

WEITERLESEN

Schätzer

Kleinsten Quadrate

VERWEISE

Gauss, CF: Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, Teile 1, 2 und suppl. Werke 4, 1–108 (1821, 1823, 1826)

Graybill, FA: Theorie und Anwendungen des linearen Modells. Duxbury, North Carolina, MA (Waldsworth und Brooks/Cole, Pacific Grove, CA) (1976)

Rao, CR: Lineare statistische Inferenz und ihre Anwendungen, 2. Aufl. Wiley, New York (1973)

Verallgemeinerte Umkehrung

Die verallgemeinerte Umkehrung ist analog zur Umkehrung einer nichtsingulären **quadratischen Matrix, wird jedoch** für eine Matrix beliebiger Dimension und Rang verwendet. Die verallgemeinerte Umkehrung wird bei der Auflösung linearer Gleichungssysteme verwendet.

GESCHICHTE

Es scheint, dass Fredholm (1903) der erste war, der das Konzept einer verallgemeinerten Umkehrung in Betracht zog. Moore definierte in seinem nach seinem Tod veröffentlichten Buch *General Analysis (1935)* eine einzelne verallgemeinerte Umkehrung . Allerdings wurde seine Arbeit erst in den 1950er Jahren genutzt, als sie aufgrund der Anwendung der verallgemeinerten Umkehrung auf Probleme im Zusammenhang mit der Methode der **kleinsten Quadrate einen Anstieg des Interesses erfuhr.** Im Jahr 1955 zeigte Penrose unter Wiederverwendung und Erweiterung einer 1951 von Bjerhammar veröffentlichten Arbeit, dass die verallgemeinerte Mooresche Umkehrung eine **eindeutige Matrix G** ist , die die folgenden vier Gleichungen erfüllt: 1. **A** = **A** · **G** · A; 2. **G** = **G** · **A** · G; 3. **(A** · G) = **A** · G; 4. **(G** · A) = **G** · A.

Diese einzigartige verallgemeinerte Umkehrung ist als Moore-Penrose-Umkehrung bekannt und wird mit A+ bezeichnet.

MATHEMATISCHE ASPEKTE

Dann ist×mmatrixG die verallgemeinerte Umkehrung der *m* × *n*- Matrix **A** if

$$\mathbf{A} = \mathbf{A} \cdot \mathbf{G} \cdot \mathbf{A} \quad .$$

Die Matrix **G** ist genau dann eindeutig, wenn *m* = *n* und **A** nicht singulär ist.

Die am häufigsten verwendete Schreibweise für die verallgemeinerte Umkehrung einer **Matrix A** ist Aÿ.

WEITERLESEN

Inverse Matrix

Kleinsten Quadrate

VERWEISE

Bjerhammar, A.: Rechteckige reziproke Matrizen mit besonderem Bezug auf Geodäten

Machine Translated by Google

Tic-Berechnungen. Stier. Geod. **20,** 188–220 (1951)

Dodge, Y., Majumdar, D.: Ein Algorithmus zum Finden kleinster verallgemeinerter Quadratinversen für Klassifizierungsmodelle mit beliebigen Mustern. J. Stat. Berechnen. Simul. 9, 1–17 (1979)

Fredholm, J.: Über eine Klasse funktionierender Gleichungen. Acta Math. **27,** 365–390 (1903)

Moore, EH: Allgemeine Analyse, Teil I. Mem. Bin. Philos. Soc., Philadelphia, PA, S. 147–209 (1935)

Penrose, R.: Eine verallgemeinerte Inverse für Matrizen. Proz. Camb. Philos. Soc. **51,** 406–413 (1955)

Rao, CR: Eine Anmerkung zur verallgemeinerten Umkehrung einer Matrix mit Anwendungen auf Probleme in der mathematischen Statistik. J. Roy. Stat. Soc. Ser. B **24,** 152–158 (1962)

Rao,CR:Kalkül der verallgemeinerten Inversen von Matrizen. I. Allgemeine Theorie. Sankhya **A29,** 317–342 (1967)

Rao, CR, Mitra, SK: Generalisierte Umkehrungen von Matrizen und ihren Anwendungen. Wiley, New York (1971)

Verallgemeinerte lineare Regression

Eine Erweiterung des linearen Regressionsmodells auf flexiblere Einstellungen hinsichtlich der zugrunde liegenden Annahmen, die die lineare Regression erfordert. Anstatt beispielsweise davon auszugehen, dass die Fehler gleiche Varianzen aufweisen sollten, könnten wir die folgenden Formen annehmen: • *Heteroskedastizität: Die* im Modell beobachteten Fehler (auch Residuen genannt) können unterschiedliche Varianzen zwischen den Beobachtungen aufweisen tionen (oder zwischen den verschiedenen Gruppen von Beobachtungen);

• *Autokorrelation:* Es kann eine **Korrelation** geben **Beziehung** zwischen den Fehlern in der Differenz HNO-Beobachtungen.

MATHEMATISCHE ASPEKTE

Wir betrachten ein allgemeines Modell **der multiplen linearen Regression:**

$$Yi = \ddot{y}0 + \sum_{j=1}^{p\ddot{y}1} \ddot{y}jXij + \ddot{y}i\ , \qquad i = 1,...,n,$$

Wo

Yi ist die **abhängige Variable, *Xji,***

j = 1,...,p ÿ 1 sind die unabhängigen Variablen, ***ÿj,***

j = 0,...,p ÿ

1 sind die zu schätzenden Parameter und ***ÿi*** ist die zufällige nicht

beobachtbare Größe **Fehler**

Begriff.

In Matrixform schreiben wir:

$$\mathbf{Y} = \mathbf{X\ddot{y}} + \ddot{\mathbf{y}} \quad ,$$

Wo:

Y ist der Vektor *(n* × 1) der Beobachtungen, die sich auf die **abhängige Variable** *(n* Beobachtungen)

beziehen, **ÿ** ist der Vektor (p×1) der zu schätzenden Parameter

ÿ ist der Vektor *(n* × 1) der Fehler,

Und

$$\mathbf{X} = \begin{matrix} 1 & X11 & \dots & X1(p\ddot{y}1) \\ \ddot{y} & \vdots & \vdots & \vdots \\ \ddot{y} & \vdots & \vdots & \vdots \\ \ddot{y} & 1 & Xn1 & \dots & Xn(p\ddot{y}1) \end{matrix} \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix}$$

Machine Translated by Google

ist die **Matrix** *(n* × p), die sich auf die unabhängigen Variablen bezieht.
In einem gewöhnlichen Regressionsmodell lauten die für die Fehler *ÿi* verwendeten Hypothesen im Allgemeinen wie folgt: • Jede **Erwartung** ist gleich Null (mit anderen Worten, die Beobachtungen sind nicht verzerrt), • Sie sind unabhängig und nicht korreliert verspätet,

• ihre **Varianz** V(ÿi) ist konstant und gleich ÿ2 (wir nehmen an, dass die Fehler gleich groß sind).

Wenn das gewählte Modell geeignet ist, muss die Verteilung der Residuen dies bestätigen Hypothesen (siehe **Varianzanalyse).**
Wenn wir feststellen, dass dies nicht der Fall ist, oder wenn wir zusätzliche Informationen erhalten, die darauf hinweisen, dass die Fehler korrelieren, können wir die verallgemeinerte lineare Regression verwenden, um eine genauere Schätzung für die Parameter ÿ0,..., ÿpÿ1 zu *erhalten* . Anschließend führen wir die Varianz-Kovarianz- **Matrix V(ÿ)** von (der Dimension *n* × n) der Fehler ein, die durch die folgende Gleichung für Zeile *i* und Spalte j definiert ist: •
V(ÿi) , wenn *i* = *j* (Varianz in ÿi), • Cov *ÿi, ÿj* **(Kovarianz** zwischen *ÿi* und *ÿj),* wenn *i* = j.

Oft muss diese Matrix geschätzt werden, aber wir betrachten zunächst den einfachsten Fall, bei dem **V (ÿ)** bekannt ist. Wir verwenden den üblichen Begriff für Varianz, ÿ2 : **V(ÿ)** = **ÿ2V.**

Schätzung des Vektors ÿ, wenn V bekannt ist

Durch die Umwandlung unseres allgemeinen Modells in ein gewöhnliches Regressionsmodell ist es möglich, die Parameter des verallgemeinerten Modells abzuschätzen, indem wir Folgendes verwenden:

$$ÿ\hat{} = \mathbf{X}\ V ÿ1 X \overset{ÿ1}{} \mathbf{X}\ V ÿ1 Y .$$

Wir können Folgendes beweisen: 1. Die Schätzung ist nicht verzerrt; das ist:

E(ÿ^) = ÿ.

2. Folgende Gleichheit wird überprüft:

$$\mathbf{V}\ ÿ\hat{} = \hat{}ÿ2\ \mathbf{X}\ V ÿ1 X \overset{ÿ1}{} ,$$

mit $ÿ\hat{}^{2} = \frac{1}{n ÿ p}\ \mathbf{ÿ}\ V ÿ1 ÿ.$

3. ÿ^ ist der beste (d. h. mit der geringsten Varianz) erwartungstreue Schätzer **von ÿ;** Dieses Ergebnis ist als verallgemeinertes Gauß-Markov-Theorem bekannt.

Schätzung des Vektors ÿ, wenn V unbekannt ist

Wenn die Varianz-Kovarianz-Matrix nicht vorhanden ist Bekanntlich wird das verwendete Modell als generalisierte Regression bezeichnet. Der erste Schritt besteht darin, **V** als Funktion eines Parameters ÿ auszudrücken , wodurch wir **V** mithilfe von V^ = V(ÿ)^ schätzen können Die Schätzungen für die Parameter erhält man dann, indem man in der oben angegebenen Formel **V** durch V^ ersetzt:

$$ÿ\hat{} = \mathbf{(X}\ V\hat{}\ \mathbf{ÿ1X)}\ ÿ1X\ V\hat{}\ \mathbf{ÿ1Y} \quad .$$

Transformation von V in drei typischen Fällen

a) Heteroskedastizität

Wir gehen davon aus, dass die Beobachtungen unterschiedlich sind ance $\frac{ÿ2}{wi}$, Dabei ist *wi* das Gewicht, das für jedes *i = 1,...,n unterschiedlich sein kann.* So:

$$\mathbf{V} = ÿ2 \begin{matrix} ÿ & \frac{1}{w1} & 0 \ldots 0 & & ÿ \\ ÿ & 0 & \frac{}{1\ w2} \ldots 0 & & ÿ \\ " & \vdots & \vdots \quad \ldots \quad \vdots & & " \\ ÿ & \cdot & \cdot \quad \ldots \quad \cdot & & ÿ \\ ÿ & 0\ 0 \ldots & & \frac{}{1\ Jahr} & ÿ \end{matrix} .$$

Indem

$$Yw = \mathbf{WY}$$

$$Xw = \mathbf{WX}$$

$$ÿw = \mathbf{Wÿ} \quad ,$$

Machine Translated by Google

wobei **W** die folgende quadratische Matrix der Ordnung n ist:

$$\mathbf{W} = \begin{pmatrix} \frac{1}{\ddot{y}w1} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\ddot{y}w2} & \dots & 0 & \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \dots & \frac{1}{\ddot{y}wn} & \end{pmatrix},$$

so dassW **W** = Vÿ1, erhalten wir das Äquiv Alent-Modell:

$$\mathbf{V}(\ddot{y}\mathbf{w}) = \mathbf{V}(\mathbf{W}\ddot{y}) = \mathbf{W}\mathbf{V}(\ddot{y})\mathbf{W} = \ddot{y}2\mathbf{WVW} = \ddot{y}2\mathbf{In},$$

wobei *In* die Identitätsmatrix der Dimension n ist. Da die Varianzen aller Fehler

in diesem neuen Modell gleich sind, können wir die Methode der kleinsten Quadrate anwenden und erhalten den Vektor der Schätzer:

$$\begin{aligned} \ddot{y}\hat{}_{w} &= (\mathbf{X}\ wXw)\ \ddot{y}1\mathrm{X}\ \mathbf{wYw} \\ &= (\mathbf{XW}\ \mathbf{WX})\ \ddot{y}1\mathrm{X}\ \mathbf{W}\ \mathbf{WY} \\ &= (\mathbf{X}\ \mathrm{V}\ddot{y}1\mathrm{X})\ \ddot{y}1\mathrm{X}\ \mathrm{V}\ddot{y}1\mathrm{Y}\ , \end{aligned}$$

also der Vektor der Schätzwerte für **Y** = Wÿ1Yw ist:

$$\begin{aligned} \mathrm{Y}\hat{} &= \mathrm{W}\ddot{y}1\mathrm{Y}\hat{}_{\ w} \\ &= \mathbf{W}\ddot{y}1\mathbf{WX}(\mathbf{X}\ \mathrm{V}\ddot{y}1\mathrm{X})\ \ddot{y}1\mathrm{X}\ \mathrm{V}\ddot{y}1\mathrm{Y} = \mathbf{X}\ddot{y}\hat{}_{w}\ . \end{aligned}$$

b) Heteroskedastizität nach Gruppen

Wir nehmen an, dass die Beobachtungen in drei Gruppen (der Größe n1, n2, n3) mit den jeweiligen Varianzen ÿ2 ÿ2 und ÿ2 3fallen . So:

$$\mathbf{V} = \begin{pmatrix} \ddot{y}2_1\ In\ 1 & \mathrm{Ein2\times n1} & \mathrm{Ein3\times n1} \\ \mathrm{On1\times n2} & \ddot{y}2_2\ In2 & \mathrm{On3\times n2} \\ \mathrm{On1\times n3} & \mathrm{On2\times n3} & \ddot{y}2_3\ In3 \end{pmatrix},$$

Dabei ist *Inj* die **Identitätsmatrix** der Ordnung *nj* und Oni×nj die Nullmatrix der Ordnung *ni ×nj.*

Die Varianzen ÿ2 ÿ2 ÿ2kann geschätzt werden

Dies wird durch die Durchführung gewöhnlicher Regressionen für die verschiedenen Gruppen von Beobachtungen und durch die Schätzung von ÿ2 mithilfe von S2 ermittelt (siehe **multiple lineare Regression).**

c) Autokorrelation erster Ordnung

Wir gehen davon aus, dass zwischen ihnen ein Zusammenhang besteht

die Fehler in verschiedenen Beobachtungen. Insbesondere gehen wir von einer Autokorrelation erster Ordnung aus; Das heißt, der Fehler *ÿi* für die Beobachtung *i* hängt vom Fehler ÿiÿ1 für die vorherige Beobachtung Wir haben dann:

$$\ddot{y}i = \ddot{y}\ddot{y}i\ddot{y}1 + ui\ , \qquad i = 2,\dots,n,$$

Dabei werden die üblichen Hypothesen für die Fehler eines Regressionsmodells auf die *Benutzeroberfläche angewendet.*

In diesem Fall kann die Varianz-Kovarianz-Matrix der Fehler ausgedrückt werden durch:

$$\mathbf{V}(\ddot{y}) = \ddot{y}2\mathrm{V} = \ddot{y}2 \frac{1}{1\ \ddot{y}\ \ddot{y}2} \begin{pmatrix} 1 & \ddot{y} & \ddot{y}2 \dots \ddot{y}n\ddot{y}1 & \ddot{y} \\ \ddot{y} & 1 & \ddot{y} \dots \ddot{y}n\ddot{y}2 & \ddot{y} \\ \vdots & \ddot{y} & \dots \dots \vdots & \vdots \\ \vdots & \vdots & \dots \dots \ddot{y} & \\ \ddot{y} & \ddot{y}n\ddot{y}1 & \ddot{y}n\ddot{y}2 \dots \ddot{y}\ 1 & \ddot{y} \end{pmatrix}.$$

Wir können V schätzen, indem wir die folgende Schätzung von ÿ in der Matrix V ersetzen:

$$\ddot{y}\hat{} = \frac{\sum_{i=2}^{N} ei\,ei\ddot{y}1}{\sum_{i=2}^{N} i\ddot{y}1} \text{(empirische Korrelation)}\ , \quad e2$$

wobei *ei* der geschätzte Fehler (Residuen) mit einer gewöhnlichen Regression ist.

Wir können die Autokorrelation erster Ordnung testen durch Durchführen eines Durbin-Watson-Tests an den Residuen (wie im Eintrag zur **seriellen Korrelation beschrieben).**

Machine Translated by Google

BEISPIELE

Korrektur für Heteroskedastizität, wenn die Daten stellen Mittelwerte dar

Um die Wirksamkeit eines neuen Nahrungsergänzungsmittels auf das Wachstum von Hühnern zu testen, wurden 40 Hühner in fünf Gruppen (der Größe nach) eingeteilt n1, *n2,...,n5)* und die verschiedenen Dosen gegeben der Ergänzung. Die erhaltenen Wachstumsdaten sind in der folgenden Tabelle zusammengefasst:

Gruppe	Anzahl Hühner	ni	Bedeuten Gewicht Yi	Bedeuten Dosis Xi
1	12		1.7	5.8
2	8		1.9	6.4
3	6		1.2	4.8
4	9		2.3	6.9
5	5		1.8	6.2

Da die Variablen Mittelwerte darstellen, ist die Variable ance des Fehlers *ÿi* ist $\frac{ÿ2}{ni}$.
Die Varianz-Kovarianz-Matrix der Fehler lautet Dann:

$$\mathbf{V(ÿ)} = ÿ2 \begin{bmatrix} \frac{1}{12} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} \end{bmatrix} = ÿ2V .$$

Die **inverse Matrix** Vÿ1 ist dann:

$$Vÿ1 = \begin{bmatrix} 12 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix},$$

unddie Schätzungen für die Parameterÿ0 und ÿ1 im Modell

$$Yi = ÿ0 + ÿ1Xi + ÿi$$

Sind:

$$ÿ\hat{} 0 = ÿ1,23$$

$$ÿ\hat{} 1 = 0,502 .$$

Dieses Ergebnis ist das gleiche wie das, das mit erhalten wurde die **gewichtete Methode der kleinsten Quadrate.**

WEITERLESEN

- **Varianzanalyse**
- **Gauß-Markov-Theorem**
- **Modell**
- **Multiple lineare Regression**
- **Nichtlineare Regression**
- **Regressionsanalyse**

VERWEISE

Bishop, YMM, Fienberg, SE, Holland, PW: Diskrete multivariate Analyse: Theorie und Praxis. MIT Press, Cambridge, MA (1975)

Graybill, FA: Theorie und Anwendungen von das lineare Modell. Duxbury, North Scit uate, MA (Waldsworth und Brooks/Cole, Pacific Grove, CA) (1976)

Rao, CR: Lineare statistische Inferenz und Seine Anwendungen, 2. Aufl. Wiley, Neu York (1973)

Generation von Zufallszahlen

Bei der Generierung **von Zufallszahlen** wird eine Reihe von Zahlen erzeugt, bei denen keine erkennbaren Muster oder Regelmäßigkeiten vorliegen, d. h. sie erscheinen willkürlich. Diese Zufallszahlen sind oft dann verwendet, um **Simulationen** durchzuführen und zu Lösen Sie Probleme, die analytisch nur schwer oder gar nicht lösbar sind, mithilfe des **Monte Carlo-Methode.**

Machine Translated by Google

GESCHICHTE

Die ersten Arbeiten zur Erzeugung von Zufallszahlen stammen von von Neumann, John (1951), der die Methode der „mittleren vier Quadrate“ einführte. Lehmer (1951) führte die einfache Kongruenzmethode ein, die auch heute noch verwendet wird.

Sowey veröffentlichte drei Artikel mit einer klassifizierten Bibliographie zur Generierung und Prüfung von Zufallszahlen (Sowey, ER 1972, 1978 und 1986).

Im Jahr 1996 schlug Dodge, Yadolah eine natürliche Methode zur Generierung von Zufallszahlen aus den Nachkommastellen in ÿ vor.
Einige Dezimalzahlen von ÿ sind in Anhang A aufgeführt.

MATHEMATISCHE ASPEKTE

Es gibt verschiedene Möglichkeiten, **Zufallszahlen zu generieren.** Die erste besteht darin, Notizen herauszuziehen, die von 0 bis 9 nummeriert sind

eine Urne nach der anderen, und sie dann wieder in die Urne zu legen.

Zur Erzeugung von **Zufallszahlen können bestimmte physikalische Systeme genutzt werden:** zum Beispiel das Rouletterad. Die erhaltenen Zahlen können

in einer Zufallszahlentabelle aufgelistet werden. Diese

Tabellen sind nützlich, wenn **mehrere Simulationen** erforderlich sind

Dies muss für verschiedene Modelle durchgeführt werden, um zu testen, welches das beste ist. In diesem Fall kann es sich um die gleiche Reihe von Zufallszahlen handeln

wiederverwendet.

Einige Computer verwenden elektronische Schaltkreise oder elektromechanische Mittel, um eine Reihe von **Zufallszahlen** bereitzustellen . Da diese

Listen sind nicht reproduzierbar. Von Neumann, John und Lehmer, DH schlugen Methoden vor, die vollständig deterministisch sind, aber eine Reihe von Zahlen liefern, die scheinbar zufällig sind. Solche Zahlen werden als Pseudozufallszahlen bezeichnet.

Diese Pseudozufallszahlen werden aus einer vorgegebenen Algebra für Mula berechnet. Mit der Lehmer-Methode lässt sich beispielsweise eine Reihe von Pseudozufallszahlen x0, *x1,...,xn,...* berechnen

$$xi = a \cdot xiÿ1 \text{ (modulo m)}, \qquad ich = 1, 2, \ldots,$$

wobei x0 = *b* und a, b, *m* gegeben sind.
Die Wahl von a, b, *m* beeinflusst die Qualität der **Stichprobe** von Pseudozufallszahlen.
Die Unabhängigkeit zwischen den Pseudozufallszahlen erfordert, dass die Länge des Zyklus, d. h. die Menge der erzeugten Zahlen, bevor dieselbe Sequenz neu gestartet wird, lang genug sein muss. Um einen möglichst langen Zyklus zu erreichen, müssen folgende Kriterien erfüllt sein:

- *a* darf m nicht teilen, - x0 darf
m nicht teilen, - *m* muss groß sein.

DOMÄNEN UND EINSCHRÄNKUNGEN

Wenn die **Simulation Beobachtungen** erfordert , die aus ungleichmäßigen Verteilungen abgeleitet werden, gibt es mehrere Techniken, mit denen jedes beliebige Gesetz aus einer **gleichmäßigen Verteilung generiert werden kann.** Einige

Diese Methoden sind auf viele Distributionen anwendbar.

andere sind spezifischer und werden für eine bestimmte Verteilung verwendet.

BEISPIELE

Zur Berechnung nutzen wir die Lehmer-Methode

die ersten Pseudozufallszahlen der Reihe, die mit *a* = 33, *b* = 7 und *m* = 1000 erstellt wurden:

x0 = 7

x1 = 7 · 33 (mod1000) = 231 x2 = 231 · 33

(mod1000) = 623 x3 = 623 · 33 (mod1000) =

559

usw.

Machine Translated by Google

Eine Reihe von Zahlen *ri* , deren Verteilung nicht wesentlich abweicht
Eine Gleichverteilung erhält man, indem man *xi* durch m dividiert:

$$ri = \frac{xi}{M}, \qquad ich = 1, 2,...$$

Daher: r0 = 0,007

r1 = 0,231

r2 = 0,623

r3 = 0,559

usw.

WEITERLESEN

Kendall, Maurice George
Monte-Carlo-Methode
Zufallszahl
Simulation
Gleichmäßige Verteilung

VERWEISE

Dodge, Y.: Ein natürlicher Zufallszahlengenerator, Int. Stat.Review, 64, 329–344 (1996)

Ripley, BD: Gedanken zum Pseudozufall Zahlengeneratoren. J. Comput. Appl. Mathematik. **31,** 153–163 (1990)

Genetische Statistik

In der genetischen Statistik handelt es sich um die Analyse genetischer Daten durchgeführt mit Methoden und Konzepten von sowohl klassische als auch abgeleitete **Statistiken** aus der Theorie stochastischer Prozesse.
Genetik ist das Studium des erblichen Charakters und zufälliger Variationen; es ist gewohnt helfen, den Transformismus in der Praxis zu erklären Bereich, zur Verbesserung der Einheiten.

GESCHICHTE

Genetik und Statistik sind seit jeher miteinander verbunden über 100 Jahre. Sir **Galton, Francis** war Ich interessiere mich sehr für Studien zur Humangenetik und insbesondere für die Eugenik von Methoden, die verwendet werden könnten, um die zu verbessern genetische Qualität einer menschlichen Population, die erfordert offensichtlich Kenntnisse über die Vererbung. Galton erfand auch Regressions- und Korrelationskoeffizienten, um sie als zu verwenden statistische Hilfsmittel bei der Untersuchung der Genetik. Diese Methoden wurden später von **Pearson, Karl weiterentwickelt.** Im Jahr 1900 entstand das Werk von Mendel wurde wiederentdeckt und klassisch Mathematische und statistische Werkzeuge für Investitionen Es wurde eine entscheidende Mendel-Genetik vorgeschlagen von Sir **Fisher, Ronald Aylmer,** Wright, Sewall und Haldane, JBS in dieser Zeit 1920–1950.

DOMÄNEN UND EINSCHRÄNKUNGEN

Genetische Statistiken werden in Bereichen wie z wie Biomathematik, Bioinformatik, Biologie, **Epidemiologie** und Genetik. Standardmethoden zur **Schätzung** und zum Testen von Hypothesen sind Wird verwendet, um genetische Daten abzuschätzen und zu testen Parameter. Die Theorie stochastischer Prozesse wird verwendet, um die Einheiten der Evolution eines Subjekts einer Bevölkerung zu untersuchen Berücksichtigen Sie die zufälligen Veränderungen in der Häufigkeit von Genen.

BEISPIELE

Das Hardy-Weinberg-Gesetz ist das zentrale theoretische Modell der Populationsgenetik. Von den Engländern unabhängig formuliert Mathematiker Hardy, Godfrey H. und die Deutscher Arzt Weinberg, W. (1908), der Gleichungen dieses Gesetzes zeigen, dass genetische Variationen, wenn sie zum ersten Mal bei einigen Individuen einer Population auftreten, nicht verschwinden bei erblicher Übertragung, aber stattdessen auch in künftigen Generationen erhalten bleiben Proportionen, die den Mendelschen Gesetzen entsprechen.

Machine Translated by Google

Das Hardy-Weinberg-Gesetz erlaubt es uns, unter
Unter bestimmten Bedingungen können genotypische
Häufigkeiten aus allelischen Häufigkeiten berechnet werden.
Mit Genotyp meinen wir die erhaltenen genetischen
Eigenschaften eines Individuums
während der erblichen Übertragung von der
die Eltern der Person. Beispielsweise haben eineiige
Zwillinge den gleichen Genotyp. Ein
Ein Allel ist eine mögliche Version eines Gens, die
aus einer Nukleotidkette besteht. Dort
Daher kann ein bestimmtes Gen bei vielen vorkommen
verschiedene Allelformen in einer bestimmten
Population. In einem Individuum, wenn jedes Gen vorhanden ist
dargestellt durch zwei Allele, die aus identischen
Nukleotiden bestehen, ist das Individuum dann
homozygot für dieses Gen, und
wenn die Allele unterschiedliche Zusammensetzungen haben,
das Individuum ist hierfür dann heterozygot
Gen.
In der Originalfassung besagt das Gesetz, dass wenn
ein Gen wird von zwei Allelen *A* und B kontrolliert,
die mit den Häufigkeiten *p* und *q* = 1 ÿ *p* auftreten
jeweils in der Bevölkerung bei Generationt,
dann sind die Häufigkeiten der Genotypen bei der
Generation t+1 durch die folgende Gleichung gegeben:

$$p2AA + 2pqAB + q2BB .$$

Die Grundhypothese ist, dass die Größe der
Die Population ist groß genug, um
Stichprobenschwankungen zu minimieren. Außerdem
darf es keine Selektion, keine Mutation, keine
Migration (kein Erwerb/Verlust eines Allels) geben
und aufeinanderfolgende Generationen müssen
diskret sein (kein Generationswechsel).
Um einen *AA*- Genotyp zu erhalten, muss das Individuum
Sie erhalten von beiden Eltern ein Allele vom Typ A.
Wenn dieser Prozess zufällig ist (so die Hypothese von
Unabhängigkeit gilt), wird dieses Ereignis eintreten
mit einer Wahrscheinlichkeit von:

$$P(AA) = pp = \text{p2} .$$

Die gleiche Logik gilt für die Wahrscheinlichkeit von
Erhalt des Genotyps *BB:*

$$P(BB) = qq = \text{q2} .$$

Für den Genotyp *AB* schließlich gibt es zwei Fälle
möglich: Das Individuum erhielt A von seinem
Vater und *B* von ihrer Mutter oder umgekehrt,
So:

$$P(AB) = pq + qp = 2pq .$$

Daher ist in einer idealen Population die
Angegeben sind pHardy-Weinberg-Proportionen
von:

AA	*AB*	*BB*
p2	*2pq*	q2 .

Diese Situation kann auf ein Gen verallgemeinert werden
mit vielen Allelen A1, A2, ..., *Ak.* Der
Häufigkeit, mit der Homozygoten *(AiAi)* auftreten
entspricht:

$$f(AiAi) = \text{p2} \quad , \quad i = 1,...,k,$$

und die Häufigkeit, mit der Heterozygoten *(AiAj)*
auftreten ist gleich:

$$f\,AiAj = 2pipj ,$$
$$i = j,\ i = 1,...,k,\ j = 1,...,k.$$

Die Häufigkeit *p* des Allels *A* in der Generation
t+1 könnte auch von Interesse sein. Durch einfaches
Zählen erhalten wir:

$$p = \frac{\text{2p2N} + 2pqN}{2N} = \text{p2} + pq$$
$$= p \cdot (p + \text{q}) = p \cdot (1 \text{ ÿ } q + \text{q}) = p .$$

Die Häufigkeit des Allels *A* in der Generation
t+1 ist identisch mit der Häufigkeit dieses Allels
in der vorherigen Generation und der ersten Generation.

WEITERLESEN

Daten

Epidemiologie
Einschätzung

G

Machine Translated by Google

Hypothesentest
Unabhängigkeit
Parameter
Bevölkerung
Statistiken
Stochastischer Prozess

VERWEISE

Galton, F.: Typische Gesetze der Vererbung. Proz. Roy. Inst. Großbritannien 8, 282–301 (1877) (nachgedruckt in: Stigler, SM (1986) The Geschichte der Statistik: Die Messung von Unsicherheit vor 1900. Belknap Press, Cambridge, MA, S. 280)

Galton, F.: Ko-Beziehungen und ihre Messungen rement, hauptsächlich aus anthropologischen Daten. Proz. Roy. Soc. London. **45,** 135–145 (1888)

Hardy, GH: Mendelsche Proportionen in eine gemischte Bevölkerung. Wissenschaft **28,** 49–50 (1908)

Smith, CAB: Statistik in der Humangenetik. In: Kotz, S., Johnson, NL (Hrsg.) Encyclo pedia of Statistical Sciences, vol. 3. Wiley, New York (1983)

Geometrische Verteilung

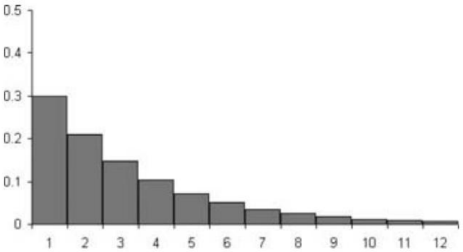
Betrachten Sie einen Prozess mit **k+1Bernoul-li-Versuchen** mit einer **Erfolgswahrscheinlichkeit** *p* und eine Ausfallwahrscheinlichkeit q.

Die **Zufallsvariable** X, entsprechend die Anzahl der Fehler *k* , die auftreten, wenn die

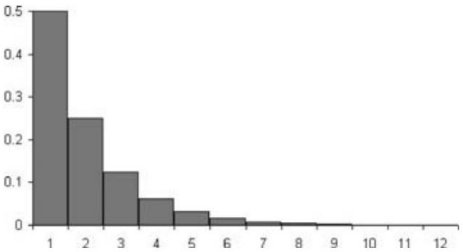
Der Vorgang wird bis zum ersten Erfolg wiederholt, folgt einer geometrischen Verteilung mit **Parameter** p, und so haben wir:

$$P(X = k) = p \cdot qk .$$

Die geometrische Verteilung ist **diskret Wahrscheinlichkeitsverteilung.**



Geometrische Verteilung, p = 0,3, q = 0,7



Geometrische Verteilung, p = 0,5, q = 0,5

Diese Verteilung wird „geometrisch“ genannt. denn die aufeinanderfolgenden Terme der oben beschriebenen **Wahrscheinlichkeitsfunktion** bilden eine geometrische Folge mit einem Verhältnis von q. Die geometrische Verteilung ist eine Besonderheit Fall der **negativen Binomialverteilung** wobei *r* = 1, d. h. der Prozess läuft bis zum ersten Erfolg.

MATHEMATISCHE ASPEKTE

Der **Erwartungswert** der geometrischen Verteilung ist per Definition gleich:

$$E[X] = \sum_{x=0}^{\infty} x \cdot P(X = x)$$

$$= \sum_{x=0}^{\infty} x \cdot p \cdot qx$$

$$= \frac{Q}{P} .$$

In der Tat:

$$E[X] = \sum_{x=0}^{\infty} x \cdot qx\ddot{y}1 \cdot p$$

Machine Translated by Google

$$= p \cdot \sum_{x=0}^{\infty} x \cdot qx\ddot{y}1$$

$$= 0 + p\,1 + 2 \cdot q + 3 \cdot q2 + \ldots .$$

Daher gilt: $1+2\cdot q+3\cdot q2+\ldots = \frac{1}{(1\ddot{y}q)2}$ seit:

$$q + q2 + q3 + \ldots = \frac{q}{q\,1\ddot{y}\,q}$$

$$q + q2 + q3 + \ldots = \frac{Q}{1\ddot{y}\,q}$$

$$1 + 2 \cdot q + 3 \cdot q2 + \ldots = \frac{1}{(1\ddot{y}\,q)^2}.$$

Die **Varianz** der geometrischen Verteilung ist per Definition gleich:

$$\mathrm{Var}(X) = E\,X2\ddot{y}\,(E[X])^2$$

$$= \sum_{x=0}^{\infty} x2 \cdot P(X = x)\ddot{y}\left(\frac{Q}{P}\right)^2$$

$$= \sum_{x=0}^{\infty} x2 \cdot p \cdot qx\ddot{y}\left(\frac{Q}{P}\right)^2$$

$$= p \cdot q(1 + 22 \cdot q + 32q2 + \ldots)\left(\frac{Q}{P}\right)^2$$

$$= \frac{q}{p2}.$$

Tatsächlich ist die Summe $1+22\cdot q+32q2+\ldots = \frac{1+q}{p2}$, seit:

$$q\,1 + 2 \cdot q + 3 \cdot q2 + \ldots = \frac{Q}{2\,(1\ddot{y}\,q)}$$

$$1 + 22 \cdot q + 32 \cdot q2 + \ldots\,1 = \frac{+\,q\,2}{(1\ddot{y}\,q)} = \frac{1+q}{p2}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die geometrische Verteilung wird relativ häufig in meteorologischen Modellen verwendet. Es ist wird auch in stochastischen Prozessen und der Theorie von Warteschlangen verwendet.

Die geometrische Verteilung kann auch in der folgenden Form ausgedrückt werden:

$$P(X = k) = p \cdot qk\ddot{y}1\ ,$$

wobei die **Zufallsvariable X** die Anzahl der Aufgaben darstellt, die erforderlich sind, um den ersten Erfolg zu erzielen (einschließlich des Erfolgs selbst).

G

BEISPIELE

Ein fairer Würfel wird geworfen und wir wollen das wissen

Wahrscheinlichkeit, dass beim vierten Wurf zum ersten Mal eine „Sechs“ geworfen wird.

Wir haben daher:

p = $\frac{1}{6}$ **(Wahrscheinlichkeit ,** dass eine Sechs geworfen wird)

q = $\frac{5}{6}$ **(Wahrscheinlichkeit ,** dass eine Sechs nicht gewürfelt wird)

k = **3** (da der vierte Wurf ein Erfolg ist, haben wir drei Fehlschläge)

Daher ist die Wahrscheinlichkeit wie folgt:

$$P(X = 3) = 6\left(\frac{1}{-}\right)\cdot\left(\frac{5}{6}\right)^3 = 0{,}0965\ .$$

WEITERLESEN

Bernoulli-Verteilung

Diskrete Wahrscheinlichkeitsverteilung

Negative Binomialverteilung

Geometrisches Mittel

Das geometrische Mittel wird als die Wurzel des Produkts von *n* nichtnegativen Zahlen definiert.

GESCHICHTE

Laut Droesbeke, J.-J. und Tassi, Ph. (1990), wurden das geometrische Mittel und das **harmonische Mittel** 1874 von Jevons, William Stanley, eingeführt.

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Sei x1, *x2,...,xn* eine Menge von *n* nichtnegativen Größen oder von *n* Beobachtungen, die sich auf eine **quantitative Variable** X beziehen. Das geometrische Mittel *G* dieser Menge ist:

$$G = \sqrt[n]{x1 \cdot x2 \cdot \quad \cdots \quad \cdot xn}.$$

Der geometrische Mittelwert wird manchmal in logarithmischer Form ausgedrückt:

$$\ln(G) = \frac{\sum_{i=1}^{N} \ln(xi)}{N},$$

oder

$$G = \exp\left(\frac{1}{N} \cdot \sum_{i=1}^{N} \ln(xi)\right).$$

Wir stellen fest, dass der Logarithmus des geometrischen Mittels einer Menge positiver Zahlen das **arithmetische Mittel** der Logarithmen dieser Zahlen ist (oder das **gewichtete arithmetische Mittel** im Fall gruppierter Beobachtungen).

DOMÄNEN UND EINSCHRÄNKUNGEN

In der Praxis wird das geometrische Mittel hauptsächlich zur Berechnung des **Mittelwerts** einer Gruppe von Verhältnissen oder insbesondere des Mittelwerts einer Gruppe von Indizes verwendet.

Ebenso wie das arithmetische Mittel berücksichtigt das geometrische Mittel jede **Beobachtung** einzeln. Sie verringert jedoch den Einfluss von Ausreißern auf den Mittelwert, weshalb sie manchmal der Arithmetik vorgezogen wird
bedeuten.

Ein wichtiger Aspekt des geometrischen Mittels besteht darin, dass es nur für positive Zahlen gilt.

BEISPIELE

Hier haben wir ein Beispiel für die Verwendung des geometrischen Mittels: 1987, 1988 und 1989 erzielt ein Unternehmer einen Gewinn von 10.000, 20.000 bzw. 160.000 Euro. Basierend auf diesen Informationen wollen wir die durchschnittliche Steigerungsrate dieser Gewinne ermitteln.

Von 1987 bis 1988 verdoppelte sich der Gewinn, von 1988 bis 1989 verachtfachte er sich. Wenn wir einfach das **arithmetische Mittel** dieser beiden Zahlen berechnen, erhalten wir

$$x = \frac{2 + 8}{2} = 5,$$

und wir kommen zu dem Schluss, dass sich der Gewinn jährlich im Durchschnitt verfünffachte. Wenn wir jedoch den Gewinn von 1987 (10.000 Euro) nehmen und ihn jährlich mit fünf multiplizieren, erhalten wir einen Gewinn von 50.000 im Jahr 1988 und 250.000 Euro im Jahr 1989.

Diese beiden Gewinne sind im Vergleich zu den tatsächlichen Gewinnen viel zu groß.
Wenn wir andererseits den Mittelwert anhand des geometrischen Mittels dieser beiden Steigerungen anstelle des arithmetischen Mittels berechnen, erhalten wir:

$$G = \sqrt{2 \cdot 8} = \sqrt{16} = 4,$$

und wir können mit Recht sagen, dass sich die Gewinne im Durchschnitt jährlich vervierfachten. Wendet man diese mittlere Steigerungsrate auf den anfänglichen Gewinn von 10.000 Euro im Jahr 1987 an, so erhält man 40.000 Euro im Jahr 1988 und 160.000 Euro im Jahr 1989. Auch wenn

Der Gewinn im Jahr 1988 ist zu hoch, er liegt unter dem vorherigen Ergebnis, und der Gewinn im Jahr 1989 ist es auch jetzt richtig.
Um die Verwendung der Formel für die zu veranschaulichen

Um den geometrischen Mittelwert in logarithmischer Form darzustellen, ermitteln wir nun den Mittelwert der folgenden Indizes:

112, 99, 105, 96, 85, 100 .

Machine Translated by Google

Wir berechnen (anhand einer Logarithmentabelle):

log (112) = 2,0492 log
(99) = 1,9956 log (105)
= 2,0212 log (96) = 1,9823
log (85) = 1,9294 log
(100) = 2,0000

11.9777

Wir erhalten dann:

$$\log (G) = = 1{,}9963 \frac{11{,}9777}{6}.$$

Daher ist $G = 99{,}15$.

WEITERLESEN

- **Arithmetisches Mittel**
- **Harmonische Mittel**
- **Bedeuten**
- **Maß für die zentrale Tendenz**

VERWEISE

Droesbeke, JJ, Tassi, PH: Histoire de la Statistique. Editionen „Que sais-je?“ Press es universitaires de France, Paris (1990)

Jevons, WS: Die Prinzipien der Wissenschaft: Eine Abhandlung über Logik und wissenschaftliche Methoden (zwei Bände). Macmillan, London (1874)

Geometrisch Standardabweichung

Die geometrische Standardabweichung einer Reihe quantitativer Beobachtungen ist ein **Maß für die Streuung.** Sie entspricht der Abweichung der Beobachtungen um die **Geometrie**
bedeuten.

GESCHICHTE

Siehe **L1- Schätzung.**

MATHEMATISCHE ASPEKTE

Sei x1, *x2,...,xn* eine Menge von *n* Beobachtungen, die sich auf eine **quantitative Variable** X beziehen.

Die geometrische Standardabweichung, bezeichnet mit ÿg, wird wie folgt berechnet:

$$\log \ddot{y}g = \frac{1}{N} \sum_{i=1}^{N} (\log xi \ddot{y} \log G)^{2}.$$

wobei $G = \ddot{y}n$ x1 · x2 · ric- ⋯ · *xn* ist das Geomet Mittelwert der Beobachtungen.

WEITERLESEN

- **L1 -Schätzung**
- **Maß der Streuung**
- **Standardabweichung**

Geostatistik

Geostatistik ist die Anwendung von Statistiken auf Probleme in der Geologie und Hydrologie. Geostatistiken behandeln naturgemäß räumliche Daten mit bekannten Koordinaten. Räumliche hierarchische Modelle folgen dem Prinzip: lokal modellieren, global analysieren. Auf allen Ebenen der Hierarchie werden einfache bedingte Modelle konstruiert (lokale Modellierung). Das Ergebnis ist ein gemeinsames Modell, das sehr komplex sein kann, eine Analyse ist aber dennoch möglich (globale Analyse). Die Geostatistik umfasst den in der Theorie verwendeten Satz sowie die Techniken und statistischen Anwendungen, die zur Analyse und Vorhersage der Verteilung der Werte der Variablen im Raum und (letztendlich) in der Zeit verwendet werd

GESCHICHTE

Der Begriff Geostatistik wurde erstmals von Hart (1954) in einem geografischen Kontext verwendet

G

Machine Translated by Google

die Anwendung bestimmter statistischer Techniken
auf Beobachtungen hervorzuheben
eine regionale Verteilung. Die ersten geostatistischen
Konzepte wurden von Matheron formuliert
(1963) an der Ecole des Mines de Paris
um die Reserven eines Bergmanns abzuschätzen
al aus räumlich verteilten Daten (Beobachtungen).

DOMÄNEN UND EINSCHRÄNKUNGEN

Geostatistik wird in einer Vielzahl von Bereichen eingesetzt
Bereiche, einschließlich Vorhersage (von Niederschlag,
Bodenporosität, Konzentrationen von
Schwermetalle) und bei der Bildbearbeitung
von Satelliten.

WEITERLESEN

Geodaten

Raumstatistik

VERWEISE

Atteia, O., Dubois, J.-P., Webster, R.: Geostatistische
Analyse der Bodenkontamination in
der Schweizer Jura. Umgebung. Umweltverschmutzung. **86,** 315–
327 (1994)

Cressie, NAC: Statistik für räumliche Daten.
Wiley, New York (1991)

Goovaerts, P.: Geostatistik für die Natur
Ressourcenbewertung. Universität Oxford
sity Press, Oxford (1997)

Isaaks, EH, Srivastava, RM: Eine Einführung
Einführung in die Angewandte Geostatistik. Oxford
University Press, Oxford (1989)

Journel, A., Huijbregts, CJ: Bergbaugeostatistik.
Akademisch, New York (1981)

Matheron, G.: Die Theorie der Variablen
regionalisiert und ihre Anwendungen. Mas Sohn,
Paris (1965)

Oliver,MA,Webster,R.:Kriging: amethod
der Interpolation für geografische Informationen
tionssystem. Int. J. Geogr. Inf. Syst. 4(3),
313–332 (1990)

Pannatier, Y.: Variowin 2.2. Software
für räumliche Datenanalyse in 2D.
Springer, Berlin Heidelberg New York
(1996)

Gini, Corrado

Gini, Corrado (1884–1965) wurde in Mat ta di Livenza,
Italien, geboren. Obwohl er seinen Abschluss machte
1905 schloss er sein Jurastudium ab und belegte Kurse in Mathematik
und Biologie während seines Studiums. Seine
späteren Neigungen sowohl zur Wissenschaft als auch
Statistik führte ihn dazu, vorübergehend Professor für
Statistik an der Universität Cagliari zu werden
1909 und 1920 übernahm er den Lehrstuhl
für Statistik an derselben Universität. Im Jahr 1913
Er begann an der Universität Padua zu unterrichten
hat den Lehrstuhl für Statistik an der Uni übernommen
Universität von Rom im Jahr 1927.
Zwischen 1926 und 1932 war er auch der
Präsident des Zentralinstituts für Statistik
(ISTAT).
Er gründete zwei Zeitschriften. Der erste,
Metron (gegründet 1920) ist eine internationale
Statistikzeitschrift, und die zweite,
Genus (gegründet 1934) ist das Journal of
das italienische Komitee zur Untersuchung von
Bevölkerungsproblemen.
Die Beiträge von Gini auf dem Gebiet der
Bei Statistiken handelt es sich hauptsächlich um Mittelwerte
und die Variabilität sowie Assoziationen
zwischen den Zufallsvariablen. Er betrügt auch
würdigte Originalwerke aus den Bereichen Wirtschaft,
Soziologie, Demografie und Biologie.

Einige Hauptwerke und Artikel von Gini, Corrado:

1910 Konzentrations- und Abhängigkeitsanzeiger
denza. Atti della III Riunione della

Machine Translated by Google

Societa Italiana per il Progresso delle Scienze, R12, 3–210.

1912 Variabilität und Mutabilität. Studi economico-giuridici, Università di Cagliari, III 2a, R12, 211–382.

Gini-Index

Der Gini-Index ist der am häufigsten verwendete Ungleichheitsindex; Es ermöglicht uns, den Grad der Ungleichheit in der Einkommensverteilung einer bestimmten Bevölkerung zu messen. Grafisch wird der Gini-Index durch die Fläche zwischen einer Linie bei 45ÿ und der dargestellt Lorenz-Kurve (eine grafische Darstellung des kumulierten Prozentsatzes des Gesamteinkommens im Vergleich zum kumulierten Prozentsatz der Bevölkerung, die dieses Einkommen erhält, wobei das Einkommen von links nach rechts steigt). Wenn man die Oberfläche zwischen der Linie bei 45ÿ und der Lorenz-Kurve (Fläche *A* im folgenden Diagramm) durch die Gesamtoberfläche unter der Linie *(A* + B) dividiert, erhält man den Gini-Koeffizienten:

$$IG = \frac{A}{A + B}.$$

Der Gini-Koeffizient ist Null, wenn es keinen gibt Ungleichheit in der Einkommensverteilung und ein Wert von 1, wenn die Einkommensverteilung völlig ungleich ist, also 0 ÿ *IG* ÿ 1. Der Gini-Index ist einfach der Gini-Koeffizient multipliziert mit 100, um ihn in einen Prozentsatz umzurechnen. Je näher der Index also bei 100 % liegt, desto ungleicher ist die Einkommensverteilung. In entwickelten Ländern liegen die Werte des Giniindex bei etwa 40 % (siehe Beispiel).

GESCHICHTE

Der Gini-Koeffizient wurde 1912 erfunden vom italienischen Statistiker und Demografen **Gini, Corrado.**

MATHEMATISCHE ASPEKTE

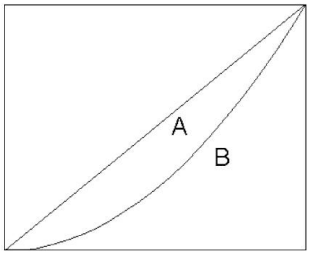
Die Gesamtoberfläche unter der Linie bei 45ÿ *(A* + B) gleich 1 und so können wir schreiben: 2 ,

$$IG = \frac{A}{A + B} = \frac{A}{\frac{1}{2}} = 2A \qquad (1)$$
$$= 2\left(\frac{1}{2} \text{ ÿ } B\right) \text{ ÿ } IG = 1 \text{ ÿ } 2B$$

Um den Wert des Gini-Koeffizienten zu berechnen, müssen wir daher B (Gesamtfläche unter der Lorenz-Kurve) berechnen.

a) *Gini-Index: diskreter Fall*

Für eine diskrete Lorenzkurve wo die Einkommensverteilung ist aufsteigend geordnet *(x1* ÿ ... ÿ xn), die Lorenz-Kurve hat folgenden Verlauf:



Um B zu berechnen, müssen wir die Oberfläche unter der Lorenz-Kurve in eine Reihe unterteilen von Polygonen (wobei das erste Polygon ein Dreieck ist und die anderen Trapeze sind). Die Oberfläche des Dreiecks ist:

$$\frac{1}{2n}\,\frac{x1}{X}.$$

Die Oberfläche des j-ten Trapezes ist:

$$\frac{1}{2n}\left(\sum_{i=1}^{j-1} \frac{xi}{X} + \sum_{i=1}^{J} \frac{xi}{X}\right)$$
$$= \frac{1}{2nX}\left(\sum_{i=1}^{j-1} xi + xj\right),$$

Machine Translated by Google

mit *j = 2,...,n*. Das haben wir dann

$$B = \frac{1}{2n} \cdot \frac{x1}{X} + n \sum_{j=2}^{j\ddot{y}1} \left(\frac{1}{2nX} \ddot{y} \ddot{y}2 \cdot \sum_{i=1}^{xi + xj} \ddot{y} \ddot{y} \right),$$

Wo:

x1,...,xn sind die Einkommen; ***X*** ist der Gesamtumsatz; **x1** $\frac{+ ... + \boldsymbol{xn}}{\boldsymbol{X}}$ ist die kumulierte Summe Umsatz, in %.

Nachdem wir diese Gleichung entwickelt haben, erhalten wir:

$$B = \frac{1}{2nX} \cdot \ddot{y}X + 2 \cdot \sum_{i=1}^{N} (n \ddot{y} i + 1)xi . \quad (2)$$

Unter Verwendung von Gl. (1) können wir dann den Gini-Index erhalten:

$$IG = 1 \ddot{y} 2B = 1 \ddot{y} \frac{2}{2nX} \ddot{y}X + 2 \cdot \sum_{i=1}^{N} (n \ddot{y} i + 1)xi$$

$$1 = 1 + \frac{}{N} \ddot{y} 2 \cdot \frac{\sum_{i=1}^{N} (n \ddot{y} i + 1)\, xi}{nX}$$

$$1 = 1 + \frac{}{N} \frac{2}{n2\ddot{y}\,(x)} \sum_{i=1}^{N} ixn\ddot{y}i+1 ,$$

wobei $\ddot{y}\,(x) = \frac{1}{N} \sum_{i=1}^{N} xi$ = mittleres Einkommen der Verteilung.

Der Gin-Koeffizient kann umformuliert werden auf viele Arten. Wir könnten zum Beispiel verwenden:

$$IG = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{xi \ddot{y} xj}{2n2\ddot{y}\,(x)} .$$

b) *Gini-Index: stetiger Fall* In diesem Fall entspricht *B* dem Integral von 0 bis 1 der Lorenz-Funktion.

Daher müssen wir dieses Integral nur in Gleichung einsetzen. (1):

$$IG = 1 \ddot{y} 2B = 1 \ddot{y} 2 \int_0^1 L\,(P)\, dP ,$$

wobei $P = F\,(y) = \int_0^j f\,(x)\, dx$ ist das pro Prozentsatz der Bevölkerung mit einem Einkommen kleiner als y. Zum Beispiel,

$$IG = \int_0^1 1 \ddot{y} (1 \ddot{y} P) \frac{\ddot{y}\ddot{y}1}{\ddot{y}} dP$$

$$= 1 \ddot{y} \frac{\ddot{y}}{2\ddot{y} \ddot{y} 1}$$

$$= 1 \ddot{y} 2 1 \ddot{y} 2\ddot{y} \ddot{y} 1 \frac{\ddot{y}}{} .$$

BEREICHE UND GRENZEN

Der Gini-Index (oder Koeffizient) ist der Parameter, der am häufigsten zur Messung des Ausmaßes der Ungleichheit in einer Einkommensverteilung verwendet wird. Es kann auch zur Messung des Diffs verwendet wer Unterschiede in der Arbeitslosigkeit zwischen Regionen.

BEISPIELE

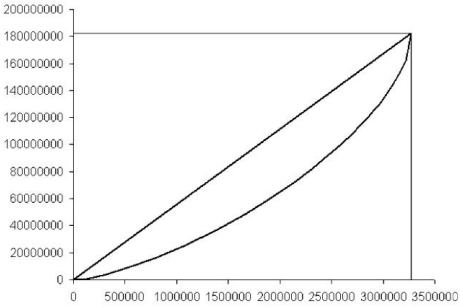
Als Beispiel für die Berechnung des Gini-Koeffizienten (oder Index) nehmen wir den diskreten Fall.

Es gilt: *IG = 1 ÿ 2B.* Denken Sie daran, dass wir in diesem Fall die Oberfläche *B* ermitteln , indem wir die Summe einer Reihe von Polygonen unter der Lorenz-Kurve berechnen.

Die Einkommensdaten für eine bestimmte Bevölkerung im Zeitraum 1993–1994 sind in der folgenden Tabelle zusammen mit den den Einkommensintervallen zugeordneten Gebieten aufgeführt:

Machine Translated by Google

Bevölkerungseinkommen	Oberfläche	
0	0	
94569	173100.4	0,000013733
37287	527244.8	0,000027323
186540	3242874.9	0,000726701
203530	4603346.0	0,00213263
266762	7329989.6	0,005465843
290524	9428392.6	0,010037284
297924	11153963.1	0,015437321
284686	12067808.2	0,02029755
251155	11898201.1	0,022956602
216164	11317647.7	0,023968454
182268	10455257.4	0,023539381
153903	9596182.1	0,022465371
127101	8560949.7	0,020488918
104788	7581237.4	0,018311098
85609	6621873.3	0,01597976
70108	5773641.3	0,013815408
57483	5021336.9	0,011848125
47385	4376126.7	0,010140353
39392	3835258.6	0,008701224
103116	11213361.3	0,024078927
71262	9463751.6	0,017876779
46387	7917846.5	0,012313061
50603	20181113.8	0,0114625049



Die Fläche *B* ist also gleich:

$$B = \frac{1}{2nX} \cdot \ddot{y}X + 2 \cdot \sum_{i=1}^{N} (n \ddot{y} i + 1) \cdot xi$$

$$= 0{,}315246894 .$$

Wir haben dann:

$$IG = 1 \ddot{y} 2B = 1 \ddot{y} 2 \cdot 0{,}315246894$$

$$= 0{,}369506212 .$$

WEITERLESEN

Gini, Corrado

Einfache Indexnummer

Gleichmäßige Verteilung

VERWEISE

Greenwald, D. (Hrsg.): Encyclopédie Wirtschaft. Economica, Paris, S. 164–165 (1984)

Flückiger, Y.: Sozioökonomische Analyse des kantonalen Unterschieds der Chomage. Le chômage en Suisse: bilan et perspectives. Forum Helvet. 6, 91–113 (1995)

Gütetest

Durch die Durchführung eines Anpassungstests an einer **Probe** können wir feststellen, ob die Die beobachtete Verteilung entspricht einer bestimmten **Wahrscheinlichkeitsverteilung (z .** B. der **Normalverteilung** oder **Poisson-Verteilung).**

Dadurch können wir herausfinden, ob Die beobachtete Stichprobe wurde aus einer **Population** gezogen , die dieser Verteilung folgt.

GESCHICHTE

Siehe **Chi-Quadrat-Anpassungstest** und **Kolmogorov-Smirnov-Test.**

MATHEMATISCHE ASPEKTE

Das Ziel des Anpassungstests besteht darin Bestimmen Sie, ob eine beobachtete **Probe** vorhanden war wird aus einer **Grundgesamtheit** gezogen , die einer bestimmten **Wahrscheinlichkeitsverteilung folgt.**

Machine Translated by Google

Der verwendete Prozess beinhaltet normalerweise einige **Hypothesentest:**

Nullhypothese H0 : $F = F0$

Alternativhypothese H1 : $F = F0$

wobei *F* eine unbekannte **Verteilungsfunktion** ist

tion der Grundgesamtheit und F0 ist
die angenommene Verteilungsfunktion.
Der Anpassungstest beinhaltet einen Vergleich
die empirische Verteilung mit der Vermuteten
Verteilung. Wir lehnen die Nullhypothese ab, wenn
die empirische Verteilung ist nicht nah genug
zur vermuteten Verteilung. Das Präzise
Regeln, die bestimmen, ob die Nullhypothese gilt
Je nach Art wird die Anfrage abgelehnt oder akzeptiert
der verwendete Test.

BEISPIELE

Es gibt viele Anpassungsgütetests, darunter den **Chi-Quadrat-Anpassungsgütetest** und
der **Kolmogorov-Smirnov-Test.**

WEITERLESEN

- **Chi-Quadrat-Anpassungstest**
- **Hypothesentest**
- **Kolmogorov-Smirnov-Test**

Gosset, William Sealy

Gosset, William Sealy, besser bekannt unter seinem Pseudonym „Student“, wurde 1876 in Canterbury, England, geboren. Nach dem Studium Mathematik und Chemie am New College, Oxford begann er als Brauer zu arbeiten GuinnessBrauereieninDublinim Jahr 1899.Guin ness war ein Unternehmen, das Forschung bevorzugte, und stellte ihm seine Labore zur Verfügung Brauer und ihre Chemiker. Tatsächlich im Jahr 1900 eröffnete das „Guinness Research Laborato ry“, in dem einer der größten Chemiker beschäftigt war Damals war Horace Brown als Regisseur tätig. Arbeiten, die in diesem Labor durchgeführt wurden Tory bezog sich auf die Qualität und die Kosten von die vielen Gersten- und Hopfensorten.
In diesem Umfeld entwickelte sich Gosset eröffnete sein Interesse an **Statistik.**
In Oxford hatte Gosset Mathematik studiert, und so wurde er oft von seinen Kollegen bei Guinness hinzugezogen, wenn bestimmte mathematische Probleme auftauchten. Das führte dazu, dass er sich mit der Fehlertheorie beschäftigte **und** sich beraten ließ **Pearson, K.,** den er im Juli 1905 kennenlernte, weiter dieses Thema.
Die Hauptschwierigkeit, mit der Gosset konfrontiert war Während seiner Arbeit gab es kleine Stichprobengrößen. Er Daher beschloss ich, eine geeignete Methode für den Umgang mit den Daten zu finden diese kleinen **Proben.**
In den Jahren 1906–1907 verbrachte Gosset ein Jahr lang in Zusammenarbeit mit **Pearson, K.,** in Pearsons Laboratorium am University College London.
Versuch, Methoden im Zusammenhang mit dem zu entwickeln wahrscheinlicher **Fehler** des **Mittelwerts.**
Im Jahr 1907 wurde Gosset dafür verantwortlich gemacht Guinness' Experimental Brewery und er nutzte dazu die **Student-Tabelle ,** die er zuvor experimentell abgeleitet hatte Bestimmen Sie die beste Gerstensorte zum Brauen. Guinness, vorsichtig bei der Veröffentlichung Geschäftsgeheimnisse erlaubten es Gosset nur, sein Werk unter dem Pseudonym eines von beiden zu veröffentlichen „Schüler“ oder „Student“, und so wählte er das Latein ter.

Er starb 1937 und hinterließ wichtige Werke, die alle unter dem Pseudonym veröffentlicht wurden von „Student“.

Gossets Werk blieb lange Zeit weitgehend unbeachtet viele Jahre, und nur wenige Wissenschaftler nutzten ihn Tisch abseits der Guinness-Forscher Brauereien und die der Rothamsted Experimental Station (wo **Fisher, RA** arbeitete) . als Statistiker).

Machine Translated by Google

Einige der Hauptwerke und Artikel von Gosset, WS:

1907 Über den Fehler beim Zählen mit einem Hämozytometer. Biometrie 5, 351–360.

1908 Der wahrscheinliche Fehler eines Mittelwerts. Biometrie 6, 1–25

1925 Neue Tabellen zur Prüfung der Bedeutung von Beobachtungen. Metron 5, 105–120.

1942 Pearson, ES und Wishart, J. (Hrsg.) Gesammelte Aufsätze der Studierenden (Vorwort von McMullen, L.). Cambridge University Press, Cambridge (herausgegeben vom Biometrika Office, University College London).

WEITERLESEN

Studentenverteilung
Studententisch
Schülertest

VERWEISE

Fisher-Box, J.: Guinness, Gosset, Fisher und kleine Proben. Stat. Wissenschaft. 2, 45–52 (1987)

McMullen, L., Pearson, ES: William Sealy Gosset, 1876–1937. In: Pearson, ES, Kendall, M. (Hrsg.) Studies in the History of Statistics and Probability, vol. I. Griffin, London (1970)

Plackett, RL, Barnard, GA (Hrsg.): Student: A Statistical Biography of William Sealy Gosset (Basierend auf Schriften von ES Pear, Sohn). Clarendon, Oxford (1990)

Griechisch-lateinische Quadrate

Siehe **griechisch-lateinisches quadratisches Design.**

Griechisch-lateinisches quadratisches Design

Ein griechisch-lateinisches Quadratdesign ist ein **Versuchsplan,** bei dem die Versuchseinheiten auf drei verschiedene Arten gruppiert sind. Man erhält es durch Überlagerung zweier lateinischer Quadrate gleicher Größe.

Wenn jeder lateinische Buchstabe genau einmal mit einem griechischen Buchstaben übereinstimmt, sind die beiden lateinischen Quadratmuster orthogonal. Zusammen bilden sie ein griechisch-lateinisches quadratisches Design.

In diesem Design erscheint jedes **Behandlungszeichen** (lateinischer Buchstabe) nur einmal in jeder Zeile, einmal in jeder Spalte und einmal mit jedem griechischen Buchstaben.

GESCHICHTE Der Bau eines griechisch-lateinischen Platzes geht auf Euler, Leonhard (1782) zurück. Ein Buch von **Fisher, Ronald Aylmer** und Yates, Frank (1963) gibt griechisch-lateinische Tabellen der Ordnung 3 bis zur Ordnung 12 (ohne die Ordnung sechs). Auch das Buch von Denes und Keedwell (1974) enthielt umfassende Informationen über griechisch-lateinische Quadratdesigns. Dodge und Shah (1977) behandeln den Fall der **Schätzung ,** wenn Daten fehlen.

DOMÄNEN UND EINSCHRÄNKUNGEN

Griechisch-lateinische quadratische Designs werden verwendet, um die Auswirkungen von drei Quellen systematischer Fehler zu reduzieren.

Es gibt griechisch-lateinische quadratische Designs für jede Größe n außer $n = 1$, 2 und 6.

BEISPIELE

Wir wollen drei verschiedene Benzinsorten testen. Dafür haben wir drei Fahrer und drei Fahrzeuge. Allerdings ist der Test zu aufwändig, um ihn an einem Tag durchzuführen: Wir müssen das Experiment über drei Tage durchführen.

In diesem Fall ist der Einsatz von Fahrern und Fahrzeugen überflüssig

Machine Translated by Google

Drei verschiedene Tage könnten zu systematischen Fehlern führen. Um diese Fehler zu beseitigen, erstellen wir daher einen Versuchsplan mit griechisch-lateinischen Quadraten. In diesem Fall wird jede Benzinsorte **(Behandlung)** nur einmal bei jedem Fahrer, einmal bei jedem Fahrzeug und einmal täglich getestet.

In diesem griechisch-lateinischen Quadratdesign stellen wir die verschiedenen Benzinarten durch die *lateinischen Buchstaben A, BandC und* die verschiedenen Tage durch ÿ, ÿ und ÿ dar.

	Treiber		
	1	2	3
Fahrzeuge 1	Aÿ	Bÿ	Cÿ
2	Bÿ	Cÿ	Aÿ
3	Cÿ	Aÿ	Bÿ

Eine **Varianzanalyse** wird uns dann zeigen, ob nach Eliminierung der Auswirkungen der Zeilen (Fahrzeuge), der Spalten (Fahrer) und der griechischen Buchstaben (Tage) ein signifikanter Unterschied zwischen den Benzinsorten besteht.

WEITERLESEN

Varianzanalyse

Versuchsplanung

Lateinische quadratische Designs

REFERENZEN

Dénes, J., Keedwell, AD: Lateinische Quadrate und ihre Anwendungen. Akademisch, New York (1974)

Dodge, Y., Shah, KR: Schätzung von Parametern in lateinischen Quadraten und griechisch-lateinischen Quadraten mit fehlenden Beobachtungen. Kommun. Stat. Theorie A6(15), 1465–1472 (1977)

Euler, L.: Recherches sur une nouvelle espèce de carrés magiques. Verh. Zeeuw.
General Weten. Vlissengen 9, 85–239 (1782)

Fisher, RA, Yates, F.: Statistische Tabellen für Biologie, Landwirtschaft und Medizin Forschung, 6. Aufl. Hafner (Macmillan), New York (1963)

Grafische Darstellung

Grafische Darstellungen umfassen eine Vielzahl von Techniken, die zur Verdeutlichung, Interpretation und Analyse von Daten durch das Plotten von Punkten und das Zeichnen von Liniensegmenten, Oberflächen und anderen geometrischen Formen oder Symbolen verwendet werden. Der Zweck eines Diagramms ist die schnelle Visualisierung eines Datensatzes. Es sollte beispielsweise das allgemeine Verhalten des untersuchten Phänomens klar veranschaulichen und alle wichtigen Faktoren hervorheben. Es kann beispielsweise zum Übersetzen oder Vervollständigen einer **Häufigkeitstabelle verwendet werden.**
Daher ist die grafische Darstellung eine Form der Datendarstellung.

GESCHICHTE Das Konzept, einen Punkt im Koordinatenraum zu zeichnen, geht zumindest auf die alten Griechen zurück, aber wir mussten bis zur Arbeit von Descartes und René warten, bis Mathematiker dieses Konzept untersuchten.
Laut Royston, E. (1970), einem deutschen Mathematiker namens Crome, war AW einer der ersten, der die grafische Darstellung in der **Statistik verwendete.** Er nutzte es zunächst als Lehrmittel.

In seinen Werken *„Geographisch-statistische Darstellung der Staatskräfte“ (1820)* und „ Über *die Größe und Bevölkerung der gesamten Europäischen Staaten“ (1785)* verwendete Crome verschiedene Arten der grafischen Darstellung, darunter das **Kreisdiagramm.**

Royston, E. (1970) zitiert auch Playfair, W., dessen Werk *The Commercial and Polit*

Auch *der ische Atlas* (1786) bezog sich auf verschiedene

Machine Translated by Google

grafische Darstellungen, insbesondere die **Linie Diagramm.** Playfair war sehr interessiert daran internationale Handelsbilanz und illustriert seine Studien mit verschiedenen Grafiken wie das **Histogramm** und das Kreisdiagramm. Zunächst wurde der Begriff Histogramm verwendet Zeit von **Pearson, Karl** und Guerry (1833) scheint zu den ersten gehört zu haben, die es verwendet haben das Liniendiagramm; in seinem *Essai sur la Statis tique Morale de la France,* veröffentlicht in Im Jahr 1833 beschrieb er in Paris die Frequenzen von Verbrechen nach ihren Merkmalen; diese Studie stellte eine der ersten Anwendungen einer Frequenzverteilung dar.
Schmid, Calvin F. (1954) illustriert und beschreibt die verschiedenen Arten grafischer Darstellungen in seinem *Handbook of Graphic Präsentation.*

MATHEMATISCHE ASPEKTE

Die Art der verwendeten Grafik hängt davon ab die Art der darzustellenden **Daten ,** die Art der untersuchten **Variablen** und das Ziel der lernen:

- Eine **quantitative Grafik** ist besonders wichtig nützlich für die Darstellung qualitativer kategorialer Variablen. Solche Grafiken umfassen das **Kreisdiagramm,** das **Liniendiagramm** und das **Bilddiagramm.**
- Eine Häufigkeitsgrafik ermöglicht es uns, die (diskrete oder kontinuierliche) **Häufigkeitsverteilung** einer **Quantität** darzustellen **Variable.** Zu diesen Grafiken gehören die **his Togramm** und das **Stamm- und Blattdiagramm .**
- Eine kartesische Grafik verwendet ein System von Achsen, die senkrecht zueinander stehen und kreuzen sich an einem Punkt, der als „Ursprung“ bekannt ist. Ein solches Koordinatensystem heißt „kartesisch“.

WEITERLESEN

Häufigkeitstabelle

Quantitatives Diagramm

G

VERWEISE

Crome, AFW: Über die Grösse und Bevölkerung der sämtlichen europäischen Staaten. Weygand, Leipzig (1785)

Crome, AFW: Geographisch-statistische Darstellung der Staatskräfte. Weygand, Leipzig (1820)

Fienberg, SE: Grafische Methode in der Statistik. Bin. Stat. **33,** 165–178 (1979)

Guerry, AM: Essay über die statistische Moral de la France. Crochard, Paris (1833)

Playfair, W.: Das Kommerzielle und Politische Atlas. Schauplatz, London (1786)

Royston, E.: Eine Anmerkung zur Geschichte der grafische Darstellung von Daten. In: Pear Son, ES, Kendall, M. (Hrsg.) Studien in der Geschichte der Statistik und Wahrscheinlichkeit, vol. ICH. Griffin, London (1970)

Schmid, CF: Handbuch der grafischen Präsentation. Ronald Press, New York (1954)

Machine Translated by Google

Hájek, Jaroslav

Hájek, Jaroslav wurde 1926 in Pode Brady, Böhmen, geboren. Von Beruf war er Statistikingenieur und promovierte 1954. Von 1954 bis 1964 arbeitete er als Forscher am Institut für Mathematik

der Tschechoslowakischen Akademie der Wissenschaften. Anschließend wechselte er an die Karls-Universität in Prag, wo er von 1966 bis zu seinem frühen Tod 1974 Professor war. Die Hauptwerke von Hájek, J. betreffen die Stichprobenwahrscheinlichkeitstheorie und die Rangtesttheorie. Insbesondere entwickelte er eine asymptotische Theorie der Statistik linearer Ränge. Er war der erste, der das Konzept der Invarianz auf die Theorie des Rangtests anwendete.

Einige Hauptwerke und Artikel von Hájek, Jaroslav:

1955 Einige Rangverteilungen und ihre Anwendungen. Cas. Pest. Matte. 80, 17–31 (auf Tschechisch); Übersetzung in (1960) Wählen. Übers. Mathematik. Stat. Probab., 2, 41–61.

1958 Einige Beiträge zur Theorie der Wahrscheinlichkeitsstichprobe. Stier. Inst. Int. Stat., 36, 127–134.

1961 Einige Erweiterungen des Wald–Wolfowitz-Noether-Theorem. Ann. Mathematik. Stat. 32, 506–523.

1964 Asymptotische Theorie der abweisenden Stichprobe mit unterschiedlichen Wahrscheinlichkeiten aus einer endlichen Grundgesamtheit. Ann. Mathematik. Stat. 35, 1419–1523.

1965 Erweiterung des Kolmogorov-Smirnov-Tests auf Regressionsalternativen. In: Neyman, J. und LeCam, L. (Hrsg.) Bernoulli–Bayes–Laplace: Proceedings of an International Seminarar, 1963. Springer, Berlin Heidelberg New York, S. 45–60.

1981 Dupac, V. (ed) Stichprobe aus einer endlichen Bevölkerung. Marcel Dekker, New York.

Harmonische Mittel

Das harmonische **Mittel** von *n* Beobachtungen ist definiert als *n* dividiert durch die Summe der Umkehrwerte aller Beobachtungen.

GESCHICHTE

Siehe **geometrisches Mittel.**
Die Beziehung zwischen dem harmonischen Mittel, dem geometrischen Mittel und dem **arithmetischen Mittel** wird von Mitrinovic, DS (1970) beschrieben.

MATHEMATISCHE ASPEKTE

Seien x1, *x2,...,xn n* Größen ungleich Null oder *n* Beobachtungen, die sich auf eine **Größe** beziehen

Machine Translated by Google

Variable X. Das harmonische Mittel *H* davon

n Mengen wird wie folgt berechnet:

$$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{xi}} .$$

Wenn {xi}i=1,...,n eine endliche Reihe positiver Zahlen darstellt, geben wir Folgendes an:

$$\text{Mindest } xi \ddot{y} H \ddot{y} G \ddot{y} \bar{x} \ddot{y} \max \quad xi .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Das harmonische Mittel wird in **der Statistik selten verwendet.** Manchmal kann es jedoch nützlich sein, wie zum Beispiel in folgenden Fällen:

- Wenn eine Reihe von Investitionen zu Dif

  Es gibt verschiedene Zinssätze, und sie alle geben das gleiches Einkommen, der einzigartige Satz, zu dem alle des in diesen Investitionen gebundenen Kapitals

  Es muss investiert werden, um den gleichen Ertrag zu erwirtschaften, der durch die Gesamtheit der Investitionen gegeben ist

  gleich dem harmonischen Mittel der Einzelraten.

- Angenommen, wir haben eine Gruppe verschiedener Materialien und jedes Material kann gekauft werden zu einem bestimmten Preis pro Materialmenge (wobei der Preis pro Menge für jedes Material unterschiedlich sein kann). Wir kaufen dann von jedem Material eine bestimmte Menge und geben dafür aus jeweils der gleiche Geldbetrag. In diesem

  In diesem Fall der Durchschnittspreis pro Menge über alle Materialien werden durch das harmonische Mittel angegeben

  der Preise pro Menge für alle thematischen Rials.

- Eine Eigenschaft des harmonischen Mittelwerts ist das

  Es ist weitgehend unempfindlich gegenüber Ausreißern viel größere Werte als die anderen Daten.

  Betrachten Sie beispielsweise die folgenden Werte: 1, 2, 3, 4, 5 und 100. Hier beträgt das harmonische Mittel 2,62 und das **arithmetische Mittel** 19,17. Allerdings ist die

  Das harmonische Mittel ist viel empfindlicher zu Ausreißern, wenn sie viel kleinere Werte als der Rest der Daten haben. So für

  die Beobachtungen 1, 6, 6, 6, 6, 6 erhalten wir *H* = 3,27, wobei das arithmetische Mittel

  entspricht 5,17.

BEISPIELE

Drei Investitionen, die jeweils den gleichen Ertrag bringen Einkommen haben die folgenden Zinssätze: 5 %, 10 % und 15 %.

Das harmonische Mittel ergibt den Zinssatz bei dem das gesamte Kapital vorhanden sein müsste investiert, um die gleiche Summe zu produzieren Einkommen als die drei separaten Investitionen:

$$H = \frac{3}{\frac{1}{5} + \frac{1}{10} + \frac{1}{15}} = \frac{3}{\frac{11}{30}} = 8{,}18\ \% .$$

Wir stellen fest, dass sich dieses Ergebnis von dem unterscheidet **arithmetisches Mittel** von 10 % (5 + 10 + 15)/3.

Ein Vertreter kauft drei Partien Kaffee, jeweils von einer anderen Note (Qualität), bei 3, 2 und 1,5 Euro pro kg. Er kauft 200 Euro pro Jahrgangsstufe.

Der mittlere Preis pro kg Kaffee beträgt dann erhalten, indem man die Gesamtkosten durch die Gesamtkosten dividiert Gekaufte Menge:

$$\text{Durchschnittspreis} = \frac{\text{Gesamtkosten}}{\text{Gesamtmenge}} = \frac{3 \cdot 200}{66{,}66 + 100 + 133{,}33} = 2 .$$

Dies entspricht dem harmonischen Mittel von die Preise der drei verschiedenen Kaffeesorten:

$$\text{mittlerer Preis} = \frac{3}{\frac{1}{3} + \frac{1}{2} + \frac{1}{1.5}} = \frac{6}{3} = 2 .$$

WEITERLESEN

- **Arithmetisches Mittel**
- **Geometrisches Mittel**

Machine Translated by Google

Bedeuten

Maß für die zentrale Tendenz

VERWEISE

Mitrinovic, DS (mit Vasic, PM): Analytische Ungleichungen. Springer, Berlin Heidelberg New York (1970)

Hutmatrix

Die Hutmatrix ist eine **Matrix ,** die in **der Regressionsanalyse** und **der Varianzanalyse verwendet wird.** Sie ist als die Matrix definiert, die Werte umwandelt von der beobachteten **Variablen** in Schätzungen ermittelt mit der Methode der **kleinsten Quadrate .** Wenn also eine lineare Regression in Matrixform durchgeführt wird, ist Yˆ der **Vektor** gebildet aus Schätzungen, die aus berechnet wurden Parameter der kleinsten Quadrate und **Y** ein Vektor von Beobachtungen in Bezug auf die **abhängige Variable ist,** dann ist Yˆ durch den Vektor **Y** multipliziert mit H gegeben , das heißt, Yˆ = **HY** wird in Y umgewandelt Ja.

GESCHICHTE

Die Hutmatrix **H** wurde 1972 von **Tukey und John Wilder** eingeführt . Ein Artikel von Hoaglin, DC und Welsch, RE (1978) gibt die Eigenschaften der Matrix **H** und auch viele Beispiele für ihre Anwendung an.

MATHEMATISCHE ASPEKTE

Betrachten Sie das folgende lineare Regressionsmodell

$$\mathbf{Y} = \mathbf{X} \cdot \ddot{\mathbf{y}} + \ddot{\mathbf{y}} .$$

Wo

Y ist ein *(n* × 1) **-Vektor** von Beobachtungen zur **abhängigen Variablen;**

X ist die **(n×p)Matrix** unabhängiger Variablen (es gibt p unabhängige Variablen);

ÿ ist der *(n* × 1)-Fehlervektor **und ;**

ÿ ist der *(p* × 1)-Vektor der zu schätzenden Parameter.

Die Schätzung ÿˆ des Vektors **ÿ** ist gegeben durch

$$\ddot{y}\hat{} = \mathbf{X} \cdot \mathbf{X}^{\ddot{y}1} \cdot \mathbf{X} \cdot \mathbf{Y} .$$

und wir können die geschätzten Werte Yˆ von **Y** berechnen , wenn wir ÿˆ kennen:

$$Y\hat{} = \mathbf{X} \cdot \ddot{y}\hat{} = \mathbf{X} \cdot \mathbf{X} \cdot \mathbf{X}^{\ddot{y}1} \cdot \mathbf{X} \cdot \mathbf{Y} .$$

Die Matrix **H** ist dann definiert durch:

$$\mathbf{H} = \mathbf{X} \cdot \mathbf{X} \cdot \mathbf{X}^{\ddot{y}1} \cdot \mathbf{X} .$$

Insbesondere wird das Diagonalelement hii definiert durch:

$$hii = xi \cdot \mathbf{X} \cdot \mathbf{X}^{\ddot{y}1} \cdot x_i .$$

wobei *xi* die i-te Linie von X ist.

BEREICHE UND EINSCHRÄNKUNGEN Die Matrix H, die es uns ermöglicht, *n* Schätzungen der **abhängigen Variablen** aus *n* Beobachtungen zu erhalten , ist eine idempotente symmetrische quadratische Matrix der Ordnung n. Das Element *(i,* j) dieser **Matrix** misst d j-te Beobachtung des i-ten vorhergesagten **Werts.** Insbesondere bewerten die Diagonalelemente die Auswirkungen der Beobachtungen auf die korre entsprechende Schätzungen der abhängigen Variablen. Der Wert jedes Diagonalelements der Matrix **H** liegt zwischen 0 und 1.

Schreiben H = (hij)fori, *j = 1,...,n,wir* haben die Beziehung:

$$hii = h2_{ii} + \sum_{\substack{j=1 \\ i=j}}^{N} h2_{ij} .$$

was auf der Grundlage der idempotenten Natur von H erhalten wird; mit anderen Worten: **H** = H2.

H

Machine Translated by Google

Daher $\mathrm{tr}(H) = \sum_{i=1}^{N} hii = p$ = Anzahl von zu schätzende Parameter.

Die Matrix **H** wird zur Ermittlung von Hebelpunkten in der Regressionsanalyse verwendet.

BEISPIELE

Betrachten Sie die folgende Tabelle, in der *Y* eine **abhängige Variable** ist , die sich auf die **unabhängige Variable** X bezieht:

X	Y
50	6
52	8
55	9
75	7
57	8
58	10

Das **Modell** der **einfachen linearen Regression** wird wie folgt in Matrixform geschrieben:

$$\mathbf{Y} = \mathbf{X} \cdot \ddot{\mathbf{y}} + \ddot{\mathbf{y}} .$$

Wo

$$\mathbf{Y} = \begin{pmatrix} 6 \\ 8 \\ 9 \\ 7 \\ 8 \\ 10 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 75 \\ 1 & 57 \\ 1 & 58 \end{pmatrix} .$$

$\ddot{\mathbf{y}}$ ist der (6 × 1) **-Fehlervektor** und $\ddot{\mathbf{y}}$ ist der (2 × 1) **-Parametervektor .**

Wir finden die **Matrix H** mit dem Ergebnis:

$$\mathbf{H} = \mathbf{X} \cdot \mathbf{X} \cdot \mathbf{X}^{\ddot{y}1} \cdot \mathbf{X} .$$

Durch schrittweise Matrixberechnungen erhalten wir:

$$(\mathbf{X} \cdot \mathrm{X})^{\ddot{y}1} = \frac{1}{2393} \cdot \begin{pmatrix} 20467 & \ddot{y}347 \\ \ddot{y}347 & 6 \end{pmatrix} = \begin{pmatrix} 8{,}5528 & \ddot{y}0{,}1450 \\ \ddot{y}0{,}1450 & 0{,}0025 \end{pmatrix} .$$

und schlussendlich:

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X} \cdot \mathrm{X})^{\ddot{y}1} \cdot \mathbf{X}$$

$$= \begin{pmatrix} \ddot{y} & & & & & & \ddot{y} \\ & 0{,}32 & 0{,}28 & 0{,}22 & \ddot{y}0{,}17 & 0{,}18 & 0{,}16 & \\ & 0{,}28 & 0{,}25 & 0{,}21 & \ddot{y}0{,}08 & 0{,}18 & 0{,}16 & \\ & 0{,}22 & 0{,}21 & 0{,}19 & 0{,}04 & 0{,}17 & 0{,}17 & \\ & \ddot{y}0{,}17 & \ddot{y}0{,}08 & 0{,}04 & 0{,}90 & 0{,}13 & 0{,}17 & \\ & 0{,}18 & 0{,}18 & 0{,}17 & 0{,}13 & 0{,}17 & 0{,}17 & \\ \ddot{y} & 0{,}16 & 0{,}16 & 0{,}17 & 0{,}17 & 0{,}17 & 0{,}17 & \ddot{y} \end{pmatrix} .$$

Wir stellen beispielsweise fest, dass das Gewicht von y1, das bei der **Schätzung** von y^1 verwendet wird , 0,32 beträgt.

Anschließend überprüfen wir, **ob die Spur von He gleich** 2 ist, d. h. sie entspricht der Anzahl der Parameter des **Modells.**

$$\mathrm{tr}(H) = 0{,}32 + 0{,}25 + 0{,}19 + 0{,}90 + 0{,}17 + 0{,}17 = 2 .$$

WEITERLESEN

Hebelpunkt
Matrix
Multiple lineare Regression
Regressionsanalyse
Einfache lineare Regression

LITERATUR:

Belsley, DA, Kuh, E., Welsch, RE: Regressionsdiagnostik. Wiley, New York, S. 16–19 (1980)

Hoaglin, DC, Welsch, RE: Die Hutmatrix in Regression und ANOVA. Bin. Stat. **32,** 17–22 (und Korrektur bei **32,** 146) (1978)

Tukey, JW: Einige grafische und semigrafische Darstellungen. In: Bancroft, TA (Hrsg.) Sta tistical Papers in Honor of George W. Snedecor. Iowa State University Press, Ames, IA, S. 293–316 (1972)

Histogramm

Das Histogramm ist eine **grafische Darstellung** der Verteilung der **Daten**

Machine Translated by Google

in Klassen eingeteilt. Es besteht aus einer Reihe von Rechtecken und ist eine Art Häufigkeitsdiagramm. Jeder Datenwert wird sortiert und in einem geeigneten Klassenintervall platziert . Die Anzahl der Datenwerte innerhalb jedes Klassenintervalls bestimmt

die **Häufigkeit** (oder relative Häufigkeit) dieses Klassenintervalls.

Jedes Rechteck im Histogramm stellt eine Datenklasse dar. Die Breite des Rechtecks entspricht der Breite des Klassenintervalls und die Oberfläche des Rechtecks repräsentiert das Gewicht der Klasse.

GESCHICHTE Der Begriff Histogramm wurde erstmals 1895 von **Pearson, Karl** verwendet.
Siehe auch **grafische Darstellung.**

MATHEMATISCHE ASPEKTE

Der erste Schritt bei der Erstellung eines Histogramms besteht in der Darstellung der **Daten** in Form einer **Häufigkeitstabelle.**
Dies erfordert, dass die Klassenintervalle **festgelegt** und die Datenwerte sortiert und sortiert werden

in den Klassen platziert, was es ermöglicht, die **Häufigkeiten** der Klassen zu berechnen . Anschließend werden die Klassenintervalle und Häufigkeiten zur **Häufigkeitstabelle hinzugefügt.** Anschließend verwenden wir die Häufigkeitstabelle, um das Histogramm zu erstellen. Wir unterteilen die horizontale Achse des Histogramms in Intervalle, wobei die Breiten dieser Intervalle denen der Klassenintervalle entsprechen. Anschließend zeichnen wir die Rechtecke auf das Histogramm.
Die Breite jedes Rechtecks entspricht der Breite der Klasse, der es entspricht. Die Höhe des Rechtecks ist so, dass die Oberfläche des Rechtecks gleich der relativen Häufigkeit der entsprechenden Klasse ist.

Die Summe der Flächen des Rechtecks

gles muss gleich 1 sein.

DOMÄNEN UND EINSCHRÄNKUNGEN

Histogramme werden verwendet, um einen **Datensatz** in einer visuellen Form darzustellen, die leicht verständlich ist. Sie ermöglichen die Visualisierung und Extraktion bestimmter allgemeiner Merkmale (z. B. typischer **Werte,** des Bereichs oder der Form der Daten).
Durch die Überprüfung der Form eines Histogramms können wir das **Wahrscheinlichkeitsmodell** erkennen, dem die Daten folgen **(Normalverteilung, logarithmische Normalverteilung** usw.).
Mit einem Histogramm ist es auch möglich, unerwartetes Verhalten oder abnormale **Werte** zu erkennen .
Diese Art der **grafischen Darstellung** wird am häufigsten in den Wirtschaftswissenschaften verwendet, da es sich jedoch um eine äußerst einfache Möglichkeit zur Visualisierung eines Datensatzes handelt , wird sie auch in vielen anderen Bereichen verwendet.
Wir können die relative **Häufigkeit** auch in einem Histogramm darstellen. In diesem Fall entspricht die Höhe jedes Rechtecks der relativen Häufigkeit der entsprechenden Klasse dividiert durch die Länge des **Klassenintervalls.** Wenn wir in diesem Fall die Oberflächen aller Rechtecke im Histogramm summieren, erhalten wir Einheit.

BEISPIELE

Die folgende Tabelle enthält Rohdaten **zum** durchschnittlichen Jahresniederschlag in 69 Städten der USA. Wir werden diese Daten zur Feststellung verwenden

eine **Häufigkeitstabelle** und dann ein entsprechendes Histogramm.

Jährliche durchschnittliche Niederschläge in 69 Städten der USA (in Zoll)

Mob. 67,0	Schick. 34,4	St.L. 35.9	
Heilig. 17,2	Loui. 43,1	At.C. 45,5	
Wash. 38,9	Detr. 31,0	Char. 42.7	
Boise 11,5	KC 37,0	Col. 37,0	
Welche. 30,6	Konz. 36.2	Prov. 42,8	
Bost. 42,5	NY 40,2	Dall. 35.9	
Jack. 49,2	Klev. 35,0	Nord. 44.7	
Reno 7.2	Pitt.	36,2	Chey. 14.6

Machine Translated by Google

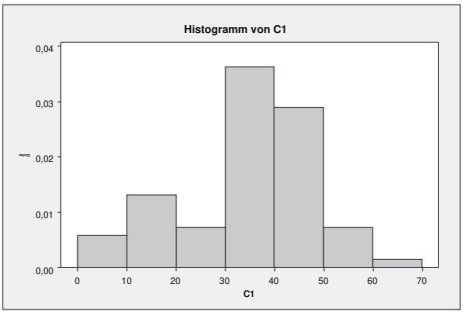
Polieren.	36,1 Nash. 46,0 LR		48,5
Cinc.	39,0 Wurzelholz.	32,5 Hart.	43.4
Phil.	39,9 Milw.	29,1 Atl.	48.3
Memp. 49,1 Phoen. 7,0 Ind.			38.7
SL	15.2 Denv. 13,0 Port.		40.8
Verkohlen.	40,8 Miami 59,8 Dul.		30.2
Jun.	54,7 Peor.	35,1 G.Fls 15,0	
SF	20,7 N.Or. 56,8 Albq		7.8
Jack.	54,5 SSM 31,7 Ral.		42,5
Sprach. 17,4 LA		14,0 Omaha 30,2	
Ok.C. 31,4 Wilm. 40,2 Alb.			33.4
Col.	46,4 Hon. 22,9 Bism. 16.2		
El Paso 7,8 D.Mn. 30,8 Hafen.			37.6
S-Tac 38,8 Balt.		41,8 S.Fls	24.7
SJ	59,2 Mn-SP 25,9 Hstn. 48.2		

Quelle: US Census Bureau (1975) Statistical Abstract of the United States. US-Volkszählung Büro, Washington, D.C.

Diese Daten können durch die folgende **Häufigkeitstabelle dargestellt werden:**

Klassenhäufigkeit		Relative Häufigkeit
0–10	4	0,058
10–20	9	0,130
20–30	5	0,072
30–40	24	0,348
40–50	21	0,304
50–60	4	0,058
60–70	2	0,030
Gesamt	69	1.000

Wir können nun das Histogramm erstellen:



Die horizontale Achse ist in Klassen unterteilt, und in diesem Fall sind die relativen **Häufigkeiten** gegeben durch die Höhe der Rechtecke, weil Die Klassen haben alle die gleiche Breite.

WEITERLESEN

- **Häufigkeitsverteilung**
- **Frequenzpolygon**
- **Häufigkeitstabelle**
- **Grafische Darstellung**
- **Ogive**
- **Stamm-Blatt-Diagramm**

VERWEISE

Dodge, Y.: Einige Schwierigkeiten bei der nichtparametrischen Schätzung einer Dichtefunktion. J. Offic. Stat. 2(2), 193–202 (1986)

Freedman, D., Pisani, R., Purves, R.: Statistik. Norton, New York (1978)

Pearson, K.: Beiträge zur mathematischen Evolutionstheorie. II: Schräglage Variation in homogenem Material. In: Karl Pearsons frühe statistische Arbeiten. Cambridge University Press, Cambridge, S. 41–112 (1948). Erstveröffentlichung in 1895 in Philos. Trans. Roy. Soc. London. Ser. A **186,** 343–414

Homogenitätstest

Ein Thema, das oft berücksichtigt werden muss bei der Analyse der erhaltenen **kategorialen Daten** für viele Gruppen ist die Homogenität der Gruppen. Mit anderen Worten, wir brauchen um herauszufinden, ob es signifikante gibt Unterschiede zwischen diesen Gruppen in Bezug auf eine oder mehrere qualitative kategoriale Variablen. Ein Homogenitätstest kann zeigen ob die Unterschiede signifikant sind oder nicht.

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Wir betrachten den Chi-Quadrat-Homogenitätstest
Hier handelt es sich um eine spezielle Art von **Chi-Quadrat prüfen.**

Sei ich die Anzahl der betrachteten Gruppen und
J die Anzahl der berücksichtigten Kategorien.
Lassen:

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$$ entsprechen der Gruppengröße

$$n_{\cdot j} = \sum_{i=1}^{I} n_{ij};$$

$$n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$$ sei die Gesamtzahl von Beobachtungen

nij sei die empirische **Häufigkeit** (d. h. die Anzahl der beobachteten Vorkommnisse) entsprechend Gruppe *i* und Kategorie j;

mij ist die theoretische Häufigkeit entsprechend der Gruppe *i* und der Kategorie j, die unter der Annahme von Homogenität unter den Gruppen, gleich:

$$m_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N}.$$

Wenn wir die Daten in Form einer **Kontingenztabelle** mit *I* Zeilen und *J* Spalten darstellen,
wir können das *ni berechnen.* die dazu beitragen
Summe der Elemente der Zeile *i* und des *nj* das tragen zur Summe aller Elemente bei
der Spalte j.
Wir berechnen

$$\ddot{y}2_c = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}\ddot{y}\, m_{ij2}}{m_{ij}}.$$

und der Chi-Quadrat-Homogenitätstest ist
so ausgedrückt: Wir lehnen das ab
Homogenitätshypothese (bei einer Signifikanz
Niveau von 5%), wenn der Wert $\ddot{y}2_c$ ist dann größer
der Wert der ÿ2 (Chi-Quadrat)-Verteilung
mit *(J* ÿ 1) · *(I* ÿ 1) Freiheitsgraden.

Hinweis: Wir haben hier dasselbe angenommen
Für jede Kombination wird eine bestimmte Anzahl an Einheiten getestet
Zuordnung von Gruppe und Kategorie. Allerdings können wir
Ich möchte unterschiedliche Anzahlen von Einheiten für Diff testen
verschiedene Kombinationen. In diesem Fall, wenn ja
Um einen Anteil *pij* an Einheiten für Gruppe *i* und Kategorie
j festzulegen, reicht es aus, *mij* durch *ni zu ersetzen. · pij.*

BEISPIELE

Wir betrachten eine im pharmazeutischen Bereich durchgeführte Studie, an der 100 Personen beteiligt sind, die an einer bestimmten Krankheit leiden. In um die Wirkung eines Arzneimittels zu untersuchen
Für die Behandlung wurden 100 Personen nach dem Zufallsprinzip ausgewählt. Die Hälfte von ihnen wurde einer Kontrollgruppe zugeordnet und erhielt ein Placebo.
Die anderen Patienten erhielten die medizinische Behandlung.
Dann die Anzahl der gesunden Menschen in jedem
Die Gruppe wurde anschließend 24 Stunden lang überwacht
Verabreichung der Behandlung. Die Ergebnisse sind
finden Sie in der folgenden Tabelle.

Beobachtet Frequenz	Gesund für 24 Std.	Nicht gesund	Gesamt
Placebo	2	48	50
Behandlung	9	41	50

Die theoretischen Frequenzen erhält man durch
vorausgesetzt, dass der allgemeine Gesundheitszustand
wäre für beide Gruppen gleich gewesen, wenn
Es wurde keine Behandlung durchgeführt. In diesem Fall
wir erhalten m11 = m21 = = *5,5* und $\frac{50\cdot 11}{100}$
m12 = m22 = $\frac{50\cdot 89}{100}$ 100 = *44,5.*
Den Wert von ÿ2 berechnen wir durch Vergleich
die theoretischen Frequenzen mit den beobachteten
Frequenzen:

$$\ddot{y}2_c = \frac{(2\ddot{y}\,5{,}5)^2}{5.5} + \frac{(48\ddot{y}\,44{,}5)^2}{44{,}5} + \frac{(9\ddot{y}\,5{,}5)^2}{5{,}5} + \frac{(41\ddot{y}\,44{,}5)^2}{44{,}5}$$

$$= 5{,}005.$$

Machine Translated by Google

Wenn wir uns dann auf die ÿ2- Verteilungstabelle für einen Freiheitsgrad beziehen, erhalten wir den Wert $ÿ^2_{0,05} = 3,84$ für ein Signifikanzniveau von 5 %, Dies ist kleiner als der von uns berechnete Wert, $ÿ^2 = 5,005$. Wir kommen zu dem Schluss, dass die Gruppen nicht homogen waren und die Behandlung effizient ist.

WEITERLESEN

Analyse kategorialer Daten
Varianzanalyse
Kategoriale Daten
Chi-Quadrat-Verteilung
Chi-Quadrat-Test
Frequenz

REFERENZ

Siehe **Analyse kategorialer Daten.**

Hotelling, Harold

Hotelling, Harold (1895–1973) gilt als einer der Pioniere auf dem Gebiet der Wirtschaftsmathematik im Zeitraum 1920–1930. Er führte den multivariaten T2- Test, die Hauptkomponentenanalyse und die kanonische Korrelationsanalyse ein.

Er studierte an der University of Washington, wo er 1919 einen BA in Journalismus erhielt. Anschließend wechselte er an die Princeton University

Dort promovierte er 1924 in Mathematik. Im selben Jahr begann er seine Lehrtätigkeit an der Stanford University. Seine Anwendungen der Mathematik in den Sozialwissenschaften betrafen zunächst Journalismus und Politikwissenschaft, dann verlagerte er seinen Schwerpunkt auf Bevölkerung und Vorhersage.
1931 wechselte er an die Universität Kolumbien, wo er sich aktiv am Aufbau der Statistikabteilung beteiligte. Während der Während des Zweiten Weltkriegs führte er statistische Untersuchungen für das Militär durch.
1946 wurde er von der North Carolina University in Chapel Hill angestellt, um dort eine Statistikabteilung aufzubauen.

Einige Hauptwerke und Artikel von Hotelling, Harold:

1933 Analyse eines Komplexes statistischer Variablen mit Hauptkomponenten. J. Educ. Psychol., 24, 417–441 und 498–520.

1936 Beziehung zwischen zwei Mengen von Variablen. Biometrie 28, 321–377.

Huber, Peter J.

Huber, Peter J. wurde 1934 in Wohlen (Schweiz) geboren. Während seines Studiums und seiner Promotion in Mathematik am Eidgenössischen Polytechnikum Zürich erbrachte er hervorragende Leistungen, wo er für die wissenschaftliche Qualität seiner Dissertation die Silbermedaille erhielt. Er arbeitete als Professor für Mathematische Statistik Tics an der Eidgenössischen Polytechnischen Schule Zürich. Anschließend zog er in die USA und arbeitete als Gastprofessor an den renommiertesten Universitäten (Princeton, Yale, Berkeley). 1977 wurde er zum Professor ernannt Angewandte Mathematik am Massachusetts Institute of Technology. Er ist Mitglied der renommierten American Academy of Arts and Sciences, der Bernoulli Society und der National Science Foundation in den USA, in denen ausländische Mitglieder äußerst selten sind.
Seit der Veröffentlichung seines Artikels „Robuste Schätzung eines Standortparameters“ im Jahr 1964 gilt er als Begründer von robuste Statistiken.

Machine Translated by Google

Huber, Peter J. erhielt den Titel Docteur HonorisCausa der Universität Neuchâtel 1994.

Einige Hauptwerke und Veröffentlichungen von Huber, Peter J.:

1964 Robuste Standortschätzung Parameter. Ann. Mathematik. Stat. 35, 73–101.

1968 Robuste Statistik Verfahren. SIAMCBMS-NSF-Reg. Konf. Ser. Appl. Mathematik.,

1981 Robuste Statistik. Wiley, New York.

1995 Robustheit: Wo stehen wir jetzt? Student, Band 1, 75–86.

Hypergeometrische Verteilung

Die hypergeometrische Verteilung beschreibt die **Erfolgswahrscheinlichkeit ,** wenn eine Reihe von Objekte werden aus einer Population gezogen (die enthält einige Objekte, die einen Fehler darstellen während die anderen für Erfolg stehen), ohne Ersatz.
Es wird daher zur Beschreibung eines **Zufalls** verwendet **Experiment ,** bei dem es nur zwei mögliche Ergebnisse gibt: „Erfolg“ und „Misserfolg“.

Betrachten Sie eine Menge von *N* **Ereignissen ,** die vorhanden sind *M* „Erfolge“ und *N* ÿ *M* „Misserfolge“. Der **Zufallsvariable** X, entsprechend der Anzahl der erzielten Erfolge, wenn wir *n* ziehen Ereignisse ohne Ersatz folgen einer hypergeometrischen Verteilung mit den **Parametern** N, *M* und n, bezeichnet mit *H(N, M,* n).
Die hypergeometrische Verteilung ist eine **diskrete Wahrscheinlichkeitsverteilung.**

Die Anzahl der Möglichkeiten, die *n* Ereignisse haben können aus *N* Ereignissen gezogen ist gleich:

$$C_{nN} = \binom{N}{N} = \frac{N!}{N! \cdot (n \ddot{y} N)!} .$$

Die Anzahl der Elementarereignisse **hängt** davon ab auf *X* und ist:

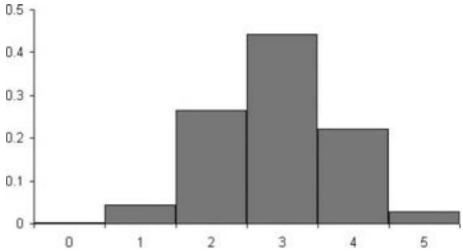
$$C_{xM} \cdot C_{n\ddot{y}x\,N\ddot{y}M} .$$

was **die folgende Wahrscheinlichkeitsfunktion** ergibt :

$$P(X = x) = \frac{C_{xM} \cdot C_{n\ddot{y}x\,N\ddot{y}M}}{C_{nN}} ,$$

für $x = 0, 1,...,n$

(wobei $Cu_v = 0$, wenn $u <$ vgemäß Konvention) .



Hypergeometrische Verteilung, N = 12, M = 7, n = 5

MATHEMATISCHE ASPEKTE

Betrachten Sie die **Zufallsvariable** $X = X1 + X2 + ... + Xn,$ wobei:

$$Xi = \begin{cases} 1, \text{ wenn die i-te Ziehung erfolgreich ist} \\ 0, \text{ wenn die i-te Zeichnung fehlgeschlagen ist} \end{cases} .$$

In diesem **Fall ist die Wahrscheinlichkeitsverteilung für** *Xi* ist:

Xi	0	1
P (Xi)	$\frac{N \ddot{y} M}{N}$	$\frac{M}{N}$

Damit ist der **Erwartungswert** von *Xi* angegeben von:

$$E[Xi] = \sum_{j=1}^{2} xjP(xi = xj)$$

$$= 0 \cdot \frac{N \ddot{y} M}{N} + 1 \cdot \frac{M}{N}$$

$$= \frac{M}{N} .$$

Machine Translated by Google

Unter Ausnutzung der Tatsache, dass X = *X1+X2+...+Xn,* wir haben:

$$E[X] = \sum_{i=1}^{N} E[Xi] = \sum_{i=1}^{N} \frac{M}{N} = n\,\frac{M}{N}.$$

Die **Varianz** von *Xi* ist per Definition:

$$\mathrm{Var(Xi)} = E\,X2 \;\ddot{y}\; (E[Xi])^2 = \frac{M}{N} \;\ddot{y}\; \left(\frac{M}{N}\right)^2 = \frac{M(N \;\ddot{y}\; M)}{N2}.$$

Da die *Xi, i = 1, 2,...,n* abhängig sind **Zufallsvariablen** sollte die **Kovarianz** sein bei der Berechnung berücksichtigt werden **Varianz** von X.

Die **Wahrscheinlichkeit ,** dass *Xi* und *Xj (i* = j) sind beide Erfolge sind gleich:

$$P(Xi = 1, Xj = 1) = \frac{M(M \;\ddot{y}\; 1)}{N(N \;\ddot{y}\; 1)}.$$

Wenn wir *V = Xi ·Xj setzen,* sind die **Werte** von *V* und die Zugehörige **Wahrscheinlichkeiten** sind:

V	0	1
P (V)	$1 \;\ddot{y}\; \frac{M(M \;\ddot{y}\; 1)}{N(N \;\ddot{y}\; 1)}$	$\frac{M(M \;\ddot{y}\; 1)}{N(N \;\ddot{y}\; 1)}$

Der **erwartete Wert** von *V* ist daher:

$$E[V] = 0 \cdot 1 \;\ddot{y}\; \frac{M(M \;\ddot{y}\; 1)}{N(N \;\ddot{y}\; 1)} + 1 \cdot \frac{M(M \;\ddot{y}\; 1)}{N(N \;\ddot{y}\; 1)}.$$

Die **Kovarianz** *von Xi* und Xj ist per Definition: gleich:

$$\mathrm{Cov(Xi,}\ Xj) = E[Xi \cdot Xj] \;\ddot{y}\; E[Xi] \cdot E[Xj] = \frac{M(M \;\ddot{y}\; 1)}{N(N \;\ddot{y}\; 1)} \;\ddot{y}\; \left(\frac{M}{N}\right)^2 = \;\ddot{y}\; \frac{M(N \;\ddot{y}\; M)}{N2(N \;\ddot{y}\; 1)} = \;\ddot{y}\; \frac{1}{N \;\ddot{y}\; 1}\,\mathrm{Var(Xi)}.$$

Wir können nun die **Varianz** von X berechnen :

$$\begin{aligned}\mathrm{Var(X)} &= \sum_{i=1}^{N} \mathrm{Var(Xi)} + 2 \sum_{j=1\ i<j}^{N} \mathrm{Cov(Xi,}\ Xj) \\ &= \sum_{i=1}^{N} \mathrm{Var(Xi)} + n(n \;\ddot{y}\; 1)\mathrm{Cov(Xi,}\ Xj) \\ &= n\,\mathrm{Var(Xi)} \;\ddot{y}\; \frac{n \;\ddot{y}\; 1}{N \;\ddot{y}\; 1}\,\mathrm{Var(Xi)} \\ &= n\mathrm{Var(Xi)}\;\frac{N \;\ddot{y}\; n}{N \;\ddot{y}\; 1} \\ &= n\,\frac{M(N \;\ddot{y}\; M)}{N2}\,\frac{N \;\ddot{y}\; n}{N \;\ddot{y}\; 1} \\ &= \frac{N \;\ddot{y}\; n}{N \;\ddot{y}\; 1}\,N\frac{M}{N}\,\frac{1 \;\ddot{y}\; \frac{M}{N}}{}.\end{aligned}$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die hypergeometrische Verteilung ist häufig in der Qualitätskontrolle eingesetzt. Angenommen, eine Produktionslinie produziert *N* Produkte, die dann einer Verifizierung unterzogen werden. Daraus wird eine **Stichprobe** der Größe *n* entnommen Produktcharge und die Anzahl der fehlerhaften Produkte in dieser Probe werden notiert. Es ist möglich, dies zu verwenden, um (durch **Schlussfolgerung)** zu erhalten: Angaben zur voraussichtlichen Gesamtzahl fehlerhafte Produkte in der gesamten Charge.

BEISPIELE

Eine Box enthält 30 Sicherungen, davon 12 defekt. Wenn wir zufällig fünf Sicherungen nehmen, sind die **Wahrscheinlichkeit ,** dass keiner von ihnen defekt ist, ist gleich:

$$P(X = 0) = \frac{Cx\ MOnyhx}{Cn_N} = \frac{C0\ 12C5\ 18}{C5_{30}} = 0{,}0601.$$

Machine Translated by Google

WEITERLESEN

Bernoulli-Verteilung

Binomialverteilung

Diskrete Wahrscheinlichkeitsverteilung

Hypothese

Eine statistische Hypothese ist eine Aussage über die Verteilung(en) einer oder mehrerer **Zufallsvariablen.** Dabei kann es sich um die **Parameter** einer gegebenen Verteilung oder die **Wahrscheinlichkeitsverteilung** einer untersuchten **Population** handeln.

Die Gültigkeit der Hypothese wird überprüft, indem **Hypothesentests** anhand von **Beobachtungen** durchgeführt werden, die für eine **Stichprobe** der untersuchten **Population gesammelt wurden.**

Bei der Durchführung von **Hypothesentests** zur **Wahrscheinlichkeitsverteilung** der untersuchten **Grundgesamtheit** wird die Hypothese, dass die untersuchte Grundgesamtheit einer bestimmten Wahrscheinlichkeitsverteilung folgt, als **Nullhypothese bezeichnet.** Die Hypothese, die bestätigt, dass die Grundgesamtheit keiner gegebenen Wahrscheinlichkeitsverteilung folgt, wird Alternativhypothese **(oder** Gegenhypothese) genannt. Wenn wir **Hypothesentests** für die **Parameter** einer **Verteilung durchführen, wird** die Hypothese, dass die untersuchten Parameter einem gegebenen **Wert** entsprechen , als **Nullhypothese bezeichnet.** Die Hypothese, die besagt, dass der Wert des Parameters von diesem vorgegebenen Wert abweicht, wird Alternativhypothese genannt .

Die **Nullhypothese** wird üblicherweise mit H0 und die **Alternativhypothese** mit H1 bezeichnet .

GESCHICHTE

Beim **Hypothesentest** wird die zu testende Hypothese als **Nullhypothese bezeichnet.** Den Begriff „null“ verdanken wir **Fisher, RA** (1935). Als er dieses Konzept vorstellte, erwähnte er das bekannte Problem der Teeverkostung, bei dem eine Dame behauptete, am Geschmack erkennen zu können, ob die Milch oder der Tee zuerst in ihre Tasse gegossen wurde. Die zu prüfende Hypothese war, dass der Geschmack überhaupt nicht von der Reihenfolge der Teezubereitung beeinflusst wurde.

Ursprünglich wurde die Nullhypothese üblicherweise so verstanden, dass eine bestimmte **Behandlung** keine Wirkung hat oder dass es keinen Unterschied gibt Zusammenhang zwischen den Wirkungen verschiedener Behandlungen gen.

Heutzutage wird die Nullhypothese im Gegensatz zur **Alternativhypothese hauptsächlich verwendet, um anzugeben, dass die Hypothese getestet werden muss.** Siehe auch **Hypothesentests.**

BEISPIELE

Bei vielen Problemen geht es darum, ein **Experiment** zu wiederholen , das zwei mögliche Ergebnisse liefert. Ein Beispiel hierfür ist das Geschlecht eines neugeborenen Kindes. In diesem Fall interessiert uns der Anteil von Jungen und Mädchen in einer bestimmten **Bevölkerung.** Betrachten Sie p, den Anteil der Mädchen, den wir anhand einer beobachteten **Stichprobe** schätzen möchten . Um festzustellen, ob die Anteile neugeborener Jungen und Mädchen gleich sind, stellen wir die statistische Hypothese auf, dass $p =$

WEITERLESEN

Alternative Hypothese

Varianzanalyse

Hypothesentest

Nullhypothese

VERWEISE

Fisher, RA: Das Design von Experimenten. Oliver & Boyd, Edinburgh (1935)

Machine Translated by Google

Hypothesentest

Das Testen von Hypothesen ist ein Verfahren, das es uns (abhängig von bestimmten Entscheidungsregeln) ermöglicht, eine **Ausgangshypothese,** die sogenannte **Nullhypothese,** zu bestätigen oder diese Nullhypothese zugunsten der **Alternativhypothese abzulehnen.**

GESCHICHTE

Die Theorie hinter dem Hypothesentest wurde im Zuge der Untersuchung entwickelt. Die ersten Schritte wurden unternommen, als Arbeiten zu erscheinen begannen, die die Bedeutung (oder Bedeutungslosigkeit) einer Gruppe von Beobachtungen diskutierten. Einige Beispiele solcher Werke stammen aus dem 18. Jahrhundert, darunter die von Arbuth Nott, J. (1710), Bernoulli, Daniel (1734) und **Laplace, Pierre Simon de** (1773). Diese Werke wurden im 19. Jahrhundert häufiger gesehen, beispielsweise die von Gavarett (1840) und **Edgeworth, Francis Y.** (1885).

Die Entwicklung des Hypothesentests erfolgte parallel zur Schätztheorie . Das Testen von Hypothesen scheint zuerst von Forschern in den experimentellen Wissenschaften und im Managementbereich entwickelt worden zu sein. Beispielsweise wurde der **Student-Test von Gosset und William Sealy** während seiner Zeit bei Guinness entwickelt .

Neyman, JerzyandPearson und Egon Sharpe entwickelten die mathematische Theorie des Hypothesentests, die sie in einem 1928 in der Zeitschrift Biometrika veröffentlichten Artikel vorstellten. Sie waren die ersten, die erkannten, dass die rationale Entscheidung, die beim Hypothesentest getroffen werden muss, zwischen der **Nullhypothese,** die wir testen wollen, und einer **Alternativhypothese liegen muss.** Ein zweiter grundlegender Artikel über die Theorie des Hypothesentests wurde 1933 von denselben Mathematikern veröffentlicht, in dem sie ebenfalls zwischen einem **Fehler vom Typ I** und einem **Fehler vom Typ II unterschieden.**

Die aus der Zusammenarbeit zwischen Neyman, J. und Pearson, ES hervorgegangenen Werke werden in Pearson (1966) und in der von Reid (1982) veröffentlichten Biografie von Neyman beschrieben.

MATHEMATISCHE ASPEKTE

Das Testen einer Hypothese einer **Stichprobe** umfasst im Allgemeinen die folgenden Schritte: 1.

Formulieren Sie die Hypothesen: • Die **Nullhypothese** H0, • Die **Alternativhypothese** H1.

2. Bestimmen Sie das **Signifikanzniveau** ÿ des prüfen.
3. Bestimmen Sie die **Wahrscheinlichkeitsverteilung ,** die der **Stichprobenverteilung entspricht.**
4. Berechnen Sie den **kritischen Wert** der Null Hypothese und leiten Sie den **Ablehnungsbereich** oder den **Akzeptanzbereich ab.**
5. Legen Sie die Entscheidungsregeln fest:
   - Wenn die in der Stichprobe beobachteten **Statistiken** im Akzeptanzbereich liegen, lehnen wir die Nullhypothese H0 nicht ab;
   - Wenn die am Sam beobachteten Statistiken Befinden sich viele im Ablehnungsbereich, lehnen wir die Nullhypothese H0 für die Alternativhypothese H1 ab.
6. Treffen Sie die Entscheidung, die Nullhypothese auf der Grundlage der beobachteten Stichprobe zu akzeptieren oder abzulehnen.

BEREICHE UND EINSCHRÄNKUNGEN

Die häufigsten Arten von Hypothesentests werden im Folgenden beschrieben.

1. *Hypothesentest einer Stichprobe:* Wir wollen testen, ob der **Wert** eines **Parameters** ÿ der **Grundgesamtheit** mit einem vermuteten Wert identisch ist. Die Hypothesen lauten wie folgt:

$$H0 : \ddot{y} = \ddot{y}0 ,$$

$$H1 : \ddot{y} = \ddot{y}0 .$$

Machine Translated by Google

wobei ÿ0 der angenommene Wert von ist unbekannter Parameter ÿ.

2. *Hypothesentest an zwei Stichproben:* Die Ziel ist es in diesem Fall herauszufinden, ob zwei Populationen, die beide beschrieben werden von eines bestimmten Parameters sind unterschiedlich. Sei ÿ1 und ÿ2 sind Parameter, die die Populationen 1 bzw. 2 beschreiben. Dann können wir formulieren Sie die folgenden Hypothesen:

$$H0 : \ddot{y}1 = \ddot{y}2 ,$$
$$H1 : \ddot{y}1 = \ddot{y}2 .$$

oder

$$H0 : \ddot{y}1 \ddot{y} \ddot{y}2 = 0 \quad H1 : \quad ,$$
$$\ddot{y}1 \ddot{y} \ddot{y}2 = 0 .$$

3. *Hypothesentest von mehr als zwei Proben:* Wie bei einem Test, der an zwei durchgeführt wurde Proben werden Hypothesentests durchgeführt an mehr als zwei Proben zu bestimmen ob diese Populationen unterschiedlich sind, basierend auf dem Vergleich desselben Parameters aus allen getesteten Populationen.
Testen Sie in diesem Fall die folgende Hypothese ses:

$$H0 : \ddot{y}1 = \ddot{y}2 = \quad \cdots \quad = \ddot{y}k ,$$

H1 : Die Werte von $\ddot{y}i$ $(i = 1, 2,...,k)$ sind nicht alle identisch.

Dabei sind $\ddot{y}1,..., \ddot{y}k$ die unbekannten Parameter der Populationen und k die Anzahl der zu vergleichenden Populationen.

In der Hypothesentesttheorie und in der Praxis können wir zwischen zwei Typen unterscheiden von Tests: ein parametrischer Test und ein nichtparametrischer Test.

Parametrische Tests

Ein parametrischer Test ist ein Hypothesentest, der für jeden eine bestimmte Form voraussetzt Verteilungen im Zusammenhang mit den zugrunde liegenden Populationen. Dieser Fall liegt beispielsweise dann vor, wenn Diese Populationen folgen einer **Normalverteilung.**

Der Studententest ist ein Beispiel für einen parametrischen Test. Dieser Test vergleicht die Mittelwerte von zwei normalverteilte Populationen.

Nichtparametrischer Test

Ein nichtparametrischer Test ist ein Hypothesentest wobei eine Angabe nicht erforderlich ist parametrische Form der Verteilung der Grundgesamtheit.
Es gibt viele Beispiele dieser Art Test, einschließlich des **Zeichentests,** dem **Wilcoxon Test,** der **signierte Wilcoxon-Test,** der **Mann–Whitney- Test, der Kruskal-Wallis-Test** und der **Kolmogorov-Smirnov-Test.**

BEISPIELE

Beispiele für parametrische Hypothesentests finden Sie unter **Binomialtest, Fisher-Test** oder **Student-Test.** Beispiele für nichtparametrische Hypothesentests finden Sie bei **Kolmogorov–Smirnov-Test, Kruskal-Wallis-Test, Mann-Whitney-Test, Wilcoxon-Test, unterzeichneter Wilcoxon-Test** und **Vorzeichentest.**

WEITERLESEN

Akzeptanzregion
Alternative Hypothese
Nichtparametrischer Test
Nullhypothese
Einseitiger Test
Parametrischer Test
Ablehnungsbereich
Stichprobenverteilung
Signifikanzniveau
Zweiseitiger Test

VERWEISE

Arbuthnott, J.: Ein Argument für die göttliche Vorsehung, entnommen aus der konstanten Regelmäßigkeit bei Geburten beider Geschlechter beobachtet. Phi los. Trans. **27,** 186–190 (1710)

Machine Translated by Google

Bernoulli, D.: Quelle ist die Ursache der Neigung des Planeten (. . .). Empf. Stückbericht. Prix Acad. Roy. Wissenschaft. 3, 95–122 (1734)

Edgeworth, FY: Methoden der Statistik. Jubiläumsband der Royal Statistical Society, London (1885)

Gavarret, J.: Principes généraux de statis tiquemédicale. Beché Jeune & Labé, Paris (1840)

Laplace, PS de: Erinnerung an die moyenne inclinaison des orbites des comètes. Mém. Acad. Roy. Wissenschaft. Paris 7, 503–524 (1773)

Lehmann, EL: Testing Statistical Hypothe ses, 2. Aufl. Wiley, New York (1986)

Neyman, J., Pearson, ES: Zur Verwendung und Interpretation bestimmter Testkriterien für statistische Schlussfolgerungen, Teile I und II. Biometrie 20A, 175–240, 263–294 (1928)

Neyman, J., Pearson, ES: Zum Problem der effizientesten Tests von Statistiken Cal-Hypothesen. Philos. Trans. Roy. Soc. London. Ser. A **231,** 289–337 (1933)

Pearson, ES: Die Neyman-Pearson-Geschichte: 1926–34. In: David, FN (Hrsg.) Research Papers in Statistics: Festschrift für J. Neyman. Wiley, New York (1966)

Reid, C.: Neyman – Aus dem Leben. Springer, Berlin Heidelberg New York (1982)

Machine Translated by Google

Idempotente Matrix

Siehe **Matrix.**

Identitätsmatrix

Siehe **Matrix.**

Vorfall

Siehe **Inzidenzrate.**

Inzidenzrate

Die **Inzidenz** einer Krankheit wird definiert als

Dabei handelt es sich um die Zahl neuer Krankheitsfälle
Erscheinen während eines bestimmten Zeitraums
die Individuen einer **Population.** Diese Vorstellung
ähnelt dem von „stream“.

Die Inzidenzrate wird als relativ zur Bevölkerungsgröße definiert
die Zeit; Dieser Wert wird durch eine Relation ausgedrückt
auf eine Anzahl von Personen und auf eine Dauer.
Die Inzidenzrate ist die Inzidenz, *die ich* geteilt habe
durch die Anzahl der krankheitsgefährdeten Personen.

GESCHICHTE

Siehe **Prävalenzrate.**

MATHEMATISCHE ASPEKTE

Formal haben wir:

$$\text{Inzidenzrate} = \frac{\text{Risiko}}{\text{Dauer der Beobachtung}}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Eine genaue Inzidenzrate berechnet in
eine **Bevölkerung** und während eines bestimmten Zeitraums wird
als Index der Existenz eines interpretiert werden
Epidemie in dieser Bevölkerung.

BEISPIELE

Berechnen wir die **Inzidenzrate** der Brust
Krebs bei einer Population von 150.000 Frauen
beobachtet während 2 Jahren, gemäß der folgenden
Tabelle:

	Nummer von Fälle	**Nummer von Fächer**	**Vorfall Rate / 100000 / Jahr**
Gruppe	*[A]*	*[B]*	$[\frac{A}{2B}]$
Nicht ausgesetzt	40	35100	57,0
Passiv Raucher	140	55500	126.1
Aktiv Raucher	164	59400	138,0
Gesamt	344	150000	114,7

Der Zähler der Inzidenzrate ist
das gleiche wie das **Risiko:** die Zahl der Neuen

Machine Translated by Google

Fälle (oder Zwischenfälle) von Brustkrebs während der 2-jährigen Beobachtung in der Präferenzgruppe. Der Nenner der Inzidenzrate ergibt sich aus der Multiplikation der Anzahl der Personen in der Präferenzgruppe mit der Beobachtungsdauer (hier 2 Jahre). Normalerweise wird die Rate als 1 Jahr ausgedrückt.

WEITERLESEN

- **Zurechenbares Risiko**
- **Vermeidbares Risiko**
- **Ursache und Wirkung in der Epidemiologie**
- **Quoten und Quotenverhältnis**
- **Prävalenzrate**
- **Relatives Risiko**
- **Risiko**

VERWEISE

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Little Brown, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Morabia, A.: L'Épidémiologie Clinique. Ausgaben „Que sais-je?“. Presses Univer Sitaires de France, Paris (1996)

Unvollständige Daten

Siehe **fehlende Daten.**

Unabhängigkeit

Der Begriff der Unabhängigkeit kann zwei Bedeutungen haben: Die erste bezieht sich auf die **Ereignisse** eines **Zufallsexperiments.** Es werden zwei Ereignisse erwähnt unabhängig sein, wenn die Verwirklichung des zweiten Ereignisses nicht von der Verwirklichung des ersten abhängt.

Die Unabhängigkeit zwischen zwei **Ereignissen** *A und B* kann durch Wahrscheinlichkeiten bestimmt werden : Wenn die Wahrscheinlichkeit des Ereignisses *A* , wenn *B* eintritt, identisch ist mit der Wahrscheinlichkeit von A , wenn B nicht eintritt, dann werden A und B als unabhängig bezeichnet.

Die zweite Bedeutung des Wortes Unabhängigkeit betrifft die **Beziehung ,** die bestehen kann

zwischen **Zufallsvariablen.** Die Unabhängigkeit zweier Zufallsvariablen kann getestet werden

ed mit dem **Kendall-Test** und dem **Spearman**

prüfen.

GESCHICHTE

Der Unabhängigkeitsgedanke wurde implizit verwendet, lange bevor ein formaler Satz von **Wahrscheinlichkeitsaxiomen** aufgestellt wurde. Laut Maistrov, LE (1974) verwendete Cardano bereits die Multiplikationsregel der Wahrscheinlichkeiten *(P (A* ÿ B) = *P (A)* · *P* (B)). Maistrov, LE erwähnt auch, dass die Vorstellungen von Unabhängigkeit und **Abhängigkeit zwischen Ereignissen** Pascal, Blaise, de Fermat, Pierre und Huygens, Christiaan sehr vertraut waren.

MATHEMATISCHE ASPEKTE

Mathematisch können wir sagen, dass **Ereignis** *A* unabhängig von Ereignis *B* ist , wenn

$$P(A) = P(A \mid B),$$

Dabei ist *P(A)* die **Wahrscheinlichkeit** der Realisierung von Ereignis *A* und P(A|B) die **bedingte Wahrscheinlichkeit** von *Ereignis A* als Funktion der Realisierung *von B,* also die Wahrscheinlichkeit, dass *A* eintritt, obwohl man weiß, dass *B* eingetreten ist.

Unter Verwendung der Definition der **bedingten Wahrscheinlichkeit** kann die vorherige Gleichung geschrieben werden als:

$$P(A) = \frac{P(A \text{ ÿ } B)}{P(B)}$$

Machine Translated by Google

oder

$$P(A) \cdot P(B) = P(A \cap B),$$

was gibt

$$P(B) = P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

Mit dieser einfachen Transformation haben wir gezeigt, dass, wenn *A* unabhängig von B ist, auch *B* unabhängig von A ist. Wir werden sagen, dass **die Ereignisse** *A* und *B* unabhängig sind.

DOMÄNEN UND EINSCHRÄNKUNGEN

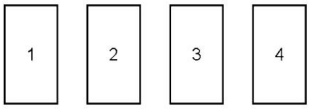
Wenn zwei **Ereignisse** *A* und *B* unabhängig sind, dann sind sie immer kompatibel. **Die Wahrscheinlichkeit ,** dass beide Ereignisse *A* und *B* gleichzeitig eintreten, lässt sich durch Multiplikation der Einzelwahrscheinlichkeiten der beiden Ereignisse ermitteln:

$$P(A \cap B) = P(A) \cdot P(B).$$

Wenn *A* und *B* jedoch abhängig sind, können sie kompatibel oder inkompatibel sein. Es gibt keine absolute Regel.

BEISPIELE

Stellen Sie sich eine Schachtel vor, die vier von 1 bis 4 nummerierte Notizen enthält, aus denen wir etwas machen werden zwei aufeinanderfolgende Zeichnungen, ohne die erste Note wieder in die Schachtel zu legen, bevor die zweite Note gezogen wird.



Der **Probenraum** kann durch die folgenden 12 Paare dargestellt werden:

= {(1, 2),(1, 3),(1, 4),(2, 1),(2, 3),
(2, 4),(3, 1),(3, 2),(3, 4), (4, 1),(4,
2),(4, 3)}.

Betrachten Sie die folgenden zwei **Ereignisse:**

A = „Notiz Nr. 3 auf der ersten Zeichnung herausziehen“

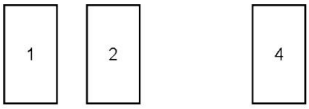
= {(3, 1),(3, 2),(3, 4)}

B = „Notiz Nr. 3 auf der zweiten Zeichnung herausziehen“

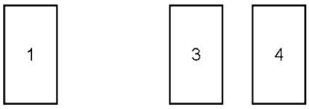
= {(1, 3),(2, 3),(4, 3)}

Sind diese beiden **Ereignisse** abhängig oder unabhängig? Wenn sie unabhängig sind, muss die **Wahrscheinlichkeit** von **Ereignis** *B* identisch mit der Wahrscheinlichkeit von Ereignis A sein, unabhängig vom Ergebnis von Ereignis A.

Wenn *A* passiert, bleiben die folgenden Notizen für die zweite Zeichnung im Feld:



Die **Wahrscheinlichkeit ,** dass *B* passiert, ist dann null, da sich Notiz Nr. 3 nicht mehr in der Box befindet. Wenn andererseits *A* nicht erfüllt ist (z. B. Anmerkung Nr. 2 wurde in der ersten Zeichnung gezeichnet), enthält die zweite Zeichnung die folgenden Anmerkungen im Feld:



Die **Wahrscheinlichkeit ,** dass *B* erfüllt wird, ist $\frac{1}{3}$. Wir können daher schlussfolgern, dass die Wahrscheinlichkeit des **Ereignisses B** vom Ereignis A abhängt oder dass die Ereignisse *A* und *B* abhängig sind. Wir können unsere Schlussfolgerung folgendermaßen überprüfen: Wenn *A* und *B* unabhängig sind, muss die folgende Gleichung überprüft werden:

$$P(A \cap B) = P(A) \cdot P(B).$$

In unserem Beispiel ist $P(A) = \frac{3}{12}$ und $P(B) = \frac{3}{12}$.

Machine Translated by Google

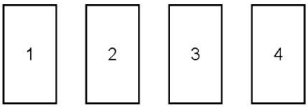
Der Schnittpunkt zwischen *A* und *B* ist null.
Deshalb haben wir

$$P(A \text{ ÿ } B) = 0 = P(A) \cdot \frac{1}{16} = P(B).$$

Wenn die Gleichung nicht bestätigt ist, dann
Ereignisse *A* und *B* sind nicht unabhängig. Deshalb
sind sie abhängig.
Betrachten Sie nun ein ähnliches Experiment, jedoch von
Legen Sie den zuvor gezogenen Zettel wieder in die Schachtel
die zweite Zeichnung.
Der **Probenraum** dieses neuen **Zufalls**
Das Experiment besteht aus den folgenden 16
Paare:

= {(1, 1),(1, 2),(1, 3),(1, 4),(2, 1),
(2, 2),(2, 3),(2, 4),(3, 1),(3, 2),
(3, 3),(3, 4),(4, 1),(4, 2),(4, 3),
(4, 4)} .

Unter diesen Bedingungen erfolgt die zweite Ziehung
aus einer Box erfolgen, die mit der von identisch ist
die erste Zeichnung:



Die Ereignisse A und B enthalten die folgenden Ergebnisse:

A = {(3, 1),(3, 2),(3, 3),(3, 4)},
B = {(1, 3),(2, 3),(3, 3),(4, 3)}.

Die Wahrscheinlichkeit sowohl *für A* als auch für B
fore Der Schnittpunkt zwischen *A* und *B* ist
beträgt 4
16 .
gegeben **durch das Ereignis AÿB** = {(3, 3)}; seine Wahrscheinlichkeit
Fähigkeit ist deshalb haben wir
ist 1
16 .

$$P(A \text{ ÿ } B) = \frac{1}{16} = P(A) \cdot P(B).$$

Da P(A ÿ B) = *P(A)* · *P(B),* können wir schlussfolgern,
dass **Ereignis** *B* unabhängig von Ereignis A ist,
oder dass die Ereignisse *A* und *B* unabhängig sind.

WEITERLESEN

- **Chi-Quadrat-Test der Unabhängigkeit**
- **Kompatibilität**
- **Bedingte Wahrscheinlichkeit**
- **Abhängigkeit**
- **Fall**
- **Kendall-Rangkorrelationskoeffizient**
- **Wahrscheinlichkeit**
- **Rangkorrelationskoeffizient nach Spearman**
- **Test der Unabhängigkeit**

VERWEISE

Maistrov, LE: Wahrscheinlichkeitstheorie – eine
historische Skizze (übersetzt und herausgegeben von Kotz, S.).
Akademisch, New York (1974)

Unabhängige Variable

In einem Regressionsmodell die Variablen, die sind
Es wird davon ausgegangen, dass sie die abhängigen
Variablen beeinflussen oder deren Variationen erklären.
sind alle unabhängige Variablen.

WEITERLESEN

- **Abhängige Variable**
- **Multiple lineare Regression**
- **Regressionsanalyse**
- **Einfache lineare Regression**

Indexnummer

In ihrer allgemeinsten Definition ist eine Indexzahl ein
Wert, der die relative Variation einer **Variablen**
zwischen zwei bestimmten Werten darstellt
Perioden (oder Situationen).
Die **einfachen Indexzahlen** sind von den
zusammengesetzten Indexzahlen zu unterscheiden .

Machine Translated by Google

• **Einfache Indexzahlen** beschreiben die relative Änderung einer einzelnen **Variablen.** •

Zusammengesetzte Indexzahlen ermöglichen es, den Vergleich der Wertemenge , die mehrere **Variablen** in einer bestimmten Situation annehmen, mit einer einzigen Zahl zu beschreiben

in Bezug auf die Wertemenge derselben Variablen in einer Referenzsituation.

Die Referenzsituation definiert die Grundlage die Indexnummer. Wir können beispielsweise sagen, dass für das Referenzjahr (Basisjahr) 1980 eine bestimmte Indexzahl den Wert 120 hat 1982.

GESCHICHTE

Die ersten Studien mit Indexzahlen stammen aus dem frühen 18. Jahrhundert.

Im Jahr 1707 unternahm der anglikanische Bischof Fleetwood die Untersuchung der Preisentwicklung zwischen den Jahren 1440 und 1700. Diese Studie, deren Ergebnisse in seinem Werk *Chronicon Preciosum* (1745) vorgestellt werden, wurde zu einem ganz bestimmten Zweck durchgeführt: während der Gründung eines Colleges im Jahr 1440– 1460 wurde eine wesentliche Klausel eingeführt, die besagte, dass jedes aufgenommene Mitglied das College verlassen musste, wenn sein Vermögen 5 Pfund pro Jahr überstieg. Fleetwood wollte wissen, ob ein solches Versprechen unter Berücksichtigung der Preisentwicklung auch drei Jahrhunderte später noch gehalten werden könne.

Er betrachtete vier Produkte – Weizen, Fleisch, Getränke und Kleidung – und untersuchte die Entwicklung ihrer Preise. Am Ende seiner Arbeit kam er zu dem Schluss, dass 5 Pfund (£) in den Jahren 1440–1460 denselben Wert hatten wie 30 Pfund (£) im Jahr 1700.

Später, im Jahr 1738, untersuchte Dutot, C. die Wertminderung des Geldes in einem Werk mit dem Titel „ Réflexions *politiques sur les Finances et le Commerce“*. Dazu untersuchte er die Einkommen zweier Könige, Ludwig XII. und Ludwig XV. Die beiden Könige hatten folgende Einkünfte:

Ludwig XII.: £ 7650000 im Jahr 1515
Ludwig XV.: 100000000 £ im Jahr 1735

Um festzustellen, welcher der Herrscher über das höchste verfügbare Einkommen verfügte, notierte Dutot, C. die unterschiedlichen Preise einer bestimmten Anzahl von Waren zu den betrachteten Zeitpunkten: ein Huhn, ein Kaninchen, eine Taube, der Wert einer Tagesarbeit, usw.

Dutots Messung war die folgende:

$$Ich = \frac{P1}{P0},$$

Dabei sind P1 und P0 die Summen der notierten Preise zu den jeweiligen Terminen. Beachten Sie, dass diese Indexnummer nicht berücksichtigt wird.
Dutot kam zu dem Schluss, dass der Reichtum von LudwigXII. war niedriger als die von LudwigXV.

Der Wertverlust des Geldes war auch Gegenstand einer Studie von Carli und Gian Rinaldo. Im Jahr 1764 untersuchte der Astronomieprofessor aus Padoue die Preisentwicklung seit der Entdeckung Amerikas.

Als Referenz nahm er drei Produkte – Getreide, Wein und Speiseöl – und untersuchte die Preise für diese Produkte im Zeitraum 1500–1750.

Carlis Messung war die folgende:

$$Ich = \frac{1}{N} \cdot \frac{P1}{P0}.$$

Die erhaltene Indexzahl ist eine **Arithmetik** Mittelwert der relativen Preise für jedes beobachtete Produkt.
Im Jahr 1798 wurde ein von Sir George Shuckburgh Evelyn unterzeichnetes Werk mit dem Titel „ An *Account of Some Endeavors to Ascertain a Standard of Weight and Measure“* veröffentlicht. Dieser Vorläufer der Indexzahlen definiert eine Preisindexzahl für die Jahre 1050 bis 1800, indem er die Begriffe Referenzjahr (1550) und relative Preise einführt. Er auch

Machine Translated by Google

beinhaltete den Preis für bestimmte Dienstleistungen in seinem **Beobachtungen.**

Es ist Lowe, Joseph, der gesehen werden sollte, nach Kendall, MG (1977), als wahrer Vater der Indexzahlen. Sein 1822 veröffentlichtes Werk mit dem Titel *The present state of England* behandelte viele Probleme im Vergleich zu die Erstellung von Indexnummern. Er führte unter anderem den Begriff des Preises ein Nachdenken zu verschiedenen Zeitpunkten *(P1* und P0) von die zum ersten Termin (Q0) beobachteten Mengen. Die Lowe-Messung ist die folgende:

$$Ich = \frac{(P1 \cdot Q0)}{(P0 \cdot Q0)} .$$

Später würde diese Indexnummer bekannt werden als **Laspeyres-Index.**

In der zweiten Hälfte des 19. Jahrhunderts erzielten die Statistiker auf diesem Gebiet große Fortschritte. Im Jahr 1863 mit *„A Serious Fall in the Value Ermittlung des Goldgehalts und seine sozialen Auswirkungen Set Forth,* Jevons, WS untersuchte die Art von **bedeuten ,** dass es bei der Erstellung eines verwendet werden soll Indexnummer. Er empfahl die Verwendung von das **geometrische Mittel.**

Zwischen 1864 und 1880 drei deutsche Wissenschaftler, Laspeyres, E. (1864, 1871), Dro bisch, MW (1871) und Paasche, H. (1874), arbeitete an der Entwicklung der Preise für Sachgüter in Hamburg, ohne Berücksichtigung von Dienstleistungen. Bei allen drei handelt es sich um unüberlegte Berechnungsmethoden.

Der **Paasche-Index** wird nach folgender Formel ermittelt:

$$Ich = \frac{(P1 \cdot Q1)}{(P0 \cdot Q1)} .$$

Palgrave, RHI (1886) schlug vor, die relativen Preise anhand des **Gesamtwerts** zu betrachten besorgt gut; Diese Methode führt zur Berechnung Geben Sie die Indexnummer wie folgt ein:

$$Ich = \frac{P1 \cdot Q1 \cdot \frac{P1}{P0}}{(P1 \cdot Q1)} .$$

Es waren die Arbeiten von Statistikern wie z I. Fisher, der 1922 eine Indexnummer definierte ber, der seinen Namen trägt und das ausmacht **dasgeometrische Mittel desLaspeyres-Index** und der **Paasche-Index:**

$$Ich = \sqrt{\frac{(P1 \cdot Q0)}{(P0 \cdot Q0)} \frac{(P1 \cdot Q1)}{(P0 \cdot Q1)}} ,$$

während im gleichen Zeitraum Marshall, A. und Edgeworth, FW schlug Folgendes für Mula vor:

$$Ich = \frac{(P1\ (Q1 + Q0))}{(P0\ (Q1 + Q0))} .$$

Beachten Sie, dass es sich bei dieser Formel um den **Mittelwert** handelt der berechneten Gewichte.

Moderne statistische Studien rufen immer noch weitgehend auf auf den Indexzahlen, die gewesen sind hier definiert.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Berechnung von Indexzahlen ist üblich Eine Methode zur Beschreibung der Veränderungen wirtschaftlicher **Variablen.** Auch wenn das aktuellste Indexzahlen betreffen im Wesentlichen den Preis Variable und die Mengenvariable können sie auch für Werte, Qualitäten usw. berechnet werden. Indexzahlen werden im Allgemeinen mit dem Bereich Wirtschaft und Wirtschaft in Verbindung gebracht weit verbreitet in anderen Tätigkeitsbereichen.

WEITERLESEN

Zusammengesetzte Indexnummer
Fisher-Index

Laspeyres-Index
Paasche-Index

Einfache Indexnummer

Machine Translated by Google

VERWEISE

Carli, GR: Del valore etc. In: Opere Scelte di Carli, vol. Ich, S. 299. Custodi, Mailand (1764)

Drobisch,MW:ÜberMittelgrössen und die Anwendbarkeit derselben auf die Berech

nung des Steigens und Sinkens des Geldwertes. Berichte über die Verhandlungen des Königs. Sachs. Ges. Wiss. Leipzig. Mathe-Phy. Klasse, 23, 25 (1871)

Dutot, C.: Politische Überlegungen zu Finanzen und Handel. Vaillant und Nicolas Prevost, Den Haag (1738)

Evelyn, Sir GS: Ein Bericht über einige Bemühungen, einen Standard für Gewicht und Maß festzulegen. Philos. Trans. 113 (1798)

Fleetwood, W.: Chronicon Preciosum, 2. Aufl. Fleetwood, London (1707)

Jevons, WS: Ein schwerwiegender Rückgang des festgestellten Goldwertes und seine sozialen Auswirkungen

Dargelegt. Stanford, London (1863)

Kendall, MG: Die frühe Geschichte der Indexzahlen. In: Kendall, M., Plackett, RL (Hrsg.) Studies in the History of Statistics and Probability, vol. II. Griffin, London (1977)

Laspeyres, E.: Hamburger Warenpreise 1850–1863 und die kalifornisch-austra

lischen Geldentdeckung seit 1848. Jahrb. Natl. Stat. 3, 81–118, 209–236 (1864)

Laspeyres, E.: Die Berechnung einer mit höherer Preissteigerung. Jahrb. Natl. Stat. **16,** 296–314 (1871)

Lowe, J.: Der gegenwärtige Zustand Englands im Hinblick auf Landwirtschaft, Handel und Finanzen. Kelley, London (1822)

Paasche, H.: Über die Preisentwicklung der letzten Jahre nach den Hamburger Börsennotierungen. Jahrb. Natl. Stat. **23,** 168–178 (1874)

Palgrave, RHI: Währung und Wertstandard in England, Frankreich und Indien usw. (1886) In: IUP (1969) British Parliamentary Papers: Session 1886: Third Report and Final Report of the Royal Commission on the Depression in Trade and Industry , mit Beweisprotokoll und Anhängen, Anhang B. Irish University Press, Shannon

Indikator

Ein Indikator ist eine Statistik **(offizielle Statistik) ,** deren Ziel es ist, einen Hinweis auf den Zustand, das Verhalten und die sich ändernde Natur eines wirtschaftlichen oder politischen Phänomens während eines bestimmten Zeitraums zu geben. Ein Indikator muss auch über die Schwankungen der Werte informiere

Veränderungen in der Natur der beobachteten Phänomene und müssen dazu dienen, die Entscheidungsträger zu unterrichten sion-Macher.

Im genauesten Fall spiegelt ein Indikator den sozialen oder wirtschaftlichen Prozess wider und schlägt vor, welche Änderungen vorgenommen werden sollten. Es handelt sich um ein Alarmsignal, das Aufmerksamkeit erregt und als Handlungsaufforderung (Lotse) dient. Viele Indikatoren sind im Allgemeinen notwendig, um ein globales Spiegelbild gegebener Phänomene und der Politik zu liefern, die man bewerten möchte. Eine Reihe von Indikatoren liefert Informationen über dasselbe Thema und wird als Indikatorsystem Katzen.

Nicht alle Statistiken sind Indikatoren; Indikatoren muss:

1. Quantifizierbar sein, was streng genommen nicht messbar bedeutet.
2. Seien Sie eindimensional (vermeiden Sie Überschneidungen mit anderen Indikatoren).
3. Decken Sie die vorrangigen Ziele ab.
4. Bewerten Sie direkt die Leistung einer Richtlinie.

Machine Translated by Google

5. Wege zur Verbesserung aufzeigen.
6. Beziehen Sie sich auf die wichtigsten Aspekte des Systems und mit Veränderungen im politischen System veränderbar sein.
7. Seien Sie überprüfbar.
8. Seien Sie verständnisvoll miteinander.
9. Seien Sie so wirtschaftlich wie möglich.
10. Seien Sie jederzeit erreichbar.
11. Seien Sie in der allgemeinen Kneipe leicht verständlich lic.
12. Seien Sie für jeden als gültig erkennbar und verlässlich.

DOMÄNEN UND EINSCHRÄNKUNGEN

Konstruktion von Indikatoren

Indikatoren können, wie alle Maße sozialer Phänomene, nicht genau und erschöpfend sein
Die Konstruktion von Indikatoren erfordert philosophische Überlegungen (Wertebeurteilung).
undtechnische und wissenschaftliche Kenntnisse, aberes basiert ebenfalls auf Logik und spiegelt oft politische Anliegen wider.

Sammlung und Auswahl von Indikatoren

Um gute Indikatoren zu sammeln, ist ein geeigneter Wirtschafts-, Sozial- und Bildungsmodell
Systeme benötigt werden. Das Modell als allgemein
Das vom OCDE angepasste Modell ist ein systemisches Modell: In einer gegebenen Umgebung untersuchen wir die Ressourcen, Prozesse und Auswirkungen eines Systems. Aber die Wahl der in einem System vereinten Indikatoren der Indikatoren hängt auch von der Logistik ab Durchführbarkeit der Datenerhebung. Dies erklärt der große Unterschied zwischen den Indikatoren. Der Anzahl der für ein Gut notwendigen Indikatoren Steuerung eines wirtschaftlichen oder sozialen Systems ist schwer zu beurteilen. Wenn zu zahlreich, die Indikatoren verlieren an Bedeutung und es besteht die Gefahr, dass sie keine wichtige Rolle mehr spielen. Aber zusammen sie muss das zu bewertende System darstellen.
Bei einer zu geringen Anzahl besteht die Gefahr, dass das System kaputt geht ungültig und nicht repräsentativ.

Interpretation von Indikatoren

Die Interpretation von Indikatoren ist für das System ebenso wichtig wie deren Auswahl, Konstruktion und Erhebung. Die Bedeutung haben
dass die Indikatoren für das Politische sorgen können verantwortlich, für die Finanzierung und den Erfolg
Bei bestimmten Projekten müssen alle Vorsichtsmaßnahmen getroffen werden genommen, um die interpretative Integrität sicherzustellen, die Qualität und die Neutralität der Indica
tors.

Validierung von Indikatoren

Bevor Sie diese Statistiken als Indikatoren verwenden der Leistung, Qualität und Gesundheit von ein Wirtschafts-, Sozial- oder Bildungssystem, sie sollten validiert werden. Wenn die Werte von Indikatoren wissenschaftlich mit mathematischen Methoden und Instrumenten berechnet werden, wird ihr Einfluss auf das zu beschreibende System und ihre Bedeutung bestimmt Zuverlässigkeit als Grundlage für einen Plan von Wirkungsweise sind nicht immer wissenschaftlich belegt.
Wenn darüber hinaus eine Reihe von Indikatoren verwendet wird, Es ist noch lange nicht klar, wie diese Indica
Tore kombinieren ihre Wirkungen, um das zu erzeugen beobachtete Ergebnisse. Die Methoden zur Validierung Indikatoren müssen noch entwickelt werden.

BEISPIELE

Ökonomische Indikatoren

Dazu gehört das Bruttoinlandsprodukt (BIP), Staatseinnahmen, Verbraucherpreis Index, Erzeugerpreisindex, Arbeitslosigkeit Zinssatz, Gesamtbeschäftigung, langfristige und kurzfristige Zinssätze, Exporte und Importe von produzierte Güter und Zahlungsbilanz.

Bildungsindikatoren

Dazu gehören die Kosten für Bildung, die Aktive im Lehrsystem, die Erwartung einer akademischen Ausbildung, die geografischen Ungleichheiten beim Zugang zum Grundstudium, die

Machine Translated by Google

allgemeines Niveau der Wehrpflichtigen,
Bildungsniveau junger Menschen nach der weiterführenden Schule
Abschluss, die Vorteile eines Bachelors
Abschluss im Hinblick auf die Arbeitssuche,
Schulpflicht für Kinder von 2 bis
5 Jahre alt, und Bildungskosten für
Weiterbildung.

Wissenschafts- und Technologieindikatoren

Hierzu zählen die Bruttoinlandsausgaben
zu Forschung und Entwicklung (GDERD),
Pro-Kopf-GDERD, GDERD als Prozentsatz des BIP,
Gesamtausgaben für Personal für Forschung und
Entwicklung, Gesamtzahl der Hochschulabschlüsse,
Prozentsatz der GDERD-Finanzierung durch Unternehmen,
Prozentsatz des GDERD, der von der Regierung
finanziert wird, Prozentsatz des GDERD, der von der Regierung finanziert wird
im Ausland und nationale Patentanmeldungen.

WEITERLESEN

Offizielle Statistiken

VERWEISE

OCDE: OCDE und Inter-Indikatoren
nationaux de l'enseignement. Center
de la Recherche et l'innovation dans
l'enseignement, OCDE, Paris (1992)

OCDE: Hauptindikatoren der Wissenschaft
et de la technologie. OCDE, Paris (1988)

Trägheit

Trägheit ist ein Begriff, der vor allem in verwendet wird
Mechanik. Je mehr Trägheit ein Körper hat
um eine Achse, desto mehr Energie muss vorhanden sein
verwendet, um diesen Körper um diesen herum in Rotation zu versetzen
Achse. Die Masse des Körpers und die (im Quadrat)
Abstand vom Körper zur Achse verwendet werden
bei der Berechnung der Trägheit. Die Defini
Die Varianz **kann** hier erkannt werden, wenn
„Masse“ wird durch **„Frequenz“** und „Abstand“ durch
„Abweichung“ ersetzt.

MATHEMATISCHE ASPEKTE

Nehmen wir an, dass wir eine Wolke aus *n* Punkten haben,
die in einem Raum der Dimension p zentriert ist ; die **Trägheit**
Matrix *V* kann mit dem **coor** konstruiert werden
Werte und **Frequenzen** der Punkte.
Die **Eigenwerte** *k1,...,kp* (geschrieben in
absteigende Reihenfolge) der **Matrix** *V* sind die
Trägheit, die durch die *p* **Fakultätsachsen** der erklärt wird
Punktewolke. Ihre Summe entspricht der **Spur**
von V ist gleich der Gesamtträgheit der Wolke.
Wir werden zum Beispiel sagen, dass die erste
Faktorialachse (oder Trägheitsachse) Folgendes erklärt:

$$100 \cdot \frac{k1}{\sum_{i=1} ki} \text{ (in \%) der Trägheit } \cdot P$$

Die ersten beiden Trägheitsachsen beispielsweise
erklären:

$$100 \cdot \frac{k1 + k2}{\sum_{i=1} ki} \text{ (in \%) der Trägheit } \cdot P$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Begriff der Trägheit oder, genauer gesagt, der
Trägheitsmatrix einer Punktwolke verwendet wird,
unter anderem, um eine **Korrespondenzanalyse**
zwischen betrachteten Punkten durchzuführen.

WEITERLESEN

Korrespondenzanalyse
Distanz
Eigenwert
Faktorielle Achse
Frequenz
Trägheitsmatrix

Machine Translated by Google

Trägheitsmatrix

Betrachten Sie eine Punktwolke in einem Raum von *p* Maße. Wir nehmen an, dass diese Wolke existiert zentriert im Ursprung des Referenzsystems; Wenn nicht, reicht es aus, eine Änderung herbeizuführen Variablen durch Subtrahieren von jedem Punkt Zentrum der Wolke. Bezeichnen wir mit **A** das **Matrix** der Koordinaten multipliziert mit der Quadratwurzel der zugehörigen **Frequenz** jeder Punkt. Eine solche Matrix hat für *n* verschiedene Punkte die Dimension *(n×p).* Vermuten dass *p* kleiner oder gleich n ist; ansonsten

Wir verwenden die **Transponierte** der Matrix **A** als neue Koordinatenmatrix.

Wir erhalten die Trägheitsmatrix **V ,** indem wir **A** mit der Transponierten von A, **A** multiplizierer**V** ist eine symmetrische Matrix der Ordnung p.

Die **Spur** dieser Matrix V ist genau die Summe **Trägheit** (oder Gesamtvarianz).

MATHEMATISCHE ASPEKTE

Es seien *n* verschiedene Punkte *Xi* von *p* Komponenten

$$Xi = \text{xi1}, xi2,...,xip , \qquad i = 1, 2,...,n,$$

Jeder Punkt *Xi* hat eine **Frequenz** *fi (i = 1, 2,...,n).*

Wenn die Punkte nicht um den Ursprung zentriert sind, sollten wir dafür sorgen, dass die sich ändernden Variablen der Verschiebung zum Ursprung von entsprechen das Zentrum der Wolke:

$$X_{_} = Xi \ddot{y} Y \quad ,$$

wobei **Y** = y1, *y2,...,yp* mit

$$yj = \frac{1}{N} \cdot \sum_{i=1}^{N} fi \cdot xij \quad .$$

Der allgemeine Term der Trägheitsmatrix *V* lautet dann:

$$vij = \sum_{k=1}^{N} fk \cdot (xki \ddot{y} yi) \cdot xkj \ddot{y} yj .$$

In Matrixform ist **V** = **A** · A, wobei die Koeffizienten der Matrix **A** sind:

$$aki = fk \cdot \overline{(xki \ddot{y} yi)} .$$

Die **Gesamtträgheit** der Punktwolke beträgt erhalten durch Berechnung der **Spur** der Trägheitsmatrix V, die gleich ist:

$$\text{Itot} = \text{tr}(V) = \sum_{i=1}^{P} vii$$

$$= \sum_{i=1}^{P} \sum_{k=1}^{N} fk \cdot (xki \ddot{y} yi)^2 .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Trägheitsmatrix einer Punktwolke ist

Wird zur Durchführung einer **Korrespondenzanalyse** verwendet zwischen zwei betrachteten Punkten.

BEISPIELE

Berücksichtigen Sie die folgenden Punkte im Raum von drei Dimensionen:

(2; 0; 0,5)
(3; 1; 2)
(ÿ2;ÿ1; ÿ1)
(ÿ1;ÿ1;ÿ0,5)
(ÿ1;ÿ1;ÿ0,5)
(ÿ1;ÿ1;ÿ0,5)

Es gibt vier verschiedene Frequenzen gleich 1, 1, 1 und 3.

Das Zentrum der Wolke wird durch ihre Punkte gebildet ist gegeben durch:

$$y1 = \frac{2 + 3 \ddot{y} 2 + 3 \cdot (\ddot{y}1)}{6} = 0 \quad ,$$

$$y2 = \frac{1 \ddot{y} 1 + 3 \cdot (\ddot{y}1)}{6} = \ddot{y}0,5 \quad ,$$

$$y3 = \frac{0,5 + 2 \ddot{y} 1 + 3 \cdot (\ddot{y}0,5)}{6} = 0 .$$

Machine Translated by Google

Die neuen Koordinaten unserer vier verschiedenen Punkte werden vergeben von:

(2; 0,5; 0,5) (3; 1,5; 2) (ÿ2; *ÿ0,5;* ÿ1) (ÿ1; *ÿ0,5;* ÿ0,5) der Frequenz 3 ,

Daraus wird A, die Matrix der Koordinaten multipliziert mit der Quadratwurzel der Frequenz, berechnet:

$$\mathbf{A} = \begin{matrix} \ddot{y} \\ \\ \ddot{y} \end{matrix} \begin{matrix} 2 & \textit{0,5} & \textit{0,5} \\ 3 & \textit{1.5} & 2 \\ \ddot{y}2 & \textit{ÿ0,5} & \ddot{y}1 \\ \ddot{y} \; \ddot{y}3 & \overline{\ddot{y}3} & \overline{\ddot{y}3} \\ & \overline{2} & \overline{2} \end{matrix} \begin{matrix} \\ \ddot{y} \\ \\ \ddot{y} \end{matrix}$$

und V, die Trägheitsmatrix:

$$\mathbf{V} = \mathbf{A} \cdot \mathbf{A} = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} 20 & 8 & \textit{10.5} \\ 8 & \textit{3,5} & \textit{4,5} \\ \textit{10,5} & \textit{4,5} & 6 \end{matrix} \begin{matrix} \\ \\ \ddot{y} \; \ddot{y} \; . \end{matrix}$$

Die **Trägheit** der Punktwolke beträgt:

Itot = tr(V) = 20 + *3,5* + 6 = *29,5* .

WEITERLESEN

Korrespondenzanalyse
Frequenz
Matrix
Verfolgen
Transponieren
Varianz

REFERENZEN

Lagarde, J. de: Initiation à l'analyse des don nées. Dunod, Paris (1983)

Inferenz

Inferenz ist eine Form der Schlussfolgerung durch Induktion, die auf der Grundlage der gesammelten Informationen erfolgt eine **Stichprobe** unter dem Gesichtspunkt der Verallgemeinerung der Informationen auf die mit dieser Stichprobe verbundene **Bevölkerung .**

GESCHICHTE

Siehe **Inferenzstatistik.**

WEITERLESEN

Inferenzstatistik

Inferenzstatistik

Die inferenzielle oder deduktive Statistik ergänzt die **deskriptive Statistik ,** da das Ziel der meisten Forschungen nicht darin besteht, eine bestimmte Anzahl von Indikatoren für eine bestimmte Stichprobe festzulegen, sondern vielmehr eine bestimmte Anzahl von Parametern zu schätzen, die die mit der behandelten Stichprobe verbundene Population charakterisieren. In der Inferenzstatistik **werden** Hypothesen in den Parametern einer **Grundgesamtheit formuliert.** Diese Hypothesen werden dann auf der Grundlage von Beobachtungen überprüft, die an einer repräsentativen **Stichprobe** der Bevölkerung gemacht wurden.

GESCHICHTE Die Ursprünge der Inferenzstatistik fallen mit denen der Wahrscheinlichkeitstheorie zusammen und entsprechen den Werken von **Bayes, Thomas** (1763), **de Moivre, Abraham** (1718), **Gauss, Carl Friedrich** (1809) und **de Laplace, Pierre Simon** (1812). .
Zu dieser Zeit gab es zahlreiche Forscher auf dem Gebiet der inferenziellen Statistik

Tics. Wir erwähnen zum Beispiel die Arbeiten von **Galton, Francis** (1889) in Bezug auf die **Korrelation** sowie die Entwicklung von Hypothesentests, die hauptsächlich auf **Pearson, Karl** (1900) und **Gosset, William** zurückzuführen sind

Machine Translated by Google

Sealey, der als „Student“ (1908) bekannt war. **Neyman, Jerzy, Pearson, Egon Sharpe** (1928) und **Fisher, Ronald Aylmer** (1956) trugen ebenfalls wesentlich zur Entwicklung der Inferenzstatistik bei.

WEITERLESEN

Satz von Bayes
Beschreibende Statistik
Einschätzung
Hypothesentest
Maximale Wahrscheinlichkeit
Optimierung
Wahrscheinlichkeit
Probenahme
Statistiken
Umfrage

REFERENZEN

Bayes, T.: Ein Aufsatz zur Lösung eines Problems in der Zufallslehre. Philos. Trans. Roy. Soc. London. **53,** 370–418 (1763). Veröffentlicht auf Betreiben von Price, R., zwei Jahre nach seinem Tod. Neuveröffentlicht mit einer Biografie von Barnard, George A. im Jahr 1958 und in Pearson, ES, Kendall, MG: Studies in the History of Statistics and Probability. Griffin, London, S. 131–153 (1970)

Fisher, RA: Statistische Methoden und Wissenschaft tific Schlussfolgerung. Oliver & Boyd, Edinburgh (1956)

Galton, F.: Natürliche Vererbung. Macmillan, London (1889)

Gauss, CF: Theoria Motus Corporum Coelestium. Werke, 7 (1809)

Gosset, SW „Student“: Der wahrscheinliche Fehler eines Mittelwerts. Biometrie 6, 1–25 (1908)

Laplace, PS de: Théorie analytique des probabilités. Courcier, Paris (1812)

Moivre, A. de: Die Lehre vom Zufall: oder eine Methode zur Berechnung der Wahrscheinlichkeit von Ereignissen im Spiel. Pearson, London (1718)

Neyman, J., Pearson, ES: Zur Verwendung und Interpretation bestimmter Testkriterien für statistische Schlussfolgerungen, Teile I und II. Biometrie 20A, 175–240, 263–294 (1928)

Pearson, K.: Über das Kriterium, dass ein gegebenes System von Abweichungen vom Wahrscheinlichen im Fall eines korrelierten Variablensystems so beschaffen ist, dass vernünftigerweise angenommen werden kann, dass es durch Zufallsstichproben entstanden ist. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, S. 339–357. Erstveröffentlichung 1900 in Philos.Mag. (5. Ser) **50,** 157–17

Innenprodukt

Das innere Produkt von **x** und y, bezeichnet mit x ·**y,** ist die Summe der Produkte zweier Vektoren, Komponente für Komponente.

Geometrisch gesehen entspricht das innere Produkt zweier Vektoren der Dimension *n* im Raum von *n* Dimensionen der Zahl, die man durch Multiplikation der **Norm** eines Vektors mit denen der Projektion eines anderen erhält. Diese Definition entspricht der Aussage, dass das Skalarprodukt zweier Vektoren dadurch erhalten wird, dass man das Produkt ihrer Normen und Kosinuswerte des durch zwei Vektoren bestimmten Winkels bildet.

MATHEMATISCHE ASPEKTE

Seien **x** und **y** zwei Vektoren der Dimension n, gegeben durch ihre Komponenten:

$$\mathbf{x} = \begin{pmatrix} x1 \\ x2 \\ \vdots \\ xn \end{pmatrix} \quad \text{und } \mathbf{y} = \begin{pmatrix} y1 \\ y2 \\ \vdots \\ yn \end{pmatrix}.$$

Machine Translated by Google

Das Skalarprodukt von **x** und y, bezeichnet mit **x** · y, ist definiert durch:

$$\mathbf{x} \cdot \mathbf{y} = [x1\ x2\ \ldots xn] \cdot \begin{bmatrix} y1 \\ y2 \\ \vdots \\ yn \end{bmatrix},$$

$$\mathbf{x} \cdot \mathbf{y} = x1 \cdot y1 + x2 \cdot y2 + \ldots + xn \cdot yn,$$

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{N} xi \cdot yi.$$

Das Skalarprodukt eines Vektors **x** mit sich selbst entspricht seiner quadratischen **Norm** (oder Länge), bezeichnet mit **x** 2:

$$\mathbf{x}\ 2 = \mathbf{x} \cdot \mathbf{x} = \sum_{i=1}^{N} x2ch.$$

Mit Hilfe der Norm können wir das Skalarprodukt berechnen

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot \mathbf{y} \cdot \cos ÿ,$$

wobei ÿ der Winkel zwischen den Vektoren **x** und ist j.

Das Skalarprodukt zweier Vektoren hat folgende Eigenschaften: • Es

ist kommutativ, das heißt für zwei Vektoren gleicher Dimension **x** und y gilt **x** · **y** = **y** ·**x.** Mit anderen Worten: Es gibt keinen Unterschied, ob wir **x** auf **y** projizieren oder umgekehrt. • Wir können

die negativen Werte annehmen, wenn beispielsweise die Projektionen zweier Vektoren, auf denen wir die Projektion durchgeführt haben, entgegengesetzte Richtungen haben.

BEISPIELE

Wir betrachten zwei Vektoren im dreidimensionalen Raum, gegeben durch ihre Komponenten:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{und } \mathbf{y} = \begin{bmatrix} 0 \\ 2 \\ ÿ1 \end{bmatrix}.$$

Das Skalarprodukt von **x** und y, bezeichnet mit **x** · y, ist definiert durch:

$$\mathbf{x} \cdot \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \\ ÿ1 \end{bmatrix}$$

$$= [1\ 2\ 3] \cdot \begin{bmatrix} 0 \\ 2 \\ ÿ1 \end{bmatrix},$$

$$\mathbf{x} \cdot \mathbf{y} = 1 \cdot 0 + 2 \cdot 2 + 3 \cdot (ÿ1) = 4 ÿ 3\ \mathbf{x} \cdot \mathbf{y} = 1,$$

.

Die **Norm** von x, bezeichnet mit **x ,** ist durch die Quadratwurzel des Skalarprodukts selbst gegeben. Daher:

$$\mathbf{x} \cdot \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$= [1\ 2\ 3] \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix},$$

$$\mathbf{x} \cdot \mathbf{x} = 12 + 22 + 32 = 1 + 4 + 9,$$

$$\mathbf{x} \cdot \mathbf{x} = 14,$$

woraus $\mathbf{x} = ÿ\sqrt{14} ÿ 3{,}74$.

Der Vektor **x** hat im dreidimensionalen Raum eine Länge von 3,74.

Auf die gleiche Weise können wir die Norm des Vektors y berechnen. Es ist gegeben durch:

$$\mathbf{j}\ 2 = y \cdot \mathbf{y} = \sum_{i=1}^{3} y2 = 22 + (ÿ1)\ 2 = 5$$

und entspricht $\mathbf{y} = ÿ\sqrt{5}$.

Somit können wir den Winkel ÿ, der zwischen den beiden Vektoren liegt, mit Hilfe der Beziehung bestimmen:

$$\cos ÿ = \frac{\mathbf{x} \cdot \mathbf{y}}{\mathbf{x} \cdot \mathbf{y}} = \frac{1}{ÿ\sqrt{14} \cdot ÿ\sqrt{5}}$$

$$ÿ\ 0{,}1195.$$

Dies ergibt einen ungefähren Winkel von *83,1ÿ* zwischen den Vektoren **x** und y.

Machine Translated by Google

WEITERLESEN

Matrix

Transponieren

Vektor

Interaktion

Der Begriff Interaktion wird in **der Versuchsplanung** verwendet , genauer gesagt in **faktoriellen Experimenten,** bei denen eine bestimmte Anzahl von Faktoren gleichzeitig untersucht werden kann. Es tritt auf, wenn die Auswirkung einer Änderung des **Niveaus** eines Faktors auf die **abhängige Variable** von den Niveaus anderer Faktoren abhängt.

GESCHICHTE Das Interaktionskonzept bei **der Gestaltung experimenteller** Methoden geht auf **Fisher, RA** (1925, 1935) zurück.

MATHEMATISCHE ASPEKTE

Das **Modell** eines Zwei-Wege-Klassifizierungsdesigns mit Interaktion lautet wie folgt:

$$Yijk = ÿ + ÿi + ÿj + (ÿÿ)ij + ÿijk \text{ , } i = 1, 2,...,a, \quad j = 1, 2,...,b \text{ und } k = 1, 2,...,nij \text{ ,}$$

Dabei ist *Yijk* die k-te **Beobachtung ,** die behandelt wird , *ij, ÿ* **der allgemeine Mittelwert** aller Behandlungen, ÿi die Auswirkung der **i-ten Stufe** von **Faktor** A, *ÿj* die Auswirkung der j-ten Stufe von Faktor B, (ÿÿ) ij die Auswirkung der Wechselwirkung zwischen der i-ten Stufe von Faktor A und j-te Ebene des Faktors B, und *ijk* ist der experimentelle Beobachtungsfehler **Yijk .**

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Wechselwirkungen zwischen zwei **Faktoren** sind sogenannte Wechselwirkungen erster Ordnung, jene zwischen Drei Faktoren werden als Interac zweiter Ordnung bezeichnet tionen usw.

BEISPIELE

Siehe **bidirektionale Varianzanalyse.**

WEITERLESEN

Varianzanalyse

Versuchsplanung

Experiment

Fisher-Verteilung

Fisher-Tisch

Fisher-Test

Zwei-Wege-Varianzanalyse

REFERENZEN

Cox, DR: Interaktion. Int. Stat. Offb. **52,** 1–25 (1984)

Fisher, RA: Statistische Methoden für Forscher. Oliver & Boyd, Edinburgh (1925)

Fisher, RA: Das Design von Experimenten. Oliver & Boyd, Edinburgh (1935)

Interquartilbereich

Der Interquartilabstand ist ein **Maß für die Streuung,** das der Differenz zwischen dem ersten und dem dritten Quartil entspricht . Es entspricht daher dem **Intervall ,** das 50 % der zentriertesten **Beobachtungen** der Verteilung enthält.

MATHEMATISCHE ASPEKTE

Betrachten Sie Q1 und Q3 als die Kacheln des ersten und dritten **Quars** einer Verteilung. Der Interquartilbereich wird wie folgt berechnet:

$$\text{Interquartilbereich} = Q3 ÿ Q1 .$$

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Interquartilabstand ist ein Maß für die Variabilität, das nicht von der Anzahl der **Beobachtungen abhängt.** Darüber hinaus ist dieses Maß am anderen Ende der **Spanne** weniger empfindlich gegenüber **Ausreißern.**

BEISPIELE

Betrachten Sie die folgenden zehn **Beobachtungen:**

0112223345

Das erste **Quartil** Q1 ist gleich 1 und das dritte Quartil Q3 gleich 3. Der Interquartilabstand ist also gleich: Interquartilabstand = Q3

ÿ Q1 = 3 ÿ 1 = 2 Wir können daraus schließen, dass · 50 % der

Beobachtungen liegen in einem **Intervall** von

Länge 2, also das Intervall zwischen 1 und 3.

WEITERLESEN

- **Maß der Streuung**
- **Quartil**
- **Bereich**

Intervall

Ein Intervall wird durch zwei Grenzen, *a* und b, bestimmt. Diese werden als Grenzwerte **bezeichnet** das Intervall.

Intervalle können geschlossen, offen, halboffen und halbgeschlossen sein.

– *Geschlossenes Intervall:*

bezeichnet *[a,* b], stellt die Menge von x dar mit:

$$a \text{ ÿ } x \text{ ÿ } b .$$

– *Offenes Intervall:*

bezeichnet mit]a, b[, repräsentiert die Menge von xmit:

$$a < x < b .$$

– *Intervall, das links halboffen oder rechts halbgeschlossen ist:* – mit]a, b] bezeichnet, stellt die Menge von *x* dar mit:

$$a < x \text{ ÿ } b .$$

– *Intervall, das rechts halboffen oder links halbgeschlossen ist:*

bezeichnet mit *[a,* b[, repräsentiert die Menge von xmit:

$$a \text{ ÿ } x < b .$$

WEITERLESEN

- **Häufigkeitsverteilung**
- **Histogramm**
- **Ogive**

Inverse Matrix

Für eine quadratische **Matrix A** der Ordnung n nennen wir die inverse Matrix von **A** die quadratische Matrix der Ordnung n , sodass das Ergebnis des Matrixprodukts von **A** und seiner Umkehrung (wie die der inversen Multiplikation von A) gleich der **Identitätsmatrix** der Ordnung n ist. Für jede quadratische Matrix ist die Umkehrmatrix, sofern vorhanden, eindeutig. Um eine invertierbare quadratische Matrix A zu haben, muss ihre Determinante ungleich

MATHEMATISCHE ASPEKTE

Sei eine **quadratische Matrix** der Ordnung *n* mit einer Determinante ungleich Null. Wir nennen die Umkehrung von **A** die eindeutige quadratische Matrix der Ordnung n, bezeichnet Aÿ1, wie zum Beispiel:

$$\mathbf{A} \cdot \text{Aÿ1} = \text{Aÿ1} \cdot \mathbf{A} = \mathbf{In} ,$$

wobei **In** die Identitätsmatrix der Ordnung n ist. Wir sagen also, dass die Matrix nicht singulär ist. Wenn det(A) = 0 ist, d. h. die Matrix A nicht invertierbar ist, sagen wir, dass sie singulär ist.

Machine Translated by Google

Die Bedingung, dass die Determinante von **A** kann nicht Null sein, wird durch die Eigenschaften der Determinante erklärt:

$$1 = \det(\mathrm{In}) = \det \mathbf{A} \cdot A\ddot{y}1$$

$$= \det(A) \cdot \det A\ddot{y}1 \quad ,$$

das heißt, det Aÿ1 = 1 $\overline{\det(A)}$.
Andererseits ist der Ausdruck Aÿ1, von A, wird mit Hilfe von det(A) $\frac{1}{\quad}$ (welcher erhalten) und ist gleich:

$$A\ddot{y}1 = i+j \cdot (\ddot{y}1)\ \det(A) \quad \cdot Dij \ ,$$

wobei *Dij* die Determinante der Ordnung *n* ÿ 1 ist, die wir erhalten, indem wir in A (der **Transpose** von A) Zeile *i* und Spalte j löschen. *Dij* heißt Cofaktor des Elements *aij.*

DOMÄNEN UND EINSCHRÄNKUNGEN

Die inverse Matrix wird hauptsächlich bei der linearen Regression in Matrixform verwendet, um die Parameterschätzungen zu bestimmen.
Im Allgemeinen können wir sagen, dass der Begriff der inversen Matrix die Lösung eines linearen Gleichungssystems ermöglicht.

BEISPIELE

Lassen Sie uns den allgemeinen Fall einer quadratischen Matrix behandeln der Ordnung 2: Sei **A** = $\begin{matrix} ab \\ cd \end{matrix}$. Der **Abschrecker** **Minant** von **A** ist gegeben durch:

$$\det(A) = a \cdot d\ \ddot{y}\ b \cdot c \ .$$

Wenn letzterer nicht Null ist, definieren wir:

$$A\ddot{y}1 = \frac{1}{a \cdot d\ \ddot{y}\ b \cdot c} \cdot \begin{matrix} d\ \ddot{y}b \\ \ddot{y}ca \end{matrix} \quad .$$

Wir können also durch direkte Berechnung verifizieren, dass A·Aÿ1 = Aÿ1·A = **I2** ist.

Sei **B** eine quadratische Matrix der Ordnung 3:

$$\mathbf{B} = \begin{matrix} 3\ 21 \\ \ddot{y}\ \ 4\ \ddot{y}1\ 0 \\ \ddot{y}\ \ 2\ \ddot{y}2\ 0 \end{matrix} \quad \ddot{y}\ \ddot{y} \ ,$$

und bestimmen wir die inverse Matrix Bÿ1 von **B ,** falls sie existiert.

Berechnen wir zunächst die Determinante von **B ,** indem wir gemäß der letzten Spalte entwickeln:

$$\det(B) = 1 \cdot [4 \cdot (\ddot{y}2)\ \ddot{y}\ (\ddot{y}1) \cdot 2] \ ,$$

$$\det(B) = \ddot{y}8 + 2 = \ddot{y}6 \ .$$

Da letzteres ungleich Null ist, können wir berechnen Bÿ1 durch Bestimmen der *Dij,* also der Determinanten der Ordnung 2, die man durch Löschen der i-ten Zeile und der j-ten Spalte der **Transponierten** e **B** von B:

$$\mathbf{B} = \begin{matrix} 342 \\ \ddot{y}\ \ 2\ \ddot{y}1\ \ddot{y}2 \quad \ddot{y} \\ \ddot{y}\ \ 100 \quad \ddot{y} \end{matrix}$$

$$D11 = \ddot{y}1 \cdot 0\ \ddot{y}\ (\ddot{y}2) \cdot 0 = 0 \quad ,$$

$$D12 = 2 \cdot 0\ \ddot{y}\ (\ddot{y}2) \cdot 1 = 2 \quad ,$$

$$D13 = 2 \cdot 0\ \ddot{y}\ (\ddot{y}1) \cdot 1 = 1 \quad ,$$

$$\vdots$$

$$D33 = 3 \cdot (\ddot{y}1)\ \ddot{y}\ 4 \cdot 2 = \ddot{y}11 \quad .$$

Somit erhalten wir die inverse Matrix Bÿ1 , indem wir das richtige Vorzeichen an das Element *Dij* anpassen [das heißt (ÿ), wenn die Summe von *i* und *j* gerade ist, und (+), wenn sie ungerade ist]:

$$B\ddot{y}1 = \frac{1}{\ddot{y}6} \cdot \begin{matrix} 0\ \ddot{y}2\ 1 \\ \ddot{y}\ \ 0\ \ddot{y}2\ 4 \\ \ddot{y}\ \ \ddot{y}6\ 10\ \ddot{y}11 \end{matrix} \quad \ddot{y}$$

$$= \frac{1}{6} \cdot \begin{matrix} 0 \quad 2\ \ddot{y}1 \\ \ddot{y}\ \ 0 \quad 2\ \ddot{y}4 \\ \ddot{y}\ \ 6\ \ddot{y}10\ 11 \end{matrix} \quad \ddot{y}\ \ddot{y} \ .$$

Machine Translated by Google

WEITERLESEN
Bestimmend
Verallgemeinerte Umkehrung
Matrix
Multiple lineare Regression
Transponieren

Umkehrung

Betrachten Sie eine **Permutation** der ganzen Zahlen *1, 2,...,n.* Wenn in einem Elementpaar (nicht unbedingt benachbart) der Permutation das erste Element größer als das zweite ist, spricht man von Inversion. Auf diese Weise ist es möglich, in einer Permutation die Anzahl der Inversionen zu zählen, also die Anzahl der Paare, in denen die große ganze Zahl vor der kleinen steht.

MATHEMATISCHE ASPEKTE

Betrachten Sie eine **Permutation** der ganzen Zahlen *1, 2,...,n* gegeben durch ÿ1, *ÿ2,..., ÿn;* Die Anzahl der Inversionen wird durch Zählen der Anzahl der Paare *(ÿi, ÿj)* mit *ÿi > ÿj bestimmt.*
Um sicherzustellen, dass wir keines vergessen, nehmen wir zunächst ÿ1 und zählen die *ÿi, i* > 1, die kleiner als ÿ1 sind, und dann nehmen wir ÿ2 und zählen die Elemente, die sich rechts davon befinden, und fahren so fort, bis ÿnÿ1; Man erhält dann die Anzahl der Inversionen der betrachteten Permutation.

DOMÄNEN UND EINSCHRÄNKUNGEN

Zur Berechnung wird die Anzahl der Inversionen verwendet

spätdie Determinante einer **quadratischen Matrix.**
Die Notation *aij* wird für das Element der Matrix verwendet, das sich am Schnittpunkt von befindet

die i-te Zeile und die j-te Spalte. Durch die

Durchführung des Produkts ··· · *a1r,* für die Prüfung

a1i · a1j · ple bestimmt die Parität der Anzahl der Inversionen der Permutation *ij...r* das Vorzeichen dem Produkt zur Berechnung der Determinante zugeschrieben.

Bei einer **Permutation** der Elemente *1, 2,...,n* ist die Anzahl der Inversionen immer eine ganze Zahl zwischen 0 und $\frac{n \cdot (n\ddot{y}1)}{2}$.

BEISPIELE

Betrachten Sie die ganzen Zahlen 1, 2, 3, 4, 5; **Die Permutation** dieser fünf ganzen Zahlen ist beispielsweise gegeben durch:

14523 .

Die Anzahl der Inversionen dieser Permutation kann bestimmt werden, indem man die Paare aufschreibt, bei denen eine Inversion vorliegt:

(4, 2)
(4, 3)
(5, 2)
(5, 3) ,

was eine Anzahl von Inversionen von 4 ergibt.

Für die **Permutation** 1 2 3 4 5 (Identität pro Mutation) finden wir, dass es keine Inversion gibt. Für die Permutation 5 4 3 2 1 erhalten wir 10 Inversionen, da jedes Paar eine Inversion ist und daher die Anzahl der Inversionen beträgt

gleich der Anzahl der **Kombinationen** von zwei von fünf Elementen, die ohne Wiederholung ausgewählt wurden, was bedeutet:

$$C_5^2 = \frac{5!}{2! \cdot 3!} = 10 .$$

WEITERLESEN
Bestimmend
Permutation

Unregelmäßige Variation

Unregelmäßige Variationen oder zufällige Variationen bilden eine von vier Komponenten einer **Zeit**

Machine Translated by Google

Serie. Sie entsprechen den Bewegungen, die unregelmäßig und meist über kurze Zeiträume auftreten.

Unregelmäßige Variationen folgen keinem bestimmten Modell und sind nicht vorhersehbar.

In der Praxis alle Komponenten von Zeitreihen, die nicht auf den Einfluss von zurückgeführt werden können

Als unregelmäßig gelten zyklische Schwankungen oder saisonale Schwankungen oder solche säkularer **Tendenz .**

GESCHICHTE

Siehe **Zeitreihen.**

MATHEMATISCHE ASPEKTE

Sei *Yt* eine **Zeitreihe;** wir können es mit Hilfe seiner Komponenten beschreiben: • **Säkulare Tendenz** *Tt* • Saisonale Schwankungen *St* • Zyklische Schwankungen *Ct* • Unregelmäßige Schwankungen *It* Wir unterscheiden: • Multiplikatives **Modell:** $Yt = Tt \cdot St \cdot Ct \cdot It$. • Additives Modell: $Yt = Tt + St + Ct + It$.

Wenn die säkulare Tendenz, die saisonalen Schwankungen und die zyklischen Schwankungen bestimmt sind, ist es möglich, die Anfangsdaten der Zeitreihe entsprechend diesen drei Komponenten anzupassen. Wir erhalten also die Werte der unregelmäßigen Variationen zu jedem Zeitpunkt *t* durch die folgenden Ausdrücke:

$$\frac{Yt}{St \cdot Tt \cdot Ct} = It \text{ (multiplikatives Modell)}$$

$Yt \ddot{y} St \ddot{y} Tt \ddot{y} Ct = It$ (additives Modell) .

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Untersuchung der Komponenten von **Zeitreihen** zeigt, dass die unregelmäßigen Variationen in zwei Kategorien eingeteilt werden können:

• Die zahlreichsten können zugeschrieben werden eine große Anzahl kleinerer Ursachen, insbesondere Messfehler, die Schwankungen kleiner Amplitude hervorrufen. • Schäden, die aus einzelnen, unfallbedingten Ereignissen größeren Ausmaßes wie Streiks, Verwaltungsentscheidungen, Finanzbooms, Naturkatastrophen usw. resultieren.

In diesem Fall führen wir zur Anwendung dieser Methode des gleitenden Durchschnitts eine erste Datenverarbeitung durch, um die Rohdaten so gut wie möglich zu korrigieren Daten.

Selbst wenn wir normalerweise davon ausgehen, dass die unregelmäßigen Schwankungen keine dauerhaften Schwankungen hervorrufen, sondern nur in einem kurzen Zeitintervall , können wir uns vorstellen, dass sie stark genug sind, um zyklischen oder anderen Schwankungen zu folgen. In der Praxis stellen wir fest, dass die unregelmäßigen Variationen dazu neigen, eine kleine Amplitude aufzuweisen und einer **Normalverteilung zu folgen.**

Das bedeutet, dass es häufig zu kleinen und selten zu großen Abweichungen kommt.

WEITERLESEN

Zyklische Schwankung
Normalverteilung
Saisonale Unterschiede
Säkularen Trend
Zeitfolgen

VERWEISE

Bowerman, BL, O'Connel, RT: Zeitreihen und Prognosen: Ein angewandter Ansatz. Duxbury, Belmont, Kalifornien (1979)

Harnett, DL, Murphy, JL: Einführung in die statistische Analyse. Addison-Wesley, Reading, MA (1975)

Machine Translated by Google

Jackknife-Methode

Bei Punktschätzungen anhand des **Konfidenzintervalls ist es oft schwierig,** die **Verzerrung** oder **den Standardfehler** der verwendeten **Schätzung** abzuschätzen **tor.** Die Gründe lassen sich im Allgemeinen in zwei Punkten zusammenfassen: (1) der Mangel an Informationen über die Form der theoretischen Datenverteilung Bution und (2) die Komplexität des theoretischen **Modells ,** wenn die Hypothese als zu stark angesehen wird. Die Jackknife-Methode ermöglicht eine Reduzierung der Bials bei der numerischen Schätzung des Standardfehlers sowie eine Konfidenz. Intervall. Dies geschieht in der Regel durch die Generierung künstlicher Teilstichproben aus den Daten der gesamten **Stichprobe.**

GESCHICHTE

Die Jackknife-Methode wurde von Quenouille (1949, 1956) und **John Wilder Tukey** (1958) entwickelt. Dank Computern, die in sehr kurzer Zeit große Datenmengen erzeugen können, ist sie mittlerweile die am weitesten verbreitete Methode.

Die Entstehung der Jackknife-Bezeichnung wird Tukey, JW (1958) zugeschrieben. Der Name dieser Methode stammt von einem Scout-Messer mit Viele Klingen können in verschiedenen Situationen eingesetzt werden.

MATHEMATISCHE ASPEKTE

Sei *x1,...,xn* eine Anzahl unabhängiger Beobachtungen einer **Zufallsvariablen.** Wir Sie möchten einen **Parameter** ÿ der **Population** schätzen . Die Klappmessertechnik ist die folgende: Der **Schätzer** ÿˆ von ÿ wird aus der Anfangsstichprobe berechnet. Letztere werden in *J* Gruppen mit *m* = ihnen aufgeteilt (zur $\frac{N}{J}$ Beobachtungen in jedem von Vereinfachung des Problems nehmen wir an, dass *n* ein Vielfaches von J ist).

Der Schätzer von ÿ nach der Jackknife-Methode ist der **Mittelwert** der Pseudowerte:

$$\hat{1}\,\bar{\ddot{y}} = \frac{}{J} \cdot \sum_{j=1}^{J} \hat{\ddot{y}}\, j\,,$$

wobei $\hat{\ddot{y}}\, j$ der *Pseudowert* genannt wird und definiert ist durch:

$$\hat{\ddot{y}}\, j = J \cdot \hat{\ddot{y}}\ \ddot{y}\ (J\ \ddot{y}\ 1) \cdot \hat{\ddot{y}} \qquad (J)$$

und die Schätzer $\hat{\ddot{y}}\ (j)\ (j = 1,...,J)$ werden auf die gleiche Weise wie ÿˆ berechnet, jedoch mit gelöschter j-ter Gruppe.

Die **Schätzung** der **Varianz** ist definiert von:

$$S2 = \mathrm{Var}(\hat{\ddot{y}})^{-} = \sum_{j=1}^{J} \frac{(\hat{\ddot{y}}j\ \ddot{y}\ \ddot{y})^{-2}}{(J\ \ddot{y}\ 1) \cdot J}\,.$$

Machine Translated by Google

Der Schätzer des **Bias** ist:

$$B(\hat{\theta})^{-} = (J-1)\cdot(\hat{\theta}_{(.)} - \theta)\hat{},$$

wobei $\hat{\theta}_{(.)}$ ist wie folgt definiert:

$$\hat{\theta}_{(.)} = \frac{1}{J}\cdot\sum_{j=1}^{J}\hat{\theta}_{(j)}.$$

Das **Konfidenzintervall** bei der **Signifikanz Level** θ ist gegeben durch:

$$\hat{\theta}^{-}\theta \frac{S\cdot tn\theta 1,1\theta\theta}{\theta\sqrt{N}}; \hat{\theta}^{-} + \frac{S\cdot tn\theta 1,1\theta\theta}{\theta\sqrt{N}}.$$

Eine der wichtigen Eigenschaften der Klappmessermethode ist die Tatsache, dass wir sie löschen die Voreingenommenheit die Laufzeit der Ordnung $\frac{1}{N}$.

Angenommen, a1, *a2*,... sind Funktionen von θ; wir können es als **mathematische Erwartung** darstellen von $\hat{\theta}$:

$$E[\hat{\theta}] = \theta + \frac{a1}{n} + \frac{a2}{n2} + \ldots$$

$$= \theta + + + \frac{a1}{m\cdot J} \quad \frac{a2}{(m\cdot J)2}$$

Und

$$E[\hat{\theta}_j] = J\cdot E[\hat{\theta}]\ \theta\ (J-1)\cdot E[\hat{\theta}_{(j)}]$$

$$= J\cdot\theta + + + \ldots\frac{a1}{m\cdot J} \quad \frac{a2}{(m\cdot J)2}$$

$$\theta\ (J-1)\cdot\theta + + \ldots \quad \frac{a1}{m\cdot(J-1)}$$

$$= J\cdot\theta + \frac{a1}{M} + +\frac{a2}{\ldots\ m2} - J\,a2 \quad \theta\ J\cdot\theta$$

$$+\theta\ \theta \quad \frac{a1}{M} \quad \frac{}{m2\cdot(J-1)\ 1} \quad \ldots$$

$$= \theta\ \theta\ a2\cdot + \ldots\frac{}{m2\cdot J\cdot(J-1)}$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Idee hinter der Klappmessermethode, die ist relativ alt, modern geworden dank Informatik. Das Jackknife generiert aus den Originaldaten Proben und ermittelt diese das **Konfidenzintervall** bei einem gewählten **Signifikanzniveau** einer unbekannten **Statistik.**

In der Praxis werden die Gruppen zusammengestellt nur in einer **Beobachtung** ($m = 1$). Die Schätzer $\hat{\theta}_{(j)}$ werden durch Streichen des j-ten berechnet Überwachung.

BEISPIELE

Eine große Ladenkette möchte das ermitteln Entwicklung seines Umsatzes zwischen 1950 und 1960.

Um eine korrekte **Schätzung zu erhalten,** nehmen wir eine **Auswahl** von 8 Verkaufsstellen, wo wir wissen die jeweiligen Umsätze.

Die Daten sind im Folgenden zusammengefasst Tisch:

Kasse	Umsatz (Mio. $)	
	1950	1960
1	6.03	7.59
2	12.42	15.65
3	11.89	14,95
4	3.12	5.03
5	8.17	11.08
6	13.34	18.48
7	5.30	7,98
8	4.23	5.61

Der Gesamtumsatz der 8 Verkaufsstellen stieg von 64,5 im Jahr 1950 auf 86,37 Mio. $ im Jahr 1960.

Daher die Schätzung des zunehmenden Faktors entspricht:

$$\hat{\theta} = \frac{86,37}{64,5} = 1,34.$$

Wir wollen den **Bias** dieser Schätzung berechnen tion sowie deren **Standardabweichung.**

Wir berechnen das $\hat{\theta}_{(j)}$ für *j* von 1 bis 8, d. h. die Erhöhungsfaktoren werden berechnet ohne die Daten der j-ten Verkaufsstelle. Wir

Machine Translated by Google

erhalten:

$$\hat{\theta}_{(1)} = \frac{86{,}37 - 7{,}59}{64{,}5 - 6{,}03} = 1{,}347 \quad ,$$

$\hat{\theta} = 1{,}358$ $\hat{\theta}_{(6)} = 1{,}327$

$\hat{\theta}_{(2)} = 1{,}358$ $\hat{\theta}_{(7)} = 1{,}324$

$\hat{\theta}_{(3)} = 1{,}325$ $\hat{\theta}_{(8)} = 1{,}340$

$\hat{\theta}_{(4)}$

$\hat{\theta}_{(5)} = 1{,}337$.

So können wir $\hat{\theta}_{(.)}$ berechnen:

$$\hat{\theta}_{(.)} = \frac{1}{J} \cdot \sum_{j=1}^{J} \hat{\theta}_{(J)} = \frac{1}{8} \cdot (10{,}716) = 1{,}339 \; .$$

Wir berechnen die Pseudowerte $\hat{\theta}$ J. Erinnere dich daran

$\hat{\theta}_J$ wird wie folgt berechnet:

$$\hat{\theta}_J = J \cdot \hat{\theta} - (J - 1) \cdot \hat{\theta}_{(j)} \; .$$

Wir finden:

$\hat{\theta}_1 = 1{,}281$ $\hat{\theta}_5 = 1{,}356$

$\hat{\theta}_2 = 1{,}207$ $\hat{\theta}_6 = 1{,}423$

$\hat{\theta}_3 = 1{,}210$ $\hat{\theta}_7 = 1{,}443$

$\hat{\theta}_4 = 1{,}436$ $\hat{\theta}_8 = 1{,}333$.

Wir berechnen den **Schätzer** von θ anhand der Buchse Messermethode:

$$\bar{\hat{\theta}} = \frac{1}{J} \cdot \sum_{j=1}^{J} \hat{\theta}_J = \frac{1}{8} \cdot (10{,}689) = 1{,}336 \; .$$

Also bestimmen wir den Bias:

$$\begin{aligned} B(\hat{\theta})^{-} &= (J - 1) \cdot (\hat{\theta}_{(.)} - \hat{\theta}) \\ &= (8 - 1) \cdot (1{,}339 - 1{,}34) \\ &= -0{,}0035 \; . \end{aligned}$$

Der **Standardfehler** ist gegeben durch:

$$\begin{aligned} S &= \sqrt{\mathrm{Var}(\hat{\theta})^{-}} \\ &= \sqrt{\frac{\sum_{j=1}^{J} (\hat{\theta}_j - \hat{\theta})^{-2}}{(J - 1) \cdot J}} \\ &= \sqrt{\frac{0{,}0649}{7 \cdot 8}} = 0{,}034 \; . \end{aligned}$$

So können wir das **Vertrauen inter** berechnen **Wert** des Steigerungsfaktors zwischen 1950 und 1960 des Umsatzes der Kette. Für ein **Signifikanzniveau** α von 5 % ist der kritische **Wert** von Die **Student-Tabelle** für 7ν Freiheit lautet:

$$t_{7,0{,}975} = 2{,}365 \; .$$

Das Konfidenzintervall für die Schätzung von

Der Erhöhungsfaktor beträgt:

$$\bar{\hat{\theta}} - \alpha \; ; \; \bar{\hat{\theta}} - \frac{S \cdot t_{7,0{,}975}}{\sqrt{n}} \quad \bar{\hat{\theta}} + \frac{S \cdot t_{7,0{,}975}}{\sqrt{n}}$$

$$1{,}34 - \frac{0{,}034 \cdot 2{,}365}{\sqrt{8}} \; ,$$

$$1{,}34 + \frac{0{,}034 \cdot 2{,}365}{\sqrt{8}} \; ,$$

das ist:

$$[1.312, 1.368].$$

Wir können also bestätigen, dass der Mittelwert zunimmt

im Umsatz ist ungefähr enthalten

zwischen 31 % und 37 % während des Zeitraums 1950–1960.

J

WEITERLESEN

Voreingenommenheit

Konfidenzintervall

Einschätzung

Schätzer

Punktschätzung

Standart Fehler

VERWEISE

Efron, B.: Das Klappmesser, der Bootstrap und die anderen Resampling-Pläne. Gesellschaft für Technische und Angewandte Mathematik, Philadelphia (1982)

Quenouille, MH: Näherungstests der Korrelation in Zeitreihen. J. Roy. Stat. Soc. Ser. B **11,** 68–84 (1949)

Machine Translated by Google

Quenouille, MH: Hinweise zur Verzerrung bei Schätzungen. Biometrie **43,** 353–360 (1956)

Tukey, JW: Voreingenommenheit und Vertrauen in nicht ganz große Stichproben. Ann. Mathematik. Stat. **29,** 614 (1958)

Jeffreys, Harold

Jeffreys, Harold wurde 1891 in Fatfield, England, geboren. Er studierte Mathematik, Physik, physikalische Chemie und Geologie am Armstrong College in Newcastle, wo er 1910 seinen Abschluss machte. Anschließend besuchte er das St. John's College der Universität Cambridge, wo er Er wurde 1914 Fellow und blieb dort 75 Jahre lang. Dort starb er 1989.
Seine erste Arbeit (über Fotografie) veröffentlichte er 1910. Die Geschichte hat insbesondere zwei seiner Werke bewahrt:

1939 Wahrscheinlichkeitstheorie. Oxford Uni versity Press, Oxford.

1946 (mit Jeffreys, BS) Methoden der Mathematischen Physik. Cambridge University Press, Cambridge; Macmillan, New York.

Gelenkdichtefunktion

Angenommen, X1 und X2 sind kontinuierliche Zufallsvariablen, die auf demselben **Stichprobenraum** definiert sind. Das

$$F(x1, x2) = P(X1 \le x1, X2 \le x2) = \int_{-\infty}^{x2}\int_{-\infty}^{x1} f(t1, t2)\, dt1, dt2$$

für alle x1, *xk*. Dann ist *f* (x1, x2) die Gelenkdichtefunktion von (X1, X2) , vorausgesetzt *f(x1,* x2) ≥ 0 und das

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x1x2)\, dx1dx2 = 1 .$$

Angenommen, *(X1,...,Xk)* sind kontinuierliche Zufallsvariablen, die auf demselben **Probenraum** und definiert sind

$$F(x1,...,xk) = P(X1 \le x1,...,Xk \le xk) = \int_{-\infty}^{xk}\cdots\int_{-\infty}^{x1} f(t1,...,tk)\, dt1 \ldots dtk$$

für alle *x1,...,xk.* Dann ist *f(x1,...,xk)* die Gelenkdichtefunktion *von (x1,...,xk),* vorausgesetzt *f(x1,...,xk)* ≥ 0 und

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x1,...,xk)\, dx1 \ldots dxk = 1 .$$

REFERENZEN

Knight, K.: Mathematische Statistik. Chap Man & Hall/CRC, London (2000)

Gemeinsame Verteilungsfunktion

Die gemeinsame Verteilungsfunktion eines **Paars von Zufallsvariablen** ist definiert als die **Wahrscheinlichkeit ,** dass die erste **Zufallsvariable** einen **Wert** kleiner oder gleich der ersten reellen Zahl annimmt und dass gleichzeitig die zweite Variable einen Wert annimmt, der kleiner oder gleich der ersten reellen Z

Machine Translated by Google

able nimmt einen Wert an, der kleiner oder gleich der zweiten reellen Zahl ist.

GESCHICHTE

Siehe **Wahrscheinlichkeit.**

MATHEMATISCHE ASPEKTE

Betrachten Sie zwei **Zufallsvariablen** *X* und Y.

Die gemeinsame Verteilungsfunktion ist wie folgt definiert:

$$Fxy\ (a,\ b) = P\ (X \ddot{y}\ a,\ Y \ddot{y}\ b)\ ,\ \ddot{y}\ \ddot{y} < a,\ b < \ddot{y} \quad .$$

WEITERLESEN

Verteilungsfunktion

Randverteilungsfunktion
Paar Zufallsvariablen

VERWEISE

Hogg, RV, Craig, AT: Einführung in die Mathematische Statistik, 2. Aufl. Macmillan, New York (1965)

Ross, SM: Einführung in Wahrscheinlichkeitsmodelle, 8. Aufl. John Wiley, New York (2006)

Gemeinsame Wahrscheinlichkeit

Verteilungsfunktion

Nehmen wir an, dass X1, X2 diskrete Zufallsvariablen sind, die auf demselben **Stichprobenraum definiert sind.** Dann wird die gemeinsame Wahrscheinlichkeitsverteilungsfunktion von $X = (X1, X2)$ wie folgt def

$$f(x1,\ x2) = P(X1 = x1,\ X2 = x2)\ .$$

Die gemeinsame Wahrscheinlichkeitsverteilungsfunktion muss die folgenden Bedingungen erfüllen:

$$\sum_{x1,x2} f\ (x1,\ x2) = 1 \quad .$$

Eine mehrfache gemeinsame Wahrscheinlichkeitsverteilungsfunktion von $X = (X1,...,Xk)$ *ist* als $f(x1,...,xk) = P(X1 = x1,...,Xk = xk)$ erfüllend definiert

$$f(x1,\ ...,\ xk) = 1\ (x1,\ ...,\ xk)$$

VERWEISE

Casella, G., Berger, RL: Statistische Inferenz. Duxbury, Pacific Grove (2001)

J

Machine Translated by Google

Kendall, Maurice George

Kendall, Maurice wurde 1907 in Ket tering, Northamptonshire, England, geboren. Er studierte Mathematik am St. John's College in Cam Bridge. Nach dem Abschluss als Mathematiker Wrangler trat 1929 in den britischen Zivildienst im Landwirtschaftsministerium ein. Er wurde 1934 zum Fellow der Gesellschaft gewählt. 1937 arbeitete er mit G. Udny Yule in der Überarbeitung seines Standardlehrbuchs für Statistik, *Einführung in die Theorie der Statistik.* Er Arbeiten Sie auch am Rangkorrelationskoeffizienten der seinen Namen trägt, **Kendalls Tau,** der führte schließlich 1948 zu einer Monographie über *die Rank-Cor-Beziehung* .

In den Jahren 1938 und 1939 begann er mit der Arbeit Bernard Babington Smith zum Problem der **Erzeugung von Zufallszahlen,** die Entwicklung beider einer der ersten frühen mechanischen Systeme Geräte zur Erzeugung zufälliger Ziffern und zur Simulation einer Reihe von Tests, beispielsweise der *Frequenz Test, Serientest und ein Pokertest,* für statistische Zufälligkeit in einem gegebenen Satz von Ziffern. Während des Krieges gelang es ihm, Band 1 der „Advanced Theory of Statistics“ herauszugeben 1943 und ein zweiter Band 1946.

1957 veröffentlichte er *Multivariate Analysis* und im selben Jahr entwickelte er auch, mit WR Buckland, ein *Wörterbuch der Statistik Bedingungen.*

1953 veröffentlichte er The Analytics of Economic Time Series, 1961 verließ er die Zeitschrift

Er studierte an der University of London und nahm eine Position als an

Geschäftsführer eines Beratungsunternehmens, Scientific Control Systems, und in der

Im selben Jahr begann die zweijährige Amtszeit des Präsidenten der Royal Statistical Society.
1972 wurde er Direktor der Welt

Fertility Survey, ein von der geförderten Projekt Internationales Statistisches Institut und das Vereinte Nationen. Er führte diese Arbeit fort

bis er 1980 krankheitsbedingt in den Ruhestand musste.

Für seine Verdienste wurde er 1974 zum Ritter geschlagen zur Theorie der Statistik und erhielt die Friedensmedaille der Vereinten Nationen im Jahr 1980

in Anerkennung seiner Arbeit am World Fertility Survey. Er wurde auch zum Fellow von gewählt der British Academy und erhielt die höchste Auszeichnung der Royal Statistical Society, die Guy-Medaille in Gold. Zum Zeitpunkt seines Todes 1983 war er Ehrenpräsident der Internationales Statistisches Institut.

Einige Hauptwerke und Artikel von Ken dall, Maurice George:

1938 (mit Babington Smith, B.) Randomness and Random Sampling Numbers. J. Roy. Stat. Soc. 101:1, 147–166.

1979 (mit Stuart, A.) Fortgeschrittene Theorie von Statistiken. Arnold.

Machine Translated by Google

1957 (mit Buckland, WR) A Dictionary statistischer Begriffe. International al Statistisches Institut, Den Haag, Niederlande.

1973 Time Series, Griffin, London.

WEITERLESEN

Kendall-Rangkorrelationskoeffizient

Zufallszahlengenerierung

Kendall-Rang Korrelationskoeffizient

Der Kendall-Rangkorrelationskoeffizient (Kendall ÿ) ist ein nichtparametrisches Maß von Korrelation.

GESCHICHTE

Dieser Rangkorrelationskoeffizient war dis bereits im frühen 20. Jahrhundert verflucht von Fechner, GT (1897), Lipps, GF (1906), und Deuchler, G. (1914).
Kendall, MG (1938) hat es nicht nur unabhängig wiederentdeckt, sondern auch untersucht unter Verwendung eines (nichtparametrischen) Ansatzes. Sein 1970 Die Monographie enthält eine vollständige, detaillierte Darstellung der Theorie sowie eine Biographie.

MATHEMATISCHE ASPEKTE

Betrachten Sie zwei Zufallsvariablen **(X,** Y) beobachtet an einer **Stichprobe** der Größe *n* mit *n* Beobachtungspaare (X1, Y1), (X2, Y2) , ..., *(Xn,* Yn). Einen Hinweis auf die Korrelation zwischen *X* und *Y* erhält man durch Ordnen der **Werte** *Xi* in aufsteigender Reihenfolge und indem die Anzahl der entsprechenden Werte *Yi* gezählt wird , die dieser Reihenfolge nicht genügen. *Q* bezeichnet die Anzahl der **Inversionen** unter den Werten von *Y* , die erforderlich sind erhalten die gleiche (aufsteigende) Ordnung wie die Werte von X.

Da es verschiedene $\frac{n(n\ddot{y}1)}{2}$ Paare gibt, die das können gebildet werden, $0 \ddot{y} Q \ddot{y}$; der Wert 0 ist 2 erhalten, wenn alle **Werte** *Yi* bereits sind in aufsteigender Reihenfolge und der Wert $\frac{n(n\ddot{y}1)}{2}$ Ist erreicht, wenn alle Werte *Yi* invers sind Reihenfolge von *Xi,* wobei jedes Paar vertauscht werden muss um die gewünschte Bestellung zu erhalten.

Der Kendall-Rangkorrelationskoeffizient, bezeichnet mit ÿ , ist definiert durch:

$$\ddot{y} = 1 \ddot{y} \frac{4Q}{n(n \ddot{y} 1)} .$$

Wenn alle Paare in aufsteigender Reihenfolge vorliegen, dann gilt:

$$\ddot{y} = 1 \ddot{y} \frac{4 \cdot 0}{n(n \ddot{y} 1)} = 1 .$$

Wenn alle Paare in umgekehrter Reihenfolge sind, dann:

$$\cdot \ddot{y} = 1 \ddot{y} \frac{4 \frac{1}{2} \cdot n(n \ddot{y} 1)}{n(n \ddot{y} 1)} = \ddot{y}1 .$$

Eine äquivalente Definition des Kendall-Ranges Der Koeffizient kann wie folgt angegeben werden: zwei **Beobachtungen** heißen übereinstimmend, wenn die zwei Mitglieder einer Beobachtung sind größer als die entsprechenden Mitglieder der anderen Beobachtung. Zum Beispiel (0,9, 1,1) und (1.5, 2.4) sind zwei übereinstimmende Beobachtungen weil *0,9 < 1,5* und *1,1* < 2,4. Zwei Beobachtungen gelten als nicht übereinstimmend, wenn die beiden Mitglieder einer Beobachtung sind entgegengesetzt Reihenfolge an die jeweiligen Mitglieder des anderen Überwachung. Zum Beispiel (0,8, 2,6) und (1,3, 2.1) sind zwei widersprüchliche Beobachtungen, weil *0,8 < 1,3* und *2,6* > 2,1.

Nc und *Nd* bezeichnen die Gesamtzahl *von* Paare übereinstimmender bzw. widersprechender **Beobachtungen .**

Machine Translated by Google

Zwei Paare, für die *Xi* = *Xj* und *Yi* = *Yj* sind, sind weder übereinstimmend noch nicht übereinstimmend und werden daher weder in *Nc* noch in *Nd gezählt.*

Mit dieser Notation sind die Kendall-Rangkoeffi

cient ist gegeben durch:

$$\ddot{y} = \frac{2(Nc \ddot{y} Nd)}{\ddot{y} 1)} \; n(n \; .$$

Beachten Sie, dass die beiden Formulierungen von ÿ genau gleich sind , wenn es keine Paare gibt, für die *Xi* = *Xj* oder *Yi* = *Yj* ist. Im umgekehrten Fall können die **Werte** beider Formeln unterschiedlich sein.

Hypothesentest

Der Kendall-Rangkorrelationskoeffizient beträgt Wird oft als statistischer Test zur Bestimmung verwendet wenn es eine Beziehung zwischen zwei **Zufällen** gibt **Variablen.** Der Test kann ein **zweiseitiger Test** sein oder ein **einseitiger Test.** Die **Hypothesen** sind:

A: Zweiseitiger Fall:

H0: *X* und *Y* sind voneinander unabhängig.

H1: Es besteht entweder eine positive oder eine negative Korrelation zwischen *X* und Y.

Es besteht eine positive Korrelation, wenn die großen **Werte** von *X* tendenziell mit den großen Werten von *Y* und die kleinen Werte von *X* mit den kleinen Werten von Y assoziiert sind.

Es besteht eine negative Korrelation, wenn die großen Werte von *X* tendenziell mit den kleinen Werten von Y verknüpft sind und umgekehrt.

B: Einseitiger Fall:

H0: *X* und *Y* sind voneinander unabhängig.

H1: Es besteht eine positive Korrelation zwischen *X* und Y.

C: Einseitiger Fall:

H0: *X* und *Y* sind voneinander unabhängig.

H1: Es besteht eine negative Korrelation zwischen *X* und Y.

Der statistische Test ist wie folgt definiert:

$$T = Nc \ddot{y} Nd .$$

Entscheidungsregeln

Die Entscheidungsregeln sind je nach den aufgestellten **Hypothesen** unterschiedlich. Deshalb gibt es in Bezug auf die vorherigen Fälle die Entscheidungsregeln A, B und C.

Entscheidungsregel A

Lehnen Sie H0 auf dem **Signifikanzniveau** ÿ ab, wenn

$$T > \text{tn,1ÿÿ} \quad \text{oder } T < tn, \ddot{y} \; ,$$

wobei *t* der **kritische Wert** des Tests ist , der durch die Kendall-Tabelle gegeben ist; ansonsten akzeptiere H0.

Entscheidungsregel B

Lehnen Sie H0 auf dem **Signifikanzniveau** ÿ ab, wenn

$$T > \text{tn,1ÿÿ} .$$

ansonsten akzeptiere H0.

Entscheidungsregel C H0 auf der **Signifikanzebene** ÿ ablehnen , wenn

$$T < tn,\ddot{y} .$$

ansonsten akzeptiere H0.

Es ist auch möglich, es zu verwenden

$$\ddot{y} = 1 \ddot{y} \; n(n \; \frac{4Q}{\ddot{y} 1)}$$

als statistischer Test.

Wenn *X* und *Y* unabhängig voneinander in einer **Population** verteilt sind , hat die genaue Verteilung von ÿ **einen erwarteten Wert von** Null und **eine Varianz** von:

$$\ddot{y}2_{\ddot{y}} = \frac{2(2n + 5)}{9n(n \ddot{y} 1)}$$

und tendiert sehr schnell zu einer Normalverteilung, wobei die Näherung für *n* ÿ 10 gut genug ist.

K

Machine Translated by Google

Um in diesem Fall **die Unabhängigkeit** bei 5 % zu testen,

Ebene reicht es beispielsweise aus, zu überprüfen, ob ÿ
liegt außerhalb der Grenzen

$$\pm 1{,}96 \cdot \ddot{y}\ddot{y}$$

und die **Unabhängigkeitshypothese** abzulehnen, wenn
das ist der Fall.

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Kendall-Rangkorrelationskoeffizient beträgt

Wird als **Hypothesentest** verwendet , um die Abhängigkeit
zwischen zwei **Zufallsvariablen zu untersuchen.** Es

kann **als ein Zeichen der Unabhängigkeit angesehen werden.**

Als nichtparametrisches Korrelationsmaß kann es auch mit
Nominalwerten verwendet werden
oder ordinale **Daten.**

Eine Korrelationsmessung zwischen zwei

Variablen müssen die folgenden Punkte erfüllen:

1. Seine **Werte** liegen zwischen ÿ1 und +1.
2. Es besteht eine positive Korrelation zwischen *X*
   und *Y* , wenn der Wert des Korrelationskoeffizienten ist
   effizient ist positiv; Eine perfekte positive Korrelation
   entspricht einem Wert von +1.
3. Es besteht eine negative Korrelation zwischen
   X und *Y* sind der Wert der Korrelation
   Koeffizient ist negativ; Eine perfekte negative Korrelation
   entspricht einem Wert
   von ÿ1.
4. Es besteht eine Nullkorrelation zwischen *X* und
   Y , wenn der Korrelationskoeffizient nahe beieinander liegt
   bis Null; Man kann auch sagen, dass *X* und *Y* sind
   nicht korreliert.

Der Kendall-Rangkorrelationskoeffizient hat

folgende Vorteile:

- Bei den **Daten** kann es sich um nichtnumerische **Beobachtungen** handeln
  solange sie klassifiziert werden können
  nach einem festgelegten Kriterium.
- Es ist leicht zu berechnen.
- Der zugehörige statistische Test funktioniert nicht
  Stellen Sie eine grundlegende **Hypothese** auf , die auf dem basiert
  Form der Verteilung der **Bevölkerung,** aus der die
  Proben entnommen werden.

Die Kendall-Tabelle gibt die theoretischen **Werte** der
Statistik ÿ des Kendall-Rangkors an

Relationskoeffizient, der als statistischer Test verwendet wird

unter der **Unabhängigkeitshypothese** von zwei
zufällige Variablen.

Eine Endalltabelle finden Sie bei Kaarsemaker

und van Wijngaarden (1953).

Hier ist ein Beispiel der Kendall-Tabelle für *n* =
4,..., 10 und ÿ = 0,01 und 0,05:

N	ÿ = 001.	ÿ = 005.
46 4		
58 6		
6	11	9
7	15	11
8	18	14
9	22	16
10	25	19

BEISPIELE

In diesem Beispiel nehmen acht Paare echte Zwillinge auf

Intelligenztests. Das Ziel besteht darin, herauszufinden, ob dies der Fall ist

Unabhängigkeit zwischen den Tests des einen

wer zuerst geboren wird und diejenigen desjenigen, der zuerst geboren wird

Zweitgeborener.

Die Daten sind in der folgenden Tabelle aufgeführt; Die

Die höchsten Punktzahlen entsprechen den besten Ergebnissen.

Zuerst ein Zwillingspaar geboren Xi	Zweite geboren Yi
1	90	88
2	75	79
3	99	98
4	60	66
5	72	64
6	83	83
7	83	88
8	90	98

Machine Translated by Google

Die Paare werden dann in aufsteigender Reihenfolge klassifiziert Reihenfolge für X und die übereinstimmenden und nicht übereinstimmenden Paare werden bestimmt. Das gibt:

Paar Zwillinge (Xi, Yi)	Übereinstimmung Paare	Discording Paare
(60,66)	6	1
(72,64)	6	0
(75,79)	5	0
(83,83)	3	0
(83,88)	2	0
(90,88)	1	0
(90,98)	0	0
(99,98)	0	0
	Nc = 23	*Nd* = 1

Der Kendall-Rangkorrelationskoeffizient beträgt gegeben von:

$$\ddot{y} = \frac{2(Nc \ddot{y} Nd)}{n(n \ddot{y} 1)} = \frac{2(23 \ddot{y} 1)}{8 \cdot 7} = 0{,}7857 .$$

Beachten Sie, dass es mehrere **Observa tionen,** für die *Xi* = *Xj* oder *Yi* = *Yj,* der Wert des Koeffizienten gegeben durch:

$$\ddot{y} = 1 \ddot{y} = 1 \ddot{y} = 0{,}9286 \qquad \frac{4Q}{n(n \ddot{y} 1)} \qquad \frac{4 \cdot 1}{56}$$

ist anders.

In beiden Fällen stellen wir einen positiven Zusammenhang fest zwischen den Intelligenztests.

Wir führen nun den **Hypothesentest durch:**

H0: Es besteht **Unabhängigkeit** zwischen den Intelligenztests eines Zwillingspaares.

H1: Es besteht eine positive Korrelation zwischen die Intelligenztests.

Wir haben ein **signifikantes Niveau** von ÿ = 0,05 gewählt. Da wir uns im Fall B befinden, wird H0 abgelehnt, wenn

$$T > t8{,}0{,}95 ,$$

wobei *T* = NcÿNd und t8,0,95 der **Wert** von ist der Kendalltisch. DaT = 22 und t8,0,95 = 14, H0 wird abgelehnt.

Daraus können wir schließen, dass es etwas Positives gibt Korrelation zwischen den Ergebnissen der Intel Intelligenztests eines Zwillingspaares.

WEITERLESEN

- **Hypothesentest**
- **Nichtparametrischer Test**
- **Test der Unabhängigkeit**

VERWEISE

Deuchler, G.: Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. Zeitung für Pädagogische Psychologie und Experimentelle Pädagogik, **15,** 114–131, 145–159, 229–242 (1914)

Fechner, GT: Kollektivmasslehre. W. Engelmann, Leipzig (1897)

Kaarsemaker, L., van Wijngaarden, A.: Tabellen zur Verwendung in der Rangkorrelation. Stat. Neerland. 7, 53 (1953)

Kendall, MG: Ein neues Maß für die Rang-Korrelation. Biometrie **30,** 81–93 (1938)

Kendall, MG: Rangkorrelationsmethoden. Griffin, London (1948)

Lipps, GF: Die Psychische Massmethod en. F. Vieweg und Sohn, Braunschweig, Deutschland (1906)

Kiefer, Jack Carl

Kiefer, Jack Carl wurde in Cincinnati geboren, Ohio im Jahr 1924. Er betrat den Bundesstaat Massachusetts Institute of Technology im Jahr 1942, aber danach Nach einem Jahr Studium der Ingenieurs- und Wirtschaftswissenschaften brach er ab, um kriegsbezogene Arbeiten zu übernehm während des Zweiten Weltkriegs. Seine Masterarbeit,

Machine Translated by Google

Sequentielle Bestimmung des Maximums einer Funktion, wurde betreut von Harold Freeman. Es war die Grundlage für seine Arbeit „Sequentielle Minimax-Suche nach ein Maximum“, das 1953 erschien die „Proceedings of the American Mathematical Society“. 1948 ging er an die Abteilung für Mathematische Statistik bei Columbia University, wo AbrahamWald war herausragend in einer Abteilung, zu der Ted Anderson, Henry Scheffé und Jack gehörten Wolfowitz. Seine Doktorarbeit verfasste er in Entscheidungstheorie unter Wolfowitz und ging 1951 ging er zusammen mit Wol Fowitz an die Cornell University. 1973 wurde Kiefer zum Ersten gewählt Horace White Professor an der Cornell Univer Stadt, eine Position, die er bis 1979 innehatte, als er ging in den Ruhestand und trat der Fakultät der University of California in Berkeley bei. Er starb am 1981 57 Jahre alt.

Kiefers Forschungsgebiet war die Gestaltung von Experimente. Die meisten seiner 100 Veröffentlichungen habe mich mit dem Thema beschäftigt. Er verfasste auch Aufsätze Ontopien in der mathematischen Statistik, einschließlich Entscheidungstheorie, stochastische Approximation, Warteschlangentheorie, nichtparametrische Inferenz, Schätzung, sequentielle Analyse und bedingte Inferenz.

Kiefer war Mitglied des Institute of Mathe matischer Statistik und der American Statis Technische Vereinigung und Präsident des Instituts für Mathematische Statistik (1969–1970). Er wurde in die American Academy of gewählt Arts and Sciences im Jahr 1972 und zum National Akademie der Wissenschaften (USA) im Jahr 1975.

Ausgewählte Werke und Veröffentlichungen von Jack Carl Kiefer:

1987 Einführung in die statistische Inferenz. Springer, Berlin Heidelberg Neu York

WEITERLESEN

Versuchsplanung

VERWEISE

Kiefer, J.: Optimale experimentelle Designs. J. Roy. Stat. Soc. Ser. B **21,** 272–319 (1959)

LD Brown, I. Olkin, J. Sacks und HP Wynn (Hrsg.): Jack Karl Kiefer, Collected Papiere. I: Statistische Schlussfolgerung und Wahrscheinlichkeit (1951–1963). New York (1985)

LDBrown, I. Olkin, J. Sacks und H.P. Wynn (Hrsg.): Jack Karl Kiefer, Collected Papers. II: Statistische Schlussfolgerung und Wahrscheinlichkeit (1964–1984). New York (1985)

LD Brown, I. Olkin, J. Sacks und HP Wynn (Hrsg.): Jack Karl Kiefer, Collected Papiere. III: Versuchsplanung. Neu York (1985)

Kolmogorov, Andrei Nikolajewitsch

Andrei Nikolaevich wurde 1903 in Tambow, Russland, als Kol Mogorov geboren Begründer der modernen Wahrscheinlichkeitsrechnung. Im Jahr 1920 er trat in die Moskauer Staatsuniversität ein und studierte Mathematik, Geschichte und Metallurgie. 1925 veröffentlichte er seinen ersten Artikel in Wahrscheinlichkeit auf den Ungleichungen des Teils Summen von Zufallsvariablen, die wurden Hauptreferenz auf dem Gebiet der stochastischen Prozesse. Er promovierte in 1929 und veröffentlichte 18 Artikel zum Recht von großen Zahlen sowie auf intuitiver Logik. Er wurde zum Professor an der Moskauer Staatlichen Universität ernannt Universität im Jahr 1931. Im Jahr 1933 veröffentlichte er seine Monographie über Wahrscheinlichkeitstheorie. 1939 wurde er zum Mitglied der Akademie gewählt mein der Wissenschaften der UdSSR. Er erhielt

Machine Translated by Google

der Lenin-Preis 1965 und der Orden von Lenin bei sechs verschiedenen Gelegenheiten sowie 1987 erhielt er den Lobatschewski-Preis. Er wurde zum Mitglied zahlreicher anderer ausländischer Akademien gewählt darunter die Rumänische Akademie der Wissenschaften (1956), die Royal Statistical Society of London (1956), der Leopoldina-Akademie of Germany (1959), der American Academy of Arts and Sciences (1959), der London Mathematical Society (1959), der American Philosophical Society (1961), dem Indian Institute of Statistics (1962) und der Hol land Academy of Sciences (1963), der Royal Society of London (1964), dem National

Academy of the United States (1967) und die Académie Française des Sciences (1968).

Ausgewählte Hauptwerke von Kolmogorov, Andrei Nikolajewitsch:

1933 Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin Heidelberg New York.

1933 Sulla determinazione empirica di una lege di distribuzione. Giornale dell'Instituto Italiano degli Attuari, 4, 83–91 (6.1).

1941 Lokale Turbulenzstruktur in inkompressible Flüssigkeiten mit sehr hoher Reynolds Nummer. Dan SSSR, 30, 229.

1941 Energiedissipation vor Ort isotrope Turbulenz. Dokl. Akad. Nauk. SSSR, 32, 16–18.

1958 (mit Uspenskii, VA) K oprede leniyu algoritma. (Auf dem Weg zur Definition eines Algorithmus). Uspechi Matematicheskikh Nauk 13(4):3–28, American Mathematical Society Übersetzungen Reihe 2(29):217–245, 1963.

1961 (mit Fomin, SV) Maßnahme, Lebesgue-Integrale und Hilbert-Raum. Natascha Artin Brunswick und Alan Jeffrey. Akademisch, New York.

1963 Zur Darstellung kontinuierlicher Funktionen vieler Variablen von Überlagerung stetiger Funktionen einer Variablen und Addition. Doklady Akademii Nauk SSR, 114, 953–956, 1957. Englische Übersetzung. Transaktionen der Mathematischen Gesellschaft, 28, 55–59.

1965 Drei Ansätze zur quantitativen Definition von Information. Wahrscheinlich Grundlagen der Informationsübertragung, 1, 1–17. Übersetzung von Problemy peredachi informatsii 1(1), 3–11 (1965).

1987 (mit Uspenskii, VA) Algorithmen und Zufälligkeit. Teoria veroyatnos tey i ee primeneniya (Wahrscheinlichkeitslehre und ihre Anwendungen), 3(32):389–412.

WEITERLESEN

Kolmogorov-Smirnov-Test

Kolmogorov-Smirnov-Test

Der Kolmogorov-Smirnov-Test ist ein nichtparametrischer **Anpassungstest** und wird verwendet bestimmen, ob sich zwei Verteilungen unterscheiden, oder ob eine zugrunde liegende Wahrscheinlichkeitsverteilung von einer hypothetischen Verteilung abweicht. Es wird verwendet, wenn wir Bitte kommen aus zwei Populationen, die sein können anders. Im Gegensatz zum **Mann-Whitney-Test** und der **Wilcoxon-Test ,** wo das Ziel liegt den Unterschied zwischen zwei Mittelwerten erkennen oder Mediane hat der Kolmogorov-Smirnovtest

Machine Translated by Google

Der Vorteil, die Verteilungsfunktionen gemeinsam zu betrachten. Der Kolmogorov-Smirnov-Test kann auch als **Anpassungstest verwendet werden.** In diesem Fall haben wir nur eine Zufallsstichprobe aus einer **Grundgesamtheit ,** deren Verteilungsfunktion spezifisch und bekannt ist.

GESCHICHTE

Der **Anpassungstest** für eine **Stichprobe** wurde von Andrey Nikolaevich Kol mogorov (1933) erfunden.
Der Kolmogorov-Smirnov-Test für zwei Proben wurde von Wladimir Iwanowitsch Smirnow (1939) erfunden.
In Massey (1952) finden wir eine Smirnov-Tabelle für den Kolmogorov-Smirnov-Test für zwei Stichproben, und in Miller (1956) finden wir eine Kol-Mogorov-Tabelle für den Anpassungstest.

MATHEMATISCHE ASPEKTE

Betrachten **Sie** zwei unabhängige Zufallsstichproben : (X1 , _ _ _ 2. Wir bezeichnen mit *F (x)* bzw. *G* (x) ihre unbekannten **Verteilungsfunktionen.**

Hypothesen

Die zu testenden Hypothesen lauten wie folgt:

A: Zweiseitiger Fall:

H0: *F (x)* = *G (x)* für jedes *x*

H1: *F (x)* = *G (x)* oder mindestens ein **Wert** von *x*

B: Einseitiger Fall:

H0: *F (x)* ÿ *G (x)* für jedes *x*

H1: *F (x)* > *G* (x) für mindestens einen Wert von *x*

C: Einseitiger Fall:

H0: *F (x)* ÿ *G (x)* für jedes *x*

H1: *F (x)* < *G* (x) für mindestens einen Wert von *x*

Im Fall A stellen wir die Hypothese auf, dass es keinen Unterschied zwischen den Verteilungen gibt

Butitionsfunktionen dieser beiden Populationen.

Beide Populationen können dann als eine Population betrachtet werden.

Im Fall B stellen wir die Hypothese auf, dass die Verteilungsfunktion von Population 1 kleiner ist als die von Population 2. Wir sagen manchmal, dass X im Allgemeinen kleiner als Y ist.

Im Fall C stellen wir die Hypothese auf, dass *X* größer als Y ist.
Wir bezeichnen mit H1 *(x)* die empirische Verteilungsfunktion der **Stichprobe** (X1, *X2,...,Xn)* und mit H2 *(x)* die empirische Verteilungsfunktion der Stichprobe (Y1, *Y2,...,Ym)* . Die statistischen Tests sind wie folgt definiert:

A: Zweiseitiger Fall

Der statistische Test T1 ist definiert als der größte vertikale **Abstand** zwischen zwei empirischen Verteilungsfunktionen:

$$T1 = \sup_x |H1\ (x) \text{ ÿ } H2\ (x)| .$$

B: One-Tail-Fall

Der statistische Test T2 ist definiert als der größte vertikale Abstand, wenn H1 *(x)* größer als H2 (x) ist:

$$T2 = \sup_x [H1\ (x) \text{ ÿ } H2\ (x)] .$$

C: One-Tail-Fall

Der statistische Test T3 ist definiert als der größte vertikale Abstand, wenn H2 *(x)* größer als H1 (x) ist:

$$T3 = \sup_x [H2\ (x) \text{ ÿ } H1\ (x)] .$$

Machine Translated by Google

Entscheidungsregel

Wir lehnen H0 auf dem **Signifikanzniveau** ÿ ab, wenn das geeigneter statistischer Test (T1, T2 oder T3) ist größer als der **Wert** der Smirnov-Tabelle mit den Parametern n, m und 1ÿÿ, die wir bezeichnen mit *tn,m,1ÿÿ,* also if

$$T1(\text{oder } T2\text{oder } T3) > tn,m,1ÿÿ .$$

Wenn wir die Anpassungsgüte eines testen wollen unbekannte Verteilungsfunktion F(x) einer Zufallsstichprobe aus einer Grundgesamtheit mit einer bestimmten und bekannten Verteilungsfunktion *Fo* (x), dann werden die Hypothesen dieselben sein Vortesten von zwei Proben, außerdassF *(x)* und *G (x)* werden durch *F (x)* und *Fo* (x) ersetzt . Wenn *H* (x) die empirische Verteilungsfunktion ist einer Zufallsstichprobe, dann die statistischen Tests

T1, T2 und T3 sind wie folgt definiert:

$$T1 = \sup_x |Fo\ (x) ÿ\ H\ (x)| ,$$

$$T2 = \sup_x [Fo\ (x) ÿ\ H\ (x)] ,$$

$$T3 = \sup_x [H\ (x) ÿ\ Fo\ (x)] .$$

Die Entscheidungsregel lautet wie folgt: H0 ablehnen das Signifikanzniveau ÿ, wenn T1 (oder T2 oder T3) ist größer als der Wert des Kolmogorov Tabelle mit den Parametern *n* und 1 ÿ ÿ, was wir mit tn,1ÿÿ bezeichnen , also if

$$T1(\text{oder } T2\text{oder } T3) > tn,1ÿÿ .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Um den Kolmogorov-Smirnov-Test durchzuführen, Folgendes ist zu beachten:

1. Beide Proben müssen stichprobenartig entnommen werden aus ihren jeweiligen **Populationen.**
2. Es muss gegenseitige **Unabhängigkeit** herrschen zwischen zwei Proben.
3. Die Maßskala muss mindestens die Ordnungszahl übertreffen.
4. Um einen exakten Test durchzuführen, müssen die Zufallsvariablen stetig sein; ansonsten der

Der Test ist weniger präzise.

BEISPIELE

Das erste Beispiel behandelt die Kolmogorov–Smirnov-Test für zwei Proben und die zweite eine für den **Anpassungstest.**
In einer Klasse zählen wir 25 Schüler: 15 Jungen und 10 Mädchen. Wir führen einen mentalen Rechentest durch, um zu sehen, ob die Jungen tendenziell besser sind als die Mädchen in dieser Domäne.
Die Daten werden in der folgenden Tabelle dargestellt; die höchsten Punktzahlen entsprechen den Ergebnissen des Tests.

Jungen (Xi)		Mädchen (Yi)	
19.8	17.5	17.7	14.1
12.3	17.9	7.1	23.6
10.6	21.1	21.0	11.1
11.3	16.4	10.7	20.3
13.3	7.7	8.6	15.7
14.0	15.2		
9.2	16.0		
15.6			

Wir testen die **Hypothese ,** nach der die Verteilungen der Ergebnisse der Mädchen und die der Jungen sind identisch. Das heisst dass die **Bevölkerung ,** aus der die Stichprobe stammt ofX istaken hat **die gleicheVerteilungsfunktion** tionasdie Bevölkerung, aus der die Stichprobe stammt von *Y* wird genommen. Daher die **Nullhypothese:**

$$H0 : F\ (x) = G\ (x) \text{ für jedes } x .$$

Wenn wir hier den zweiseitigen Fall anwenden, berechnen wir:

$$T1 = \sup_x |H1\ (x) ÿ\ H2\ (x)| ,$$

wobei H1 *(x)* und H2 *(x)* die empirischen Verteilungsfunktionen der Stichproben sind

Machine Translated by Google

(X1, *X2,...,X15)* und (Y1, *Y2,...,Y10),*
bzw. In der folgenden Tabelle haben wir
ordnete die Beobachtungen von zwei Proben ein
in aufsteigender Reihenfolge, um die Berechnungen zu vereinfachen
von H1 *(x)* ÿ H2 (x).

Xi	Yi	$H_1(x)$ ÿ $H_2(x)$
	7.1	(0 ÿ 1/10 = ÿ *0,1*
7.7		1/15 ÿ 1/10 = ÿ 0,0333
	8.6	1/15 ÿ 2/10 = ÿ 0,1333
9.2		2/15 ÿ 2/10 = ÿ 0,0667
10.6		3/15 ÿ 2/10 = 0
	10.7	3/15 ÿ 3/10 = ÿ *0,1*
	11.1	3/15 ÿ 4/10 = ÿ *0,2*
11.3		4/15 ÿ 4/10 = ÿ 0,1333
12.3		5/15 ÿ 4/10 = ÿ 0,0667
13.3		6/15 ÿ 4/10 = 0
14.0		7/15 ÿ 4/10 = 0,0667
	14.1	7/15 ÿ 5/10 = ÿ 0,0333
15.2		8/15 ÿ 5/10 = 0,0333
15.6		9/15 ÿ 5/10 = *0,1*
	15.7	9/15 ÿ 6/10 = 0
16.0		10/15 ÿ 6/10 = 0,0667
16.4		11/15 ÿ 6/10 = 0,1333
17.5		12/15 ÿ 6/10 = *0,2*
	17.7	12/15 ÿ 7/10 = *0,1*
17.9		13/15 ÿ 7/10 = 0,1667
19.8		14/15 ÿ 7/10 = 0,2333
	20.3	14/15 ÿ 8/10 = 0,1333
	21.0	14/15 ÿ 9/10 = 0,0333
21.1		1 ÿ 9/10 = *0,1*
	23.6	1 ÿ 1 = 0

Wir haben dann:

$$T1 = \sup_x |H1(x) - H2(x)|$$

$$= 0{,}2333 .$$

Der **Wert** der Smirnov-Tabelle für *n* =
15, *m* = 10 und 1 ÿ ÿ = 0,95 gleich
t15,10,0,95 = 0,5.
Somit ist T1 = 0,2333 < t15,10,0,95 = 0,5,
und H0 kann nicht abgelehnt werden. Das bedeutet, dass
Es gibt keinen signifikanten Unterschied im Niveau
des Kopfrechnens von Mädchen und Jungen.
Betrachten Sie die folgende Zufallsstichprobe von
Maß 10: X1 = 0,695, X2 = 0,937,
X3 = 0,134, X4 = 0,222, X5 = 0,239,
X6 = 0,763, X7 = 0,980, X8 = 0,322,
X9 = 0,523, X10 = 0,578.
Wir wollen durch die Kolmogorov– überprüfen
SmirnovtestiftDiese Stichprobe stammt aus **einer einheitlichen Formverteilung.** Die Verteilungsfunktion
Die Bestimmung der Gleichverteilung ergibt sich aus:

$$Fo(x) = \begin{cases} 0, \text{ wenn } x < 0 \\ x \text{ wenn } 0 \le x < 1 \\ 1 \text{ sonst .} \end{cases}$$

Die **Nullhypothese** H0 lautet dann wie folgt, wobei
F (x) die unbekannte Verteilungsfunktion der
zugehörigen Grundgesamtheit ist
zur Probe:

$$H0 : F(x) = Fo(x) \text{ für jedes } x .$$

Wenn der zweiseitige Fall angewendet wird, berechnen wir:

$$T1 = \sup_x |Fo(x) - H(x)| ,$$

wobei *H (x)* die empirische Verteilung ist
Funktion der Probe (X1, *X2,...,X10).*
In der folgenden Tabelle ordnen wir die 10
Beobachtungen zur Vereinfachung in aufsteigender Reihenfolge
Berechnung von F0 *(x)* ÿ *H* (x).

Xi	Fo (x)	H (x)	Fo (x) ÿ H (x)
0,134	0,134	0,1	0,134 ÿ *0,1* = 0,034
0,222	0,222	0,2	0,222 ÿ *0,2* = 0,022
0,239	0,239	0,3	0,239 ÿ *0,3* = ÿ 0,061
0,322	0,322	0,4	0,322 ÿ *0,4* = ÿ 0,078
0,523	0,523	0,5	0,523 ÿ *0,5* = 0,023
0,578	0,578	0,6	0,578 ÿ *0,6* = ÿ 0,022
0,695	0,695	0,7	0,695 ÿ *0,7* = ÿ 0,005
0,763	0,763	0,8	0,763 ÿ *0,8* = ÿ 0,037
0,937	0,937	0,9	0,937 ÿ *0,9* = 0,037
0,980	0,980	1,0	0,980 ÿ *1,0* = ÿ 0,020

Machine Translated by Google

Wir erhalten dann:

$$T1 = \sup_x \; |Fo(x) - H(x)| = 0{,}078 .$$

Der Wert der Kolmogorov-Tabelle für $n = 10$ und $1 - \alpha = 0{,}95$ beträgt $t_{10;0{,}95} = 0{,}409$.
Wenn T1 kleiner als $t_{10;0{,}95}$ (0,078 < 0,409) ist, kann H0 nicht abgelehnt werden. Das bedeutet, dass die Zufallsstichprobe aus einer gleichmäßig verteilten Grundgesamtheit stammen könnte.

WEITERLESEN

- **Anpassungstest**
- **Hypothesentest**
- **Nichtparametrischer Test**

REFERENZ

Kolmogorov, AN: Die empirische Bestimmung einer Vertriebslinie.
Giornale dell'Instituto Italiano degli Attuari 4, 83–91 (6.1) (1933)

Massey, FJ: Verteilungstabelle für die Abweichung zwischen zwei Stichprobenkumulationen.
Ann. Mathematik. Stat. **23,** 435–441 (1952)

Miller, LH: Tabelle der Prozentpunkte der Kolmogorov-Statistik. Marmelade. Stat. Assoc. **31,** 111–121 (1956)

Smirnov, NV: Schätzung der Abweichung zwischen empirischen Verteilungsfunktionen in zwei unabhängigen Stichproben. (Russisch).
Stier. Moskauer Universität. 2(2), 3–16 (6.1, 6.2) (1939)

Smirnov, NV: Tabelle zur Schätzung der Anpassungsgüte empirischer Verteilungen.
Ann.Math. Stat. **19,** 279–281 (6.1) (1948)

Kruskal-Wallis-Tisch

Die Kruskal-Wallis-Tabelle gibt die theoretischen Werte der **Statistik** *H* der **Kruskal-Wallis-Tabelle** an.

Wallis-Test unter der **Hypothese ,** dass dort

Es gibt keinen Unterschied zwischen den *k (k* ≥ 2) Populationen, die wir vergleichen wollen.

GESCHICHTE

Siehe **Kruskal-Wallis-Test.**

MATHEMATISCHE ASPEKTE

Sei *k* die Anzahl der Stichproben wahrscheinlich unterschiedlicher Größe n1, *n2,...,nk.* Mit *N* bezeichnen wir die Gesamtzahl der Beobachtungen:

$$N = \sum_{i=1}^{k} ni .$$

Wir ordnen die *N* Beobachtungen in aufsteigender Reihenfolge, ohne zu berücksichtigen, zu welchen Stichproben sie gehören. Anschließend vergeben wir Rang 1 für den kleinsten **Wert,** Rang 2 für den nächstgrößeren Wert usw. bis Rang N, der dem größten Wert zugewiesen wird.

Wir bezeichnen mit *Ri* die Summe der Ränge, die den Beobachtungen der Stichprobe i gegeben wurden:

$$Ri = \sum_{j=1}^{ni} R\,Xij , \qquad i = 1, 2,...,k,$$

wobei *Xij* die Beobachtung *j* der Stichprobe *i* und *R(Xij)* den entsprechenden Rang darstellt . Wenn viele Beobachtungen identisch sind und den gleichen Rang haben, geben wir ihnen einen mittleren Rang (siehe **Kruskal-Wallis-Test).** Liegen keine Durchschnittsränge vor, wird der statistische Test wie folgt definiert:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{\bar{R}2}{ni} - 3(N+1) .$$

Was bei mittleren Rängen zu tun ist, erfahren Sie im **Kruskal-Wallis-Test.**

Die Kruskal-Wallis-Tabelle gibt die Werte der **Statistik** *H* des Kruskal-Wallis-Tests in an

im Fall von drei Stichproben für unterschiedliche Werte von n1, n2 und n3 (mit n1, n2, n3 ≥ 5).

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Für Nichtberechnungen wird die Kruskal-Wallis-Tabelle verwendet parametrische Tests, die Ränge verwenden, insbesondere für Tests mit demselben Namen. Wenn die Anzahl der Stichproben *i* größer als 3 ist, können wir den **Wert** der Kruskal-Wallis-Tabelle durch die **Chi-Quadrat-Tabelle** mit kÿ1 Freiheitsgraden annähern .

BEISPIELE

Siehe Anhang D.
Ein Beispiel für die Verwendung der Kruskal-Wallis-Tabelle finden Sie unter **Kruskal-Wallis-Test.**

WEITERLESEN

Chi-Quadrat-Tisch
Kruskal-Wallis-Test
Statistische Tabelle

LITERATUR:

Kruskal, WH, Wallis, WA: Verwendung von Rängen in der Einkriterium-Varianzanalyse. Marmelade. Stat. Assoc. **47,** 583–621 und Errata, ebenda. **48,** 907–911 (1952)

Kruskal-Wallis-Test

Der Kruskal-Wallis-Test ist ein **nichtparametrischer Test ,** dessen Ziel darin besteht, festzustellen, ob alle *k* Populationen identisch sind oder ob mindestens eine der Populationen dazu neigt, Beobachtungen zu liefern, die sich von denen anderer Populationen unterscheiden.

Der Test wird verwendet, wenn wir *k* Stichproben mit *k* ÿ 2 haben , die aus *k* Populationen stammen , die unterschiedlich sein können.

GESCHICHTE

Der Kruskal-Wallis-Test wurde 1952 von Kruskal, WH und Wallis, WA entwickelt

MATHEMATISCHE ASPEKTE

Die Daten werden in *k* Stichproben dargestellt . Wir bezeichnen mit *ni* die Dimension der **Stichprobe** i, fori = *1,...,k,* und mit *N* die Gesamtzahl der Beobachtungen:

$$N = \sum_{i=1}^{k} ni .$$

Wir klassifizieren die *N* Beobachtungen in aufsteigender Reihenfolge, ohne zu berücksichtigen, ob sie zu denselben Stichproben gehören oder nicht. Dann geben wir dem kleinsten Wert Rang 1, dem nächstgrößten Wert Rang 2 und so weiter bis N, das dem größten Wert zugewiesen wird.
Sei *Xij* die j-te Beobachtung der Stichprobe i und setze *i = 1,...,k* und *j = 1,...,ni;* Wir bezeichnen dann den Rang, den *Xij* erhält, mit *R Xij* .
Wenn viele Beobachtungen den gleichen Wert haben, geben wir ihnen einen mittleren Rang. Die Summe der Ränge, die den Beobachtungen der Stichprobe *i* gegeben werden, wird mit *Ri* bezeichnet , und wir haben:

$$Ri = \sum_{j=1}^{ni} R\,Xij , \quad i = 1,...,k.$$

Wenn es keine mittleren Ränge gibt (oder wenn es eine begrenzte Anzahl davon gibt), dann ist der statistische Test wie folgt definiert:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R2}{ni} \quad \ddot{y}\ 3\,(N+1) .$$

Wenn dagegen viele mittlere Ränge vorhanden sind, muss eine Korrektur vorgenommen und Folgendes berechnet werden:

$$H = \frac{H}{1 \ddot{y} \dfrac{\sum_{i=1}^{g} 3\,t \ddot{y}\, ti\,i}{N3 \ddot{y}}} ,$$

N wobei *g* die Anzahl der Gruppen mittlerer Ränge und *ti* die Dimension i -ter dieser Gruppe ist.

Hypothesen

Das Ziel des Kruskal-Wallis-Tests besteht darin, festzustellen, ob alle Populationen identisch sind

Machine Translated by Google

oder wenn zumindest eine der Populationen dazu neigt
Geben Sie Beobachtungen ab, die sich von denen anderer
Bevölkerungsgruppen unterscheiden. Die Hypothesen lauten wie folgt:

H0: Es gibt keinen Unterschied zwischen den Kpops
ulationen.

H1: Mindestens eine der Populationen unterscheidet sich
von den anderen Populationen.

Entscheidungsregel

Wenn es 3 Proben gibt, von denen jede eine Abmessung
kleiner oder gleich 5 hat, und wenn keine vorhanden sind
mittlere Ränge (das heißt, wenn *H* berechnet wird), dann
Wir verwenden die **Kruskal-Wallis-Tabelle ,** um H0 zu testen .
Die Entscheidungsregel lautet: Wir lehnen ab
die **Nullhypothese** H0 bei der **Signifikanz**
Stufe ÿ, wenn *T* größer als der Wert ist
Tabelle mit Parametern *ni, k* ÿ 1 und 1 ÿ ÿ,
mit *hn1,n2,n3,1ÿÿ* bezeichnet , und wenn keine exakte
Tabelle verfügbar ist oder mittlere Ränge vorhanden sind, wir
kann eine Annäherung an den Wert von machen
die Kruskal-Wallis-Tabelle durch die Verteilung
des Chi-Quadrats mit *k* ÿ 1 Freiheitsgraden **(Chi-Quadrat-**
Verteilung), das heißt, wenn:

$H > hn1,n2,n3,1ÿÿ$ **(Kruskal-Wallis-Tabelle)**

oder $H > ÿ2_{kÿ1,1ÿÿ}$ **(Chi-Quadrat-Tabelle) .**

Die entsprechende Entscheidungsregel basiert auf
$H > ÿ2_{kÿ1,1ÿÿ}$ (Chi-Quadrat-Tisch).

DOMÄNEN UND EINSCHRÄNKUNGEN

Die folgenden Regeln sollten beachtet werden
Machen Sie den Kruskal-Wallis-Test:

1. Bei allen Stichproben muss es sich um Zufallsstichproben handeln
   aus ihren jeweiligen Populationen entnommen .
2. Zusätzlich zur **Unabhängigkeit** im Inneren
   Bei jeder Probe muss eine gegenseitige
   Unabhängigkeit zwischen den verschiedenen Proben bestehen.
3. Der Maßstab muss mindestens ordi sein
   nal.

Wenn der Kruskal-Wallistest uns dazu bringt, das abzulehnen
Nullhypothese H0 können wir den **Wilcox-Test** für alle
paarweise entnommenen Proben verwenden
Bestimmen Sie, welche Populationspaare dazu neigen
Sei anders.

BEISPIELE

Wir kochen Kartoffeln in 4 verschiedenen Ölen. Wir wollen
um zu überprüfen, ob die aufgenommene Fettmenge durch
Kartoffeln hängt von der Art des verwendeten Öls ab. Wir
Führen Sie 5 verschiedene Experimente mit Öl 1, 6 durch
mit Öl 2, 4 mit Öl 3 und 5 mit Öl 4 und
Wir erhalten folgende Daten:

Ölsorte			
1	2	3	4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85		70
	77		

In diesem Beispiel die Anzahl der Proben
gleich 4 *(k* = 4) mit folgenden jeweiligen Dimensionen:

n1 = 5 ,
n2 = 6 ,
n3 = 4 ,
n4 = 5 .

Die Anzahl der Beobachtungen beträgt:

$$N = 5 + 6 + 4 + 5 = 20 .$$

Wir ordnen die Beobachtungen in aufsteigender Reihenfolge
und geben Sie ihnen einen Rang von 1 bis 20
Berücksichtigen Sie die mittleren Ränge. Wir erhalten folgendes:
Tabelle mit der Rangfolge der Beobachtungen
in Klammern und am Ende jedes Beispiels
die Summe der Ränge, die der entsprechenden Stichprobe
gegeben wurden:

Machine Translated by Google

Ölsorte			
1	2	3	4
64 (4,5)	78 (14,5)	75 (11)	55 (2)
72 (10)	91 (18)	93 (19)	66 (6)
68 (7)	97 (20)	78 (14,5)	49 (1)
77 (12,5)	82 (16)	71 (9)	64 (4,5)
56 (3)	85 (17)		70 (8)
	77 (12,5)		
R1 = 37	R2 = 98	R3 = *53,5*	R4 = *21,5*

Wir berechnen:

$$H = \frac{12}{N\,(N+1)} \sum_{i=1}^{k} \frac{R2}{ni} \; ÿ \; 3\,(N+1)$$

$$= \frac{12}{20\,(20+1)} \cdot \frac{372}{5} + \frac{982}{6} + \frac{53{,}52}{4} + \frac{21.52}{5} \; ÿ \; 3\,(20+1)$$

$$= 13{,}64\,.$$

Wenn wir eine Anpassung vornehmen, um dies zu berücksichtigen

Für die mittleren Ränge erhalten wir:

$$H = \frac{H}{1 \; ÿ \; \frac{\sum_{i=1}^{G} T^3 \; ÿ \; ti}{N3 \; ÿ \; N}}$$

$$= \frac{13.64}{1 \; ÿ \; \frac{8 \; ÿ \; 2 + 8 \; ÿ \; 2 + 8 \; ÿ \; 2}{203 \; ÿ \; 20}}$$

$$= 13{,}67\,.$$

Wir sehen, dass der Unterschied zwischen *H* und *H* ist minimal.

Die Hypothesen lauten wie folgt:

H0: Es gibt keinen Unterschied zwischen den vier Öle.

H1: Mindestens eines der Öle unterscheidet sich vom Andere.

Die Entscheidungsregel lautet wie folgt: Ablehnen **Nullhypothese** H0 auf dem **Signifikanzniveau** ÿ wenn

$$H > ÿ2_{kÿ1,1ÿÿ}\,,$$

wobei $ÿ2_{kÿ1,1ÿÿ}$ ist der **Wert** des **Chi quadratischer Tisch** auf Ebene 1 ÿ ÿ und *k* ÿ 1 = 3 Freiheitsgrade.

Wenn wir ÿ = 0,05 wählen, dann ist der Wert von $ÿ2_{3,0,95}$ 7,81 beträgt und wenn *H* größer als ÿ2 ist $_{3,0,95}$ *(13.64* > 7.81), dann lehnen wir die Hypothese H0 ab.

Wenn H0 abgelehnt wird, können wir das Verfahren verwenden von mehreren Vergleichen, um zu sehen, welche Paare der Öle sind unterschiedlich.

WEITERLESEN

- Hypothesentest
- Nichtparametrischer Test
- Wilcoxon-Test

REFERENZ

Kruskal, WH, Wallis, WA: Verwendung von Dienstgraden in der Einkriterium-Varianzanalyse. Marmelade. Stat. Assoc. **47,** 583–621 und Errata, ebenda. **48,** 907–911 (1952)

Machine Translated by Google

L1 Schätzung

Die L1- Schätzmethode ist eine Alternative zur Methode der **kleinsten Quadrate ,** die in der linearen Regression zur **Schätzung** der **Parameter** eines **Modells verwendet wird.** Anstatt die **Restquadratsumme** zu minimieren , wird die Summe der **Restabsolutwerte** minimiert.

Die L1- **Schätzer** haben den Vorteil, dass sie nicht so empfindlich auf Ausreißer reagieren wie die Methode der **kleinsten Quadrate .**

GESCHICHTE Die Verwendung der L1- Schätzmethode geht in einer sehr einfachen Form auf Galileo im Jahr 1632 zurück, der sie zur Bestimmung der **Entfernung** zwischen der Erde und einem kürzlich entdeckten Stern verwendete.

Boscovich, RJ (1757) schlug zwei Kriterien für die Anpassung einer Linie

vor: • Die jeweiligen Summen der positiven und negativen **Residuen** müssen gleich sein und die optimale Linie verläuft durch den Schwerpunkt der **Beobachtungen.**

• Die Summe der Residuen in absoluten Werten muss minimal sein.

Boscovich, RJ schlug eine geometrische Methode zur Lösung der Gleichungen vor, die sich aus seinen Kriterien ergaben. Das zur **Schätzung** der **Parameter** für die beiden Kriterien von Boscovich, RJ verwendete Analyseverfahren geht auf **Laplace, PS** (1793) zurück.

Edgeworth, FY (1887) unterdrückte die erste Bedingung, indem er sagte, dass die Summe der **Residuen** null sein muss, und entwickelte eine Methode, die zur Behandlung des Falles einer **multiplen linearen Regression verwendet werden konnte.**

In der ersten Hälfte des 20. Jahrhunderts waren Statistiker nicht sehr an der linearen Regression auf Basis der L1- Schätzung interessiert. Die Hauptgründe waren:

• Die numerischen Schwierigkeiten aufgrund der Fehlen einer geschlossenen Formel.

• Das Fehlen einer asymptotischen Theorie für die L1- Schätzung im Regressionsmodell . • Die Schätzung mit der Methode der **kleinsten Quadrate** ist für **kleine Stichproben** zufriedenstellend und es ist nicht sicher, ob die L1- Schätzung besser ist.

Ab 1970 wurde das Gebiet der L1- Schätzung zum Forschungsschwerpunkt einer Vielzahl von Statistikern, und es wurden zahlreiche Arbeiten veröffentlicht. Wir erwähnen Basset, G. und Koenker, R. (1978), die die asymptotische Theorie der L1- Schätzer entwickelt haben. 1987 fand in Neuchâtel (Schweiz) ein diesen Problemen gewidmeter Kongress statt: Die erste internationale Konferenz über statistische Daten

Analyse basierend auf der L1-Norm und verwandten Methoden.

MATHEMATISCHE ASPEKTE

Im eindimensionalen (1D) Fall besteht die L1 **-Schätzmethode** darin, die zu schätzen

Machine Translated by Google

Parameter ÿ0 im **Modell**

$$yi = \ddot{y}0 + \ddot{y}i ,$$

also die Summe der **Fehler** im absoluten Wert ist minimal:

$$\min_{\ddot{y}0} \sum_{i=1}^{N} |\ddot{y}i| = \min_{\ddot{y}0} \sum_{i=1}^{N} |yi \ddot{y} \ddot{y}0| .$$

Dieses Problem läuft darauf hinaus, ÿ0 so zu finden $\sum_{i=1}^{N} |yi \ddot{y} \ddot{y}0|$ ist minimal. Der Wert von ÿ0 das minimiert diese Summe und ist der **Median** der 1D **Beispiel** y1, *y2,...,yn.*

Im **einfachen linearen** Regressionsmodell

$$yi = \ddot{y}0 + \ddot{y}1 \cdot xi + \ddot{y}i ,$$

Die mit der L1 **-Schätzmethode** geschätzten **Parameter** ÿ0 und ÿ1 werden üblicherweise durch einen iterativen **Algorithmus** berechnet , der zur Minimierung eingesetzt wird

$$\sum_{i=1}^{N} |yi \ddot{y} \ddot{y}0 \ddot{y} \ddot{y}1 \cdot xi| .$$

Mithilfe **linearer Programmiertechniken** ist es heute möglich, die L1- **Schätzwerte** auf relativ einfache Weise zu berechnen , insbesondere **im multiplen linearen Regressionsmodell .**

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Problem der L1 -Schätzung liegt im Fehlen einer allgemeinen Formel für die **Schätzung** der **Parameter,** anders als bei der Methode der **kleinsten Quadrate .** Für einen Satz von 1D- **Daten** ist das durch die L1- Schätzung ermittelte **Maß der zentralen Tendenz** der **Median,** der per Definition der **Wert** ist , der die Summe der absoluten Abweichungen minimiert.

WEITERLESEN

- **Fehler**
- **Einschätzung**
- **Schätzer**
- **Kleinsten Quadrate**
- **Mittlere absolute Abweichung**
- **Parameter**
- **Restwert**
- **Robuste Schätzung**

REFERENZEN

Basset, G., Koenker, R.: Asymptotische Theorie der Regression des kleinsten absoluten Fehlers. Marmelade. Stat. Assoc. **73,** 618–622 (1978)

Boscovich, RJ: De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex exemplaria etiam sensorum impres sa. Bononiensi Scientiarium et Artium Instituto Atque Academia Commentarii, vol. IV, S. 353–396 (1757)

Dodge, Y. (Hrsg.): Statistische Datenanalyse basierend auf der L1-Norm und verwandten Methoden. Elsevier, Amsterdam (1987)

Edgeworth FY: Eine neue Methode zur Reduzierung von Beobachtungen, die sich auf mehrere Größen beziehen. Philos.Mag. (5. Ser) **24,** 222–223 (1887)

Lagrange-Multiplikator

Die Lagrange-Multiplikatormethode ist eine klassische **Optimierungsmethode ,** die es ermöglicht, die lokalen Extrema einer Funktion zu bestimmen bestimmten Einschränkungen unterliegen. Es ist nach dem italienisch-französischen Mathematiker und benannt Astronom Joseph-Louis Lagrange.

MATHEMATISCHE ASPEKTE

Sei *f(x,* y) die Zielfunktion, die abhängig davon maximiert oder minimiert werden soll

Machine Translated by Google

g(x, y) = 0. Wir bilden die Hilfsfunktion, die Lagrangefunktion:

$$F(x, y, \ddot{y}) = f(x, y) \ddot{y} \ddot{y} g(x, y) ,$$

wobei ÿ eine unbekannte **Variable** namens ist Lagrange-Multiplikator. Wir suchen nach den Minima oder Maxima der Funktion *F* , indem wir die partiellen Ableitungen nach x, y und ÿ zu Null setzen:

$$\frac{\ddot{y}F}{\ddot{y}x} = \frac{\ddot{y}f \ddot{y}g \ddot{y} \ddot{y}}{\ddot{y} \ddot{y} = 0 \ddot{y}x \ddot{y}x \ddot{y}f \ddot{y}g}$$

$$\frac{\ddot{y}F}{\ddot{y}y} = \frac{\ddot{y}y \ddot{y}F = g(x,}{y) = 0 .}$$

$$\overline{\ddot{y}\ddot{y}}$$

Diese drei Gleichungen können nach x, y und ÿ gelöst werden. Die Lösung liefert die Punkte, an denen die eingeschränkte Funktion ein Minimum oder hat maximal. Wir müssen entscheiden, ob es so ist ein Maximum oder ein Minimum.

Wir erhalten ein Maximum, wenn

$$\frac{\ddot{y}2F}{\ddot{y}x2} \cdot \frac{\ddot{y}2F}{\ddot{y}y2} - \frac{\ddot{y}2F}{\ddot{y}x\ddot{y}y}^{2} > 0 ,$$

$$\ddot{y}2F \ddot{y}2F < 0 \text{ und } < 0 .$$

$$\overline{\quad} \ddot{y}x2 \ddot{y}y2 \overline{\quad}$$

Wir erhalten ein Minimum, wenn

$$\frac{\ddot{y}2F}{\ddot{y}x2} \cdot \frac{\ddot{y}2F}{\ddot{y}y2} - \frac{\ddot{y}2F}{\ddot{y}x\ddot{y}y}^{2} > 0 ,$$

$$\frac{\ddot{y}2F}{\longrightarrow} > 0 \text{ und } \ddot{y}x2 \quad \frac{\ddot{y}2F}{\ddot{y}y2} \longrightarrow > 0 .$$

Wenn

$$\frac{\ddot{y}2F}{\ddot{y}x2} \cdot \frac{\ddot{y}2F}{\ddot{y}y2} - \frac{\ddot{y}2F}{\ddot{y}x\ddot{y}y}^{2} \ddot{y} 0 ,$$

Der Test schlägt fehl, und um festzustellen, ob es ein Minimum oder ein Maximum gibt, müssen wir die in *(x,* y) benachbarten Funktionen untersuchen.

Die Lagrange-Multiplikatormethode wird auf eine Funktion von *n* Variablen *f* (x1, *x2,...,xn) erweitert,* die *k* Einschränkungen *gj* (x1, *x2,...,xn)* = 0, *j = 1* unterliegen , *2,...,k* mit *k* ÿ n. In ähnlicher Weise definieren wir im bivariaten Fall die Lagrange-Funktion:

$$F = f(x1, x2,...,xn) \quad k \quad \ddot{y}j \cdot gj(x1, x2,...,xn) ,$$

$$j=1$$

wobei die Parameter *ÿ1,..., ÿk* die Lagrange-Multiplikatoren sind. Wir stellen die partiellen Ableitungen auf, die zu *n + k* Gleichungen für *n + k* unbekannte Parameter führen. Diese führen zu den Koordinaten der Funktionsextreme (Minimum oder Maximum), falls vorhanden.

DOMÄNEN UND EINSCHRÄNKUNGEN

Es ist nicht möglich, die Lagrange-Multiplikator-Methode anzuwenden, wenn die Funktionen nicht differenzierbar sind oder wenn die auferlegten Einschränkungen Ungleichungen sind. In diesem Fall können wir andere Optimierungsmethoden verwenden, die **aus** der linearen Programmierung ausgewählt wurden .

BEISPIELE

Lassen

$$f(x, y) = 5x2 + 6y2 \ddot{y} x \cdot y$$

sei die Zielfunktion und

$$g(x, y) = x + 2y \ddot{y} 24 = 0$$

der auferlegte Zwang. Wir bilden den Lag-Rangian

$$F(x, y, \ddot{y}) = f(x, y) \ddot{y} \ddot{y} g(x, y) = 5x2 + 6y2 \ddot{y} x \cdot y \ddot{y} \ddot{y} \cdot (x + 2y \ddot{y} 24) ,$$

und wir setzen die partiellen Ableitungen auf Null

$$\frac{\ddot{y}F}{\ddot{y}x} = 10x \ddot{y} y \ddot{y} \ddot{y} = 0 , \quad (1)$$

Machine Translated by Google

$$\frac{\ddot{y}F}{} = 12y \ddot{y} x \ddot{y} 2\ddot{y} = 0 \quad \ddot{y}y \ddot{y}F , \qquad (2)$$

$$\frac{}{\ddot{y}\ddot{y}} = x + 2y \ddot{y} 24 = 0 . \qquad (3)$$

Durch Streichen von ÿ aus (1) und (2):

$$20x \ddot{y} 2y \ddot{y} 2\ddot{y} = 0$$
$$\ddot{y}x + 12y \ddot{y} 2\ddot{y} = 0$$
$$21x \ddot{y} 14y = 0$$
$$3x \ddot{y} 2y = 0$$

und durch Ersetzen *von 2y* durch *3x* in (3):

$$0 = x + 2y \ddot{y} 24 = x + 3x \ddot{y} 24$$
$$= 4x \ddot{y} 24$$

wir erhalten $x = 6$. Indem wir **diesen x-Wert** in (3) ersetzen, wir erhalten $y = 9$. Der kritische Punkt hat die Koordinaten (6, 9). Wir berechnen die partiellen Ableitungen zweiter Ordnung, um zu überprüfen, ob dies der Fall ist ein Maximum oder ein Minimum. Wir bekommen:

$$\frac{\ddot{y}2F}{\ddot{y}x2} = 10 , \quad \frac{\ddot{y}2F}{\ddot{y}y2} = 12 , \quad \frac{\ddot{y}2F}{\ddot{y}x\ddot{y}y} = \ddot{y}1 ,$$

$$\frac{\ddot{y}2F}{\ddot{y}x2} \cdot \frac{\ddot{y}2F}{\ddot{y}y2} \cdot \frac{\ddot{y}2F}{\ddot{y}x\ddot{y}y}^{2}$$
$$= 10 \cdot 12 \ddot{y} (\ddot{y}1) \quad 2 = 119 > 0 ,$$

$$\frac{\ddot{y}2F}{\ddot{y}x2} > 0 , \quad \frac{\ddot{y}2F}{\ddot{y}y2} > 0 .$$

Beachten Sie, dass der Punkt (6, 9) entspricht ein Minimum für die Zielfunktion *f* vorbehaltlich der oben genannten Einschränkungen. Dieses Minimum ergibt f(6, 9) = 180 + 486 ÿ 54 = 612.

WEITERLESEN

- **Lineares Programmieren**
- **Optimierung**

VERWEISE

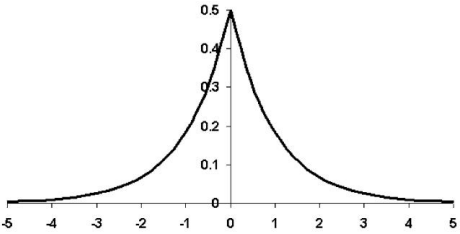
Dodge, Y. (2004) Optimierungsapplikation. Springer, Paris

Laplace-Verteilung

Die **Zufallsvariable** *X* folgt einem Laplace Verteilung, wenn ihre **Dichtefunktion** von der ist folgendes Formular:

$$f(x) = 2 \frac{1}{} \cdot \exp \quad \ddot{y}|x \ddot{y} \ddot{y}| \quad > 0 ,$$

wo ist der Dispersionsparameter **und** ÿ ist der **erwartete Wert.**



Laplace-Verteilung, ÿ = 0, = 1

Die Laplace-Verteilung ist symmetrisch um seinen erwarteten Wert ÿ, der auch der ist **Modus** und der **Median** der Verteilung.

Die Laplace-Verteilung ist **stetig Wahrscheinlichkeitsverteilung.**

GESCHICHTE

Diese **kontinuierliche Wahrscheinlichkeitsverteilung** trägt den Namen seines Autors. Im Jahr 1774 **Laplace, PS** hat einen grundlegenden Artikel geschrieben über symmetrische Verteilungen im Hinblick auf Beschreibung der **Messfehler .**

MATHEMATISCHE ASPEKTE

Der **erwartete Wert** einer **Zufallsvariablen** Gegeben ist *X* nach einer Laplace-Verteilung von:

$$E[X] = \ddot{y}$$

und die **Varianz** ist gleich:

$$\text{Var}(X) = 2 \cdot \quad ^{2} .$$

Machine Translated by Google

Die Standardform der **Dichtefunktion** der Laplace-Verteilung lautet:

$$f(x) = \ddot{y} \frac{1}{\ddot{y}2} \cdot \exp \ddot{y} \ddot{y} 2 \cdot \frac{|x \ddot{y} \ddot{y}|}{\ddot{y}} \quad \ddot{y} > 0 ,$$

mit

$$\ddot{y} = \ddot{y} \text{ und } \ddot{y} = \ddot{y} 2 \cdot .$$

Dies ermöglicht es einem, die **Verteilung** zu schreiben **Funktion** der angegebenen Laplace-Verteilung von:

$$F(x) = \begin{cases} \ddot{y} \frac{1}{2} \exp \frac{x \ddot{y} \ddot{y}}{} & \text{wenn } x \ddot{y} , \\ 1 \ddot{y} \frac{1}{\exp 2} \ddot{y} \frac{\ddot{y} x \ddot{y} \ddot{y}}{} & \text{wenn } x \ddot{y} \ddot{y} . \end{cases}$$

WEITERLESEN

Kontinuierliche Wahrscheinlichkeitsverteilung

VERWEISE

Laplace, PS de: Erinnerung an die wahrscheinliche Ursache der Ereignisse. Mem. Acad. Roy. Wissenschaft. (vorgetragen von verschiedenen Wissenschaftlern) 6, 621–656 (1774) (oder Laplace, PS de (1891) Œuvres complètes, Bd. 8. Gauthier-Villars, Paris, S. 27–65)

Laplace, Pierre Simon De

1749 in Frankreich als Sohn einer bürgerlichen Familie geboren ily, Marquis Pierre Simon de Laplace einer der Pioniere der **Statistik.** Interessiert an Mathematik, theoretischer Astronomie, **Zu den Themen Wahrscheinlichkeit** und Statistik erschienen seine ersten Veröffentlichungen in den frühen 1770er Jahren. Er wurde ein Mitglied der Akademie der Wissenschaften in 1785, und er leitete auch die Longitudes Büro. 1799 wurde er zum Minister ernannt Innere Angelegenheiten von Bonaparte, der 1806 ihm den Grafentitel verliehen. Gewählt 1816 ging er an die Französische Akademie seine wissenschaftliche Arbeit bis zu seinem Tod in Paris, im Jahr 1827.

Einige Hauptwerke und Artikel von de La Place, PS:

1774 Memoiren über die Wahrscheinlichkeit der Ursachen durch Ereignisse. Oeuvres complètes, Bd. 8, S. 27–65 (1891). Gau Thier-Villars, Paris.

1781 Mémoire sur les probabilités. Oeu vres complètes, Bd. 9 S. 383–485 (1891). Gauthier-Villars, Paris.

1805 Traité de Mécanique Céleste. Bd. 1–4. Duprot, Paris. Bd. 5. Bachelier, Paris.

1812 Théorie analytique des probabilités. Courcier, Paris.

1878–1912 Gesamtwerk. 14 Bände. Gauthier-Villars, Paris.

Laspeyres-Index

Der Laspeyres- **Index** ist ein **zusammengesetzter Index** **Anzahl** der Preise, die nach der Methode der gewichteten Summe ermittelt wurden. Diese Indexzahl stellt das Verhältnis der Summe der Preise im dar Istperiode *n* zur Preissumme in der Referenzperiode 0 zu addieren, wobei diese Summen mit den jeweiligen Mengen der Referenzperiode gewichtet werden. Daher die Indexnummer misst die relative Preisänderung der Güter, wobei die jeweiligen Mengen als unverändert betrachtet werden. Der Laspeyres-Index unterscheidet sich vom **Paasche-Index** in der Auswahl

Machine Translated by Google

Die Gewichte: Im Paasche-Index erfolgt die Gewichtung anhand der verkauften Mengen der aktuellen Periode *Qn,* während im Laspeyres Index, es erfolgt durch die verkauften Mengen in der Referenzzeitraum Q0.

GESCHICHTE

Laspeyres, Etienne, deutscher Ökonom und Statistiker französischer Herkunft, umformuliert die von Lowe entwickelte Indexzahl würde als Laspeyres-Index bekannt werden in der Mitte des 19. Jahrhunderts. In den Jahren 1864–1871 Laspeyres, Etienne arbeitete daran Waren und ihre Preise in Hamburg (he ausgeschlossene Leistungen).

MATHEMATISCHE ASPEKTE

Der Laspeyres- **Index** wird wie folgt berechnet:

$$\text{In/0} = \frac{Pn \cdot Q0}{P0 \cdot Q0},$$

wobei Pn und Qn die Preise und Mengen sind in der aktuellen Periode verkauft und P0 und Q0 sind die in der Referenz verkauften Preise und Mengen Zeitraum, ausgedrückt in Basis 100:

$$\text{In/0} = \frac{Pn \cdot Q0}{P0 \cdot Q0} \cdot 100 .$$

Bei der Summe handelt es sich um die betrachtete Ware. Auch das Laspeyres-Modell kann dazu verwendet werden Berechnen Sie einen Mengenindex (auch Volumenindex genannt). In diesem Fall sind es die Preise sind Konstanten und die Größen sind Variablen:

$$\text{In/0} = \frac{Qn \cdot P0}{Q0 \cdot P0} \cdot 100 .$$

BEISPIELE

Betrachten Sie die folgende Tabelle, die Folgendes angibt jeweilige Preise von drei Konsumgütern in Referenzjahr 0 und im laufenden Jahr n, as sowie die in der Referenz verkauften Mengen Jahr:

Produkt	verkaufte Menge im Jahr 1970 (Q_0) (Tausende)	Preis (Euro) 1970 (P_0)	1988 (Pn)
Milch	50,5	0,20	1,20
Brot	42,8	0,15	1,10
Butter	15.5	0,50	2,00

Aus der folgenden Tabelle haben wir

$$PnQ0 = 138,68 \text{ und } P0Q0 = 24,27 .$$

Produkt	PnQ_0	P_0Q_0
Milch	60,60	10.10
Brot	47.08	6.42
Butter	31.00	7,75
Gesamt	138,68	24.27

Wir können dann den Laspeyres-Index finden:

$$\text{In/0} = \frac{Pn \cdot Q0}{P0 \cdot Q0} \cdot 100$$

$$= \frac{138,68}{24.27} \cdot 100 = 571,4 .$$

Mit anderen Worten, laut den Laspeyres Index, der Preisindex dieser Güter hat stieg im Laufe des Jahres um 471,4 % (571,4 ÿ 100). betrachteter Zeitraum.

WEITERLESEN

Zusammengesetzte Indexnummer
Fisher-Index
Indexnummer
Paasche-Index

VERWEISE

Laspeyres, E.: Hamburger Warenpreise 1850–1863 und die kalifornisch-austra lischen Geldentdeckung seit 1848. Jahrb. Natl. Stat. 3, 81–118, 209–236 (1864)

Machine Translated by Google

Laspeyres, E.: Die Berechnung einer mit höherer Preissteigerung. Jahrb. Natl. Stat. **16,** 296–314 (1871)

Lateinischer Platz

Siehe **lateinische Quadratdesigns.**

Lateinische quadratische Designs

Ein lateinisches Quadratdesign ist die Anordnung von *t* Behandlungen, die jeweils *t* Mal wiederholt werden, sodass jede Behandlung genau einmal in jeder Zeile und jeder Spalte im Design erscheint. Wir bezeichnen die Behandlungen mit römischen Buchstaben. Daher wird das Design als lateinisches Quadratdesign bezeichnet. Diese Art von Design wird verwendet, um systematische Fehler aufgrund von Zeilen (Behandlungen) und Spalten zu reduzieren.

GESCHICHTE

Laut Preece (1983) reicht die Geschichte des lateinischen Quadrats bis ins Jahr 1624 zurück. In der **Statistik** führte **Fisher, Ronald Aylmer** (1925) die lateinischen Quadratdesigns ein.

BEISPIELE

Ein Bauer hat auf seinem Grundstück zwei Felder, auf denen er Mais anbauen möchte. Er möchte ein Experiment mit seinen vier verschiedenen Maissorten **(Behandlungen) durchführen.** Darüber hinaus weiß er, dass seine Felder das nicht erhalten Gleicher Sonnenschein und gleiche Luftfeuchtigkeit. Da Sonnenschein und Luftfeuchtigkeit den Maisanbau beeinflussen, gibt es zwei systematische Fehlerquellen. Der Bauer ordnet jedem kleinen Stück Land eine bestimmte Maissorte zu, so dass jede einzelne in jeder Zeile und Spalte einmal vorkommt ähm:

A: Körnung 1

B: Körnung 2

C: Körnung 3

D: Körnung 4 ·

Die folgende Abbildung zeigt das lateinische 4 × 4-Quadrat mit den römischen Buchstaben A, B, C und *D* steht für die vier oben genannten Behandlungen:

A B C D
BCDA
CDAB
DABC

Eine **Varianzanalyse** zeigt an, ob nach Streichung der Linien (Sonne) und der Spalte (Feuchtigkeit) ein signifikanter Unterschied zwischen den Behandlungen besteht.

WEITERLESEN

Varianzanalyse
Versuchsplanung

REFERENZEN

Fisher, RA: Statistische Methoden für Forschungsmitarbeiter. Oliver & Boyd, Edinburgh (1925)

Preece, DA: Lateinische Quadrate, lateinische Würfel, lateinische Rechtecke usw. In: Kotz, S., John son, NL (Hrsg.) Encyclopedia of Statis tical Sciences, vol. 4. Wiley, New York (1983)

Gesetz der großen Zahlen

Das Gesetz der großen Zahlen besagt, dass der **Mittelwert** einer Summe unabhängiger und identisch verteilter **Zufallsvariablen** gegen den **Erwartungswert** ihrer Verteilung konvergiert, wenn die **Stichprobengröße** gegen Unendlich geht.

Machine Translated by Google

GESCHICHTE

Das Gesetz der großen Zahlen wurde erstmals von **J. Bernoulli** in seinem 1713 veröffentlichten Werk *Ars Conjectandi* im Rahmen einer empirischen Definition der **Wahrscheinlichkeit aufgestellt.** Er stellte fest, dass die relative **Häufigkeit** eines **Ereignisses** bei der Wiederholung identischer Aufgaben **(Bernoulli-Verteilung)** gegen eine Zahl konvergiert, die in seiner Wahrscheinlichkeit besteht.
Die Bienaymé-Tchebychev-Ungleichung wurde zunächst von Bienaymé, J. (1853) entdeckt, bevor Tchebychev, PL (1867) einige Jahre später völlig unabhängig davon entdeckte.

Die allgemeine Version des Gesetzes der großen Zahlen wird der russischen Mathematik zugeschrieben Cian Khintchin, A. (1894–1959).

MATHEMATISCHE ASPEKTE

Stellen wir zunächst die Ungleichung von Bienaymé–Tchebychev vor:
Betrachten Sie X als **Zufallsvariable mit Erwartungswert ÿ und Varianz** ÿ2, beide endlich. Für jedes reelle ÿ > 0 gilt:

$$P\ (|X\ ÿ\ ÿ| ÿ\ ÿ)\ ÿ\ ÿ2 \quad \frac{ÿ2}{\quad},$$

Das bedeutet, dass die Wahrscheinlichkeit , dass *eine* **Zufallsvariable**

Die Bedeutung dieser Ungleichheit liegt darin, dass sie es ermöglicht, den **Wert** von zu begrenzen bestimmte **Wahrscheinlichkeiten** in Fällen, in denen nur der **Erwartungswert** der Verteilung bekannt ist, letztendlich ihre **Varianz.** Es ist offensichtlich, dass, wenn die Verteilung selbst bekannt ist, diese Grenzwerte nicht verwendet werden müssen, da der genaue Wert vorliegt dieser Wahrscheinlichkeiten können berechnet werden.
Mit dieser Ungleichung ist es möglich, die zu berechnen Konvergenz des beobachteten **Mittelwerts** einer **Zufallsvariablen** gegenüber ihrem **erwarteten Wert.** Betrachten Sie *n* unabhängige und identisch verteilte **Zufallsvariablen** X1, *X2,...,Xn* mit **Erwartungswert** ÿ und **Varianz** ÿ2; dann gilt für jedes ÿ > 0:

$$P \quad \frac{X1 + X2 + ... + Xn}{N} \quad ÿÿ >ÿ ÿÿnÿÿ0.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Gesetz der großen Zahlen unterstreicht die Tatsache, dass eine , wenn auch sehr geringe, **Wahrscheinlichkeit** einer erheblichen Abweichung zwischen dem empirischen **Mittel** einer Reihe von **Ereignissen** und ihrem **erwarteten Wert bestehen bleibt.** Aus diesem Grund wird der Begriff *schwaches Gesetz der großen Zahlen* verwendet. Tatsächlich lässt sich zeigen, dass dies nicht der Fall ist und dass die Konvergenz qua gewiss ist. Dieses Ergebnis, das stärker ist als das vorherige, wird als *starkes Gesetz der großen Zahlen* bezeichnet und lautet:

$$P \lim_{nÿÿ} \quad X^{-} = ÿ = 1 \quad .$$

BEISPIELE

Das Gesetz der großen Zahlen besagt, dass es immer möglich ist, einen **Wert** für *n* zu finden, sodass **die Wahrscheinlichkeit ,** dass X^{-} in einem **Intervall** ÿ ± ÿ enthalten ist, so groß wie gewünscht ist.
Geben wir ein Beispiel einer **Wahrscheinlichkeitsverteilung** mit **der Varianz** ÿ2 = 1. Durch die Wahl eines **Intervalls** ÿ = *0,5* und einer **Wahrscheinlichkeit** von 0,05 können wir schreiben:

$$P\ X^{-}\ ÿ\ ÿ\ ÿ\ 0{,}5 = 0{,}05\ .$$

Anhand der Ungleichung von Tschebytschow wissen wir das

$$P\ X^{-}\ ÿ\ ÿ\ ÿ\ ÿ\ ÿ\ nÿ2 \quad \frac{ÿ2}{\quad}\ .$$

Wir können daher die folgende Ungleichung feststellen:

$$0{,}05\ ÿ\ \frac{1}{n \cdot 0{,}25},$$

Machine Translated by Google

aus denen:

$$0{,}05 \cdot 0{,}25 \ddot{y} 1/n ,$$

$$n \ddot{y} 80 .$$

WEITERLESEN

Zentraler Grenzwertsatz

Konvergenzsatz

(Théorème de convergence)

VERWEISE

Bernoulli, J.: ArsConjectandi, Opus Posthu mum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Basel (1713)

Bienaymé, IJ: Überlegungen zur Entdeckung der Wahrscheinlichkeitsrechnung von Laplace anhand der Methode der geringsten Erfolge. Comptes Rendus de l'Académie des Sciences, Paris **37,** 5–13 (1853); veröffentlicht im Jahr 1867 im Journal de Liouville vor dem Preuve des Inégalités de Bienaymé-Chebyshev, Journal de Math ématiques Pures et Appliquées **12,** 158– 176

Tchebychev, PL: Des valeurs moyennes. J. Mathe. Pures Appl. 12(2), 177–184 (1867). Gleichzeitig in Russland veröffentlicht, das Mat. Sbornik 2(2), 1–9

Am wenigsten absolut Abweichungsregression

Die Methode der kleinsten absoluten Abweichungen (die LAD-Methode) ist eine der wichtigsten Alternativen zu den Methoden der kleinsten Quadrate, wenn man Regressionsparameter schätzen möchte.
Das Ziel der LAD-Regression besteht darin, einen **robusten Schätzer bereitzustellen.**

GESCHICHTE

Die Regression der kleinsten absoluten Abweichung wurde etwa 50 Jahre vor der **Methode der kleinsten Quadrate** im Jahr 1757 von **Roger Joseph Boscovich** eingeführt . Er verwendete dieses Verfahren, als er versuchte, inkohärente Maße, die zur Schätzung der Form der Erde verwendet wurden, in Einklang zu bringen. **Pierre Simon de Laplace** übernahm diese Methode 30 Jahre später, doch sie blieb im Schatten der wenigsten

QuadrateMethode entwickelt **von AdrienMarie Legendre** und **Carl Friedrich Gauss.** Die einfache Berechnung der Methode der kleinsten Quadrate machte die Methode der kleinsten Quadrate viel beliebter als die LAD-Methode. Doch in den letzten Jahren und mit Fortschritten in der statistischen Berechnung kann die LAD-Methode problemlos verwendet werden.

MATHEMATISCHE ASPEKTE

Wir behandeln hier nur die einfache LAD-Regression Fälle.

Betrachten Sie das einfache **Regressionsmodell:** *yi* = ÿ0 + *ÿ1xi* + *ÿi.* Die LAD-Methode implementiert das folgende Kriterium: Wählen Sie ÿ0 und ÿ 1 so, dass sie die Summe der Restabsolutwerte minimieren, d. h.

$$|ei| = yi \ddot{y} \ddot{y} 0 \ddot{y} \ddot{y} 1xi ,$$

wobei *ei* das i-te Residuum bezeichnet . Die Besonderheit dieser Methode besteht darin, dass es keine explizite Formel zur Berechnung von ÿ 0 und ÿ 1 gibt. LAD-Schätzungen können durch Anwendung eines iterativen **Algorithmus berechnet werden.**

Einfacher LAD-Algorithmus

Der wesentliche Punkt, den der LAD-Algorithmus nutzt, ist die Tatsache, dass die Regressionsgerade immer mindestens zwei Datenpunkte kreuzt.
Sei (x1, y1) unser gewählter Datenpunkt. Wir suchen die beste Linie, die durch diesen Punkt verläuft. Diese Linie kreuzt mindestens einen anderen Punkt (x2, y2).

Machine Translated by Google

Der nächste Schritt besteht darin, die beste Zeile mit zu suchen bezüglich der Summe $|ei|$, die diese kreuzt Zeit (x2, y2). Zumindest jetzt kreuzt sich diese Grenze ein weiterer Punkt,(x3, y3). Wir machen iterativ weiter und suchen nach der besten Linie, die wir kreuzen (x3, y3) und so weiter. Während der Iterationen die Menge|ei| nimmt ab. Wenn für eine Iteration die erhaltene Linie mit der vorherigen identisch ist, können wir daraus schließen, dass dies der Fall ist beste Zeile.

Bau des besten Linienübergangs (in Bezug auf $X_{1, j\ 1)}$ $|ei|$

Da diese Linie einen anderen Datenpunkt kreuzt, haben wir sollte den Punkt *(xk, yk)* finden , für den die Linie

$$y\ (x) = y1 + xk\ \frac{yk\ \ddot{y}\ y1}{\ddot{y}\ x1}\ (x\ \ddot{y}\ x1)\ ,$$

mit Steigung

$$\ddot{y}\ 1 = \frac{yk\ \ddot{y}\ y1}{xk\ \ddot{y}\ x1}$$

und abfangen

$$\ddot{y}\ 0 = y1\ \ddot{y}\ \ddot{y}\ 1x1\ ,$$

ist die beste in Bezug auf $|ei|$. Finden

An diesem Punkt benennen wir den $(n\ \ddot{y}\ 1)$-Kandidaten um Punkte (x2, *y2),...,(xn, yn)* auf diese Weise Das

$$\frac{y2\ \ddot{y}\ y1}{x2\ \ddot{y}\ x1}\ \ddot{y}\ \frac{y3\ \ddot{y}\ y1}{x3\ \ddot{y}\ x1}\ \ddot{y}\ \ldots\ \ddot{y}\ \frac{yn\ \ddot{y}\ y1}{xn\ \ddot{y}\ x1}\ .$$

Wir definieren $T = \sum_{i=1}^{N} |xi\ \ddot{y}\ x1|$. Die Gesuchten Punkt(xk, yk) wird nun durch den Index bestimmt *k* für welche

$$\ddot{y}\ |x2\ \ddot{y}\ x1| + \ldots + |xk\ddot{y}1\ \ddot{y}\ x1| < \frac{T}{2}$$

$$\text{„}\ |x2\ \ddot{y}\ x1| + \ldots$$

$$\text{„}\ + |xk\ddot{y}1\ \ddot{y}\ x1| + |xk\ \ddot{y}\ x1| > \frac{T}{2}\ .$$

Diese Bedingung garantiert, dass ÿ1 minimiert wird Die Quantität

$$\sum_{i=1}^{N} (yi\ \ddot{y}\ y1)\ \ddot{y}\ \ddot{y}\ 1\ (xi\ \ddot{y}\ x1)$$

analog zu |ei|für die Regressionsgeraden Durchgang durch (x1, y1). Das ÿ 0 wird berechnet so, dass sich die Regressionsgerade kreuzt (x1, y1). Wir können auch überprüfen, ob es erfolgreich ist durch den Datenpunkt *(xk,* yk). Wir haben es einfach getan den Punkt *(xk, yk)* in (x2, y2) umzubenennen und Neustart.

DOMÄNEN UND EINSCHRÄNKUNGEN

Zwei besondere Fälle können auftreten, wenn einer führt die LAD-Regression aus: das **Nichteindeutige Lösung** und die **Entartung.** Nicht einzigartig Lösung bedeutet, dass mindestens zwei Zeilen vorhanden sind Minimierung der Summe der verbleibenden Absolutwerte. Entartung bedeutet, dass die LAD Die Regressionslinie verläuft durch mehr als zwei Datenpunkte.

BEISPIELE

Nehmen wir ein Beispiel mit 14 North undMittelamerikanische Länder.Die folgenden Die Tabelle enthält die Daten für jedes Land. die Geburtenrate (Anzahl der Geburten pro 1000 Einwohner) zusammen mit der Urbanisierungsrate (Prozentsatz der Bevölkerung, die in Städten mit lebt). mehr als 100.000 Einwohner) im Jahr 1980.

Tabelle: Urbanisierungsrate und Geburtenrate

Obs.	Land	Urbanizationsrate xi	Geburt Rate yi
1	Kanada	55,0	16.2
2	Costa Rica	27.3	30.5
3	Kuba	33.3	16.9
4	USA	56,5	16.0
5	El Salvador	11.5	40.2

Machine Translated by Google

Obs.	Land	Urbanisierungsrate xi	Geburtenrate yi
6	Guatemala	14.2	38.4
7	Haiti	13.9	41.3
8	Honduras	19.0	43.9
9	Jamaika	33.1	28.3
10	Mexiko	43.2	33.9
11	Nicaragua	28.5	44.2
12	Trinidad/ Tobago	6.8	24.6
13	Panama	37,7	28.0
14	dominikanische Republik	37.1	33.1

Quelle: Birkes und Dodge (1993)

Wir versuchen, die Geburtenrate *yi* als Funktion der Urbanisierungsrate *xi* mithilfe von LAD zu schätzen Regression, die darin besteht, die LAD zu konstruieren Regressionsmodell

$$yi = \ddot{y}0 + \ddot{y}1xi + \ddot{y}i .$$

Verwendung des LAD-Algorithmus zur Schätzung von ÿ0 und ÿ1, im ersten Schritt müssen wir nach dem suchen beste Linie, die mindestens einen Punkt kreuzt, zum Beispiel Kanada (zufällig ausgewählt). Wir haben (x1, y1) = *(55,0,* 16,2). Wir berechnen die Pisten für alle 13 übrigen Länder.

$\frac{yi\ddot{y}16,2}{xi\ddot{y}55,0}$

Beispielsweise gilt für Costa Rica das erhaltene Ergebnis ist $\frac{30,5\ddot{y}16,2}{27,3\ddot{y}55,0} = 0,5162$. Der 13. Bezirk

Versuche werden jetzt in aufsteigender Reihenfolge neu angeordnet entsprechend den berechneten Steigungen, wie möglich siehe untenstehende Tabelle. Der Punkt (x2, y2) ist Mexiko, der Punkt (x3, y3) ist Nicaragua, und so weiter. Im Folgenden berechnen wir die Menge $T = |xi \ddot{y} 55,0| = 355,9$ und $\frac{T}{2} = 177,95$. Das hat mit Definieren zu tun das Land *k* , das die Bedingung erfüllt

$$\ddot{y} |x2 \ddot{y} x1| + ... + |xk\ddot{y}1 \ddot{y} x1| < \frac{T}{2},$$

$$|x2 \ddot{y} x1| + ... + |xk\ddot{y}1 \ddot{y} x1| + |xk \ddot{y} x1| > \frac{T}{2}.$$

Seit

$$\sum_{j=2}^{8} xj \ddot{y} x1 = 172,5 < \frac{T}{2}$$

Und

$$\sum_{j=2}^{9} xj \ddot{y} x1 = 216,0 > \frac{T}{2},$$

Wir können daraus schließen, dass $k = 9$ und dass die Die beste Regressionslinie verläuft durch Kanada (x1, y1) und El Salvador (x9, y9). Somit:

$$\ddot{y}\,1 = \frac{y9 \ddot{y} y1}{x9 \ddot{y} x1} = \frac{40,2 \ddot{y} 16,2}{11,5 \ddot{y} 55,0} = \ddot{y}0,552 ,$$

$$\ddot{y}\,0 = y1 \ddot{y} \ddot{y} 1x1 = 16,2 \ddot{y} (\ddot{y}0,552) \cdot 55,0$$
$$= 46,54 ,$$

und die LAD-Regressionslinie ist:

$$y(x) = 46,54 \ddot{y} 0,552x .$$

Tabelle: Berechnung zur Ermittlung der LAD-Regressionslinie

Obs.	Land	$\frac{yi}{xi}$ $\frac{j\ 1}{x_1}$	$\lvert xi \cdot x_1 \rvert$	$\sum_{j=2} xj \cdot x_1$
2	Mexiko	ÿ1,5000 11,8		11.8
3	Nicaragua	ÿ1,0566 26,5		38.3
4	Domin. Rep.	ÿ0,9441 17,9		56.2
5	Honduras	ÿ0,7694 36,0		92,2
6	Panama	ÿ0,6821 17,3		109,5
7	Haiti	ÿ0,6107 41,1		150,6
8	Jamaika	ÿ0,5525 21,9		172,5
9	El Salvador	ÿ0,5517 43,5		216,0
10	Guatemala	ÿ0,5447 40,8		256,8
11	Costa Rica	ÿ0,5162 27,7		284,5
12	Trinidad/ Tobago	ÿ0,1743 48,2		332,7
13	USA	ÿ0,1333	1.5	334,2
14	Kuba	ÿ0,0323 21,7		355,9

Der nächste Schritt besteht darin, unsere erste Schätzung zu verbessern besteht darin, die beste Linie zu finden, die El kreuzt

L

Machine Translated by Google

Salvador. Wir benennen diesen Punkt um (x1, y1) = *(11.5, 40.2)* und ordnen die 13 Länder in aufsteigender Reihenfolge entsprechend ihrer Steigung yiÿ11.5 neu xiÿ40.2 , und auf ähnliche Weise erhalten wir die beste Regressionsgerade, die sowohl durch El Sal Vador als auch durch die Vereinigten Staaten verläuft. Somit,

$$1 = = \ddot{y} \frac{40{,}2 \ddot{y}\ 16{,}0}{0{,}538 \ddot{y}\ 11{,}5 \ddot{y}\ 56{,}5} \quad ,$$

$$\ddot{y}\ 0 = 40{,}2 \ddot{y}\ (\ddot{y}0{,}538) \cdot 11{,}5 = 46{,}38 \quad ,$$

und die Regressionsgerade ist

$$y(x) = 46{,}38 \ddot{y}\ 0{,}538x \ .$$

Der dritte Schritt besteht darin, die beste Linie durch die Vereinigten Staaten zu finden. Da wir (auf ähnliche Weise wie oben) feststellen, dass diese Regressionsgerade durch El Salvador verläuft, stoppt unser Algorithmus und die Regressionsgerade *y(x) = 46,38* ÿ 0,538x ist die LAD-Regressionsgerade unseres Problems.

Wir beachten hier, dass die mit der Methode der kleinsten Quadrate für diese Daten berechnete Linie durch gegeben ist

$$y(x) = 42{,}99 \ddot{y}\ 0{,}399x \ .$$

Beachten Sie, dass Kuba, Nicaragua und Trinidad/Tobago mit dieser Regressionslinie ungewöhnlich große Restwerte aufweisen. Indem wir diese drei Punkte löschen und die Regressionsgerade der kleinsten Quadrate neu berechnen, erhalten wir:

$$y\ (x) = 48{,}52 \ddot{y}\ 0{,}528x \quad ,$$

was sehr nahe an der LAD-Regressionslinie liegt.

WEITERLESEN

- **Varianzanalyse**
- **Kleinsten Quadrate**
- **Multiple lineare Regression**
- **Regressionsanalyse**
- **Restwert**
- **Robuste Schätzung**
- **Einfache lineare Regression**

VERWEISE

Birkes, D., Dodge, Y.: Alternative Methoden der Regression. Wiley, New York (1993)

Am wenigsten bedeutsam

Differenztest

Der LSD-Test (Least Significant Difference) wird im Zusammenhang mit der **Varianzanalyse verwendet,** wenn das F-Verhältnis die Ablehnung der **Nullhypothese H0 nahelegt,** d. h. wenn die Differenz zwischen den Mittelwerten der Grundgesamtheit signifikant ist.

Dieser Test hilft dabei, die Populationen zu identifizieren, deren Mittelwerte statistisch unterschiedlich sind. Die Grundidee des Tests besteht darin, die Populationen paarweise zu vergleichen. Anschließend wird damit eine **ein- oder zweiseitige Varianzanalyse** durchgeführt , sofern die Nullhypothese bereits verworfen wurde.

GESCHICHTE

Der LSD-Test wurde von **Fisher, Ronald Aylmer** (1935) entwickelt , der wissen wollte, welche Behandlungen einen signifikanten Effekt in einer **Varianzanalyse hatten.**

MATHEMATISCHE ASPEKTE

Wenn in einer **Varianzanalyse** das F-Verhältnis zur Ablehnung der Nullhypothese H0 führt, können wir den LSD-Test durchführen, um herauszufinden, welche Mittel zur Ablehnung von H0 geführt haben . Der Test besteht in einem paarweisen Vergleich der Mittelwerte. Im Allgemeinen ist die **Standardabweichung** der Differenz zwischen

der Mittelwert der Gruppe *i* und der Mittelwert der Gruppe j,

Machine Translated by Google

für *i* = j ist gleich:

$$s2_{ICH} \quad \frac{1}{ni} + \frac{1}{nj} \quad ,$$

wobei $s2_{ICH}$ die **Schätzung** der **Varianz** ist innerhalb der Gruppen

$$2\,s_{ICH} = \frac{SCI}{N \ddot{y}\, k} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{ni} Yij \ddot{y}\, \bar{Y}i.^{2}}{N \ddot{y}\, k} \quad ,$$

wobei *SCI* die Summe der Quadrate innerhalb der Gruppen ist, Nis die Gesamtzahl der Beobachtungen, *k* die Anzahl der Gruppen ist, *Yij* die j-te Beobachtung der Gruppe i ist , $\bar{Y}i.$ ist der Mittelwert der Gruppe i, *ni* ist die Anzahl der Beobachtungen in Gruppe i und *nj* ist die Anzahl der Beobachtungen in Gruppe j.

Das Verhältnis

$$\frac{\bar{Y}i. \ddot{y}\, \bar{Y}j.}{s2_{ICH} \quad \frac{1}{ni} + \frac{1}{nj}}$$

folgt der **Student-** Verteilung mit $N \ddot{y}\, k$ Freiheitsgrade. Der Unterschied zwischen einem Mittelwertpaar ist signifikant, wenn

$$\ddot{y}\, \bar{Y} \quad J. \ddot{y}\, s2\, \bar{Y}i._{ICH} \quad \frac{1}{ni} + \frac{1}{nj} \quad \cdot tN\ddot{y}k,1\ddot{y}\ddot{y}_{\bar{2}} \quad ,$$

wobei $tN\ddot{y}k,1\ddot{y}\ddot{y}_{\bar{2}}$ den **Wert** einer **Student-** Variablen mit $N \ddot{y}\, k$ Freiheitsgraden für ein auf ÿ festgelegtes **Signifikanzniveau** bezeichnet .

Die Größe $s2\ I(\ \frac{1}{ni} + \frac{1}{nj}) \cdot tN\ddot{y}k,1\ddot{y}\ddot{y}_{\bar{2}}$ ist namens LSD.

DOMÄNEN UND EINSCHRÄNKUNGEN

Wie die **Varianzanalyse** erfordert der LSD-Test unabhängige und identisch verteilte Normalvariablen mit konstanter Varianz

ance.

BEISPIELE

Vollbutter-Croissants nehmen beim Kochen unterschiedlich viel Fett auf. Wir wollen überprüfen, ob die aufgenommene Menge von der Art des Fettgehalts (tierisches oder pflanzliches Fett) abhängt. Für unser Experiment bereiten wir vier verschiedene Fettsorten vor und kochen für jede Croissant-Sorte sechs Vollbutter-Croissants. Hier sind die Ergebnisse unseres Experiments:

Fettart			
1	2	3	4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85	63	70
95	77	76	68

Wir führen eine **einseitige Varianzanalyse** durch , die es uns ermöglicht, die **Nullhypothese** zu testen

$$H0 : \ddot{y}1 = \ddot{y}2 = \ddot{y}3 = \ddot{y}4 \ ,$$

was offensichtlich abgelehnt wird.

Da H0 abgelehnt wird, führen wir den LSD-Test durch, um herauszufinden, welche Mittel die Ablehnung von H0 verursacht haben.

In diesem Beispiel ist die Anzahl der Beobachtungen für jede Gruppe gleich *(n1* = n2 = n3 = n4 = *n* = 6). Folglich $s2_{ICH} \quad \frac{1}{ni} + \frac{1}{nj}$ wird vereinfacht zu $\frac{2s2}{N}$.

Der Wert für den entsprechenden LSD-Test ist gleich:

$$\frac{2s2}{N} \cdot tN\ddot{y}k,1\ddot{y}\ddot{y}_{\bar{2}} \quad .$$

Die Ergebnisse der **Varianzanalyse** liefern die dafür erforderlichen Rechenelemente

L

Machine Translated by Google

Finden Sie den Wert von s^2_{es}:

$$s^2_{es} = \frac{SCI}{N - k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2}{N - k} = \frac{2018}{20} = 100{,}9 .$$

Wenn wir ein **Signifikanzniveau** $\alpha = 0{,}05$ wählen, ist der **Student-** Wert für $N - k = 20$ Freiheitsgrade gleich:

$$t_{N-k,1-\alpha/2} = t_{20,0{,}975} = 2{,}086 .$$

Daher ist der Wert für die LSD-Statistik:

$$LSD = \sqrt{\frac{2s^2_{es}}{N}} \cdot t_{20,0{,}975} = \sqrt{\frac{2 \cdot 100{,}9}{6}} \cdot 2{,}086 = 12{,}0976 .$$

Der Unterschied zwischen einem Mittelwertpaar $\bar{Y}_i$ und $\bar{Y}_j$ ist signifikant, *wenn*

$$|\bar{Y}_{i.} - \bar{Y}_{j.}| > 12{,}0976 .$$

Wir können die Ergebnisse in einer Tabelle zusammenfassen

Erfassen der Mittelwertdifferenzen, der Werte für die LSD-Statistik und der Schlussfolgerung.

Unterschied	LSD Signifikant
$\bar{Y}_1 - \bar{Y}_2$ = |72 − 85| = 13	12,0976 Ja
$\bar{Y}_1 - \bar{Y}_3$ = |72 − 76| = 4	12,0976 Nr.
$\bar{Y}_4 - \bar{Y}_1$ = |72 − 62| = 10	12.0976 Nr
$\bar{Y}_3 - \bar{Y}_2$ = |85 − 76| = 9	12,0976 Nr.
$\bar{Y}_2 - \bar{Y}_4$ = |85 − 62| = 23	12,0976 Ja
$\bar{Y}_3 - \bar{Y}_4$ = |76 − 62| = 14	12,0976 Ja

Wir verifizieren, dass nur die Differenzen $\bar{Y}_1 - \bar{Y}_2$., $\bar{Y}_2 - \bar{Y}_4$. und $\bar{Y}_3 - \bar{Y}_4$. sind bedeutsam.

WEITERLESEN

- Varianzanalyse
- Studentenverteilung
- Studententisch

VERWEISE

Dodge, Y., Thomas, DR: Zur Leistung nichtparametrischer und normaltheoretischer Mehrfachvergleichsverfahren. Sankhya **B42,** 11–27 (1980)

Fisher, RA: Das Design von Experimenten. Oliver & Boyd, Edinburgh (1935)

Miller, RG, Jr.: Simultaneous Statistical Inference, 2. Aufl. Springer, Berlin Heidelberg New York (1981)

Kleinsten Quadrate

Siehe **Methode der kleinsten Quadrate.**

Methode der kleinsten Quadrate

Die Methode der kleinsten Quadrate besteht darin, die Summe der quadrierten Residuen zu minimieren. Letztere entsprechen den quadrierten Abweichungen zwischen geschätzten und beobachteten Werten.

GESCHICHTE

Die Methode der kleinsten Quadrate wurde von **Legendre, Adrien-Marie** (1805) eingeführt.

MATHEMATISCHE ASPEKTE

Betrachten Sie die **multiple lineare Regression Modell:**

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j \cdot X_{ji} + \varepsilon_i, \qquad i = 1,\ldots,n,$$

Dabei ist Y_i die **abhängige Variable,** $X_{ji}, j = 1,\ldots,p - 1$ sind die erklärenden oder unabhängigen Variablen, $\beta_j, j = 0,\ldots,p - 1$ sind

Machine Translated by Google

die zu schätzenden Parameter und *ÿi* ist der **Fehlerbegriff .**

Die Parameterschätzung mit der Methode der kleinsten Quadrate besteht darin, die Summe der quadrierten Fehler (Residuen) zu minimieren:

$$\text{Minimum } f\,ÿ0,\ldots,\ ÿpÿ1\ ,$$

Wo:

$$f\,ÿ0,\ldots,\ ÿpÿ1 = \sum_{i=1}^{N} ÿ2_{ich.}$$

Die Restquadratsumme ist gleich:

$$f\,ÿ0,\ldots,\ ÿpÿ1 = \sum_{i=1}^{N} ÿ2$$

$$= \sum_{i=1}^{N} ÿ \left(ÿYi ÿ\ ÿ0\ ÿ - \sum_{j=1}^{pÿ1} ÿj \cdot Xji \right)^2 ÿ .$$

Wir finden die geschätzten Parameter der Funktion, indem wir die partiellen Ableitungen dieser Funktion in Bezug auf jeden Parameter mit Null gleichsetzen eter:

$$\frac{ÿf\ ÿ0,\ldots,\ ÿpÿ1\ ÿ0}{ÿf} = 0\ ,$$

$$\frac{ÿ0,\ldots,\ ÿpÿ1\ ÿ1}{} = 0\ ,$$

$$\vdots$$

$$\frac{ÿf\ ÿ0,\ldots,\ ÿpÿ1}{ÿpÿ1} = 0\ .$$

Angenommen, wir haben *n* Beobachtungen und $p = 2$:

$$(X1, Y1),(X2, Y2),\ldots,(Xn, Yn)\ .$$

Die Gleichung bezüglich Yi und Xi kann anhand des linearen Modells definiert werden:

$$Yi = ÿ0 + ÿ1 \cdot Xi + ÿi\ , \qquad i = 1,\ldots,n.$$

Die Summe der quadrierten Abweichungen von der geschätzten Regressionsgeraden beträgt:

$$f\,(ÿ0,\ ÿ1) = \sum_{i=1}^{N} ÿ2$$

$$= \sum_{i=1}^{N} 2\ (yi\ ÿ\ ÿ0\ ÿ\ ÿ1 \cdot Xi) \quad .$$

Mathematisch bestimmen wir ÿ0 und ÿ1 , indem wir die partiellen Ableitungen der Funktion *f* nach den Parametern ÿ0 und ÿ1 bilden und sie gleich Null setzen.
Partielle Ableitung nach ÿ0:

$$\frac{ÿf}{} = ÿ2 \cdot ÿÿ0 \sum_{i=1}^{N} (Yi\ ÿ\ ÿ0\ ÿ\ ÿ1 \cdot Xi)\ .$$

Partielle Ableitung nach ÿ1:

$$\frac{ÿf}{ÿÿ1} = ÿ2 \cdot \sum_{i=1}^{N} Xi \cdot (yi\ ÿ\ ÿ0\ ÿ\ ÿ1 \cdot Xi)\ .$$

Daher werden die geschätzten **Werte** für ÿ0 und ÿ1, bezeichnet als ÿˆ 0 und ÿˆ 1, als Lösung für die folgenden Gleichungen angegeben:

$$\sum_{i=1}^{N} Yi\ ÿ\ ÿˆ\ 0\ ÿ\ ÿˆ\ 1 \cdot Xi = 0 \quad ,$$

$$\sum_{i=1}^{N} Xi \cdot Yi\ ÿ\ ÿˆ\ 0\ ÿ\ ÿˆ\ 1 \cdot Xi = 0\ .$$

Wenn wir diese beiden Gleichungen entwickeln, erhalten wir:

$$\sum_{i=1}^{N} Yi = n \cdot ÿˆ\ 0 + ÿˆ\ 1 \cdot \sum_{i=1}^{N} Xi$$

Und

$$\sum_{i=1}^{N} (Xi \cdot Yi) = ÿˆ\ 0 \cdot \sum_{i=1}^{N} Xi + ÿˆ\ 1 \cdot \sum_{i=1}^{N} X2_{ich,}$$

die wir die **Normalgleichungen** der Regressionsgeraden der kleinsten Quadrate nennen .

L

Machine Translated by Google

Gegeben

$$X^- = \frac{\sum_{i=1}^{N} Xi}{N} \quad \text{und} \quad Y^- = \frac{\sum_{i=1}^{N} Yi}{N},$$

Die Lösung der Normalengleichungen liefert die Schätzungen der folgenden Parameter:

$$\hat{y}0 = Y^- - \hat{y}1 \cdot X,$$

$$\hat{y}1 = \frac{n \cdot \sum_{i=1}^{n}(Xi \cdot Yi) - \sum_{i=1}^{n} Xi \cdot \sum_{i=1}^{n} Yi}{n \cdot \sum_{i=1}^{N} X2 - \left(\sum_{i=1}^{N} Xi\right)^2} = \frac{\sum_{i=1}^{n} Xi Y^- - Yi Y^-}{\sum_{i=1}^{N} Xi X^-}.$$

Die Kleinste-Quadrate-Schätzungen für die abhängige Variable sind:

$$\hat{Y}i = \hat{y}0 + \hat{y}1 \cdot Xi, \quad i = 1,\ldots,n$$

oder

$$\hat{Y}i = Y^- + \hat{y}1 \cdot Xi - X^-, \quad i = 1,\ldots,n.$$

Gemäß dieser Gleichung verläuft die Regressionsgerade der kleinsten Quadrate durch den Punkt X^-, Y^-. Dies wird als Schwerpunkt oder Schwerpunkt für die Streuwolke der Datenpunkte bezeichnet.

Wir können das multiple lineare Regressionsmodell auch in Form von Vektoren und Matrizen ausdrücken:

$$\mathbf{Y} = \mathbf{X} \cdot \hat{\mathbf{y}} + \hat{\mathbf{y}},$$

Dabei ist **Y** der $(n \times 1)$-Antwortvektor **(die** abhängige Variable), **X** die $(n \times p)$ **-Matrix** der unabhängigen Variablen , **ŷ** der $(n \times 1)$-Vektor des Fehlerterms und **ŷ** der $(p \times 1)$ Vektor der zu schätzenden Parameter.

Die Anwendung der Methode der kleinsten Quadrate besteht in der Minimierung der folgenden Gleichung tion:

$$\begin{aligned}\hat{\mathbf{y}} \cdot \hat{\mathbf{y}} &= (\mathbf{Y} - \hat{\mathbf{y}}\,\mathbf{X} \cdot \hat{\mathbf{y}}) \cdot (\mathbf{Y} - \hat{\mathbf{y}}\,\mathbf{X} \cdot \hat{\mathbf{y}}) \\ &= \mathbf{Y} \cdot \mathbf{Y} - \hat{\mathbf{y}} \cdot \mathbf{X} \cdot \mathbf{Y} - \hat{\mathbf{y}}\,\mathbf{Y} \cdot \mathbf{X} \cdot \hat{\mathbf{y}} \\ &\quad + \hat{\mathbf{y}} \cdot \mathbf{X} \cdot \mathbf{X} \cdot \hat{\mathbf{y}} - 2 \cdot \hat{\mathbf{y}} \cdot \mathbf{X} \cdot \mathbf{Y} + \hat{\mathbf{y}} \cdot \mathbf{X} \cdot \mathbf{X} \cdot \hat{\mathbf{y}}.\end{aligned}$$

Die Normalgleichungen sind in Matrixform gegeben durch:

$$\mathbf{X} \cdot \mathbf{X} \cdot \hat{\mathbf{y}} = \mathbf{X} \cdot \mathbf{Y}.$$

Wenn **X** · **X** invertierbar ist, werden die Parameter **ŷ** geschätzt durch:

$$\hat{\mathbf{y}} = \mathbf{X} \cdot \mathbf{X}^{\hat{y}1} \cdot \mathbf{X} \cdot \mathbf{Y}.$$

BEISPIELE

Siehe **einfache lineare Regression.**

WEITERLESEN

Einschätzung

Multiple lineare Regression

Regressionsanalyse

Einfache lineare Regression

REFERENZEN

Legendre, AM: Neue Methoden zur Bestimmung der Umlaufbahnen von Kometen.

Courcier, Paris (1805)

Legendre, Adrien Marie

Der 1752 in Paris geborene französische Mathematiker Legendre Adrien Marie trat die Nachfolge **von de Laplace, PS,** als Professor für Mathematik an, zunächst an der Ecole Militaire und dann an der Ecc Ecole Normale von Paris. Er war interessiert Studium der Theorie und Praxis der Astronomie und Geodäsie. Ab 1792 war er Teil der

Machine Translated by Google

Französische Kommission, die mit der Messung der Länge eines Meridianquadranten vom Nordpol bis zum Äquator durch Paris beauftragt ist.

Er starb 1833, aber in der Geschichte der **Statistik** blieb sein Name mit der Veröffentlichung seines Werks im Jahr 1805 verbunden, das einen Anhang mit dem Titel „Über die Methode der **kleinsten Quadrate** “ enthielt. Diese Veröffentlichung war der Auslöser der Kontroverse zwischen ihm und **Gauss, CF,** der mit ihm um die Anerkennung der Entdeckung dieser Methode wetteiferte.

Hauptwerk von Legendre, AM:

1805 Neue Methoden zur Bestimmung der Umlaufbahnen von Kometen. Courcier, Paris.

VERWEISE

Plackett, RL: Studien zur Geschichte der Wahrscheinlichkeit und Statistik. In: Kendall, M., Plackett, RL (Hrsg.) Die Entdeckung der Methode der kleinsten Quadrate. Bd. II. Griffin, London (1977)

Lehmann, Erich

Lehmann, Erich wurde 1917 in Straßburg, Frankreich, geboren und ging 1940 in die Vereinigten Staaten. Er hat Grundlegendes gemacht

Beiträge zur Theorie der Statistik. Seine Forschungsinteressen sind die statistische Theorie sowie die Geschichte und Philosophie der Statistik. Er ist Autor von *Basic Concepts of Probability and Statistics, Elements of Finite Proba bility, Nonparametrics: Statistical Methods Based on Ranks, Testing Statistical Hypotheses* und *Theory of Point Estima tion.* Er ist Mitglied der American Acade

Er ist Mitglied des Master of Arts and Sciences und der National Academy of Science (NAS).

wurde 1978 in die NAS gewählt. Er wurde Herausgeber der Annals of Mathematical Statis tics und Präsident des Institute of Mathe matisch Statistik.

Ausgewählte Werke und Veröffentlichungen von Erich Lehmann:

1983 Theory of Point Estimation, 2. Aufl. Wiley, NY.

1986 Testen statistischer Hypothesen. 2. Aufl. In: Reihe zur Wahrscheinlichkeitsrechnung und mathematischen Statistik. Wiley, NY.

1993 (mit Hodges, JL)Testen statistischer Hypothesen. Chapman & Hall, NY.

WEITERLESEN

Einschätzung

Hypothesentest

Signifikanzniveau

Siehe **Signifikanzniveau.**

Hebelpunkt

In der **Regressionsanalyse** nennen wir den Hebelpunkt eine **Beobachtung** *i* , für die der **geschätzte** Antwortwert Yˆi durch den **Wert** der entsprechenden **unabhängigen Variablen** *Xi* beeinflusst wird . Der Begriff Hebelpunkt entspricht dem Begriff Hebelwirkung in der Physik. Das heißt, selbst eine kleine Änderung des Beobachtungswerts kann die Schätzung von yi ernsthaft *beeinträchtigen* .

MATHEMATISCHE ASPEKTE

Die diagonalen Elemente der **Matrix H** in der Regressionsanalyse spiegeln den Einfluss der **Beobachtung** auf die Schätzung von wider

Machine Translated by Google

die Antwort. Wir assoziieren prinzipiell die Diagonalelemente *hii* der Matrix **H** mit dem Begriff des Hebelpunkts. Da die **Spur** von MatrixH gleich der Anzahl der zu schätzenden Parameter ist, erwarten wir, dass *hii* nahe bei *p/ n liegt.* Der *hii* sollte nahe bei p/n liegen , wenn alle Beobachtungen den gleichen Einfluss haben zu den Schätzungen. Wenn für eine Beobachtung *ich* wir Erhalten Sie *hii* größer als *p/ n,* dann wird die Beobachtung *i* als Hebelpunkt betrachtet. In der Praxis wird eine Beobachtung *i* als Punkt mit hoher Hebelwirkung bezeichnet, wenn:

$$hii > \frac{2p}{N} .$$

Betrachten Sie das folgende lineare Regressionsmodell in Matrixnotation:

$$\mathbf{Y} = \mathbf{X} \cdot \ddot{\mathbf{y}} + \ddot{\mathbf{y}} \quad ,$$

Dabei ist **Y** der *(n* × 1)-Antwortvektor **(die** abhängige Variable), **X** die *(n* × p) **-Matrix** der unabhängigen Variablen , **ÿ** der *(n* × 1)-Vektor des Fehlerterms und **ÿ** der *(p* ×1) Vektor der zu schätzenden Parameter.

Die **Schätzung** ÿˆ des **Vektors ÿ** ist gegeben durch:

$$\ddot{y}\hat{} = \mathbf{XX}^{\ddot{y}1}\mathbf{XY} \quad .$$

Matrix **H** ist definiert durch:

$$\mathbf{H} = \mathbf{XXX}^{\ddot{y}1}\mathbf{X} \quad .$$

Insbesondere sind die Diagonalelemente *hii* definiert durch:

$$hii = xi\,\mathbf{XX}^{\ddot{y}1}x_{ich,}$$

wobei *xi* die i-te Linie von X ist.

Den Hebelpunkt nennen wir jede Beobachtung *i,* für die gilt:

$$hii = xi\,\mathbf{XX}^{\ddot{y}1} > 2 \cdot x\,i\,\frac{p}{N} .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Huber (1981) bezeichnet Hebelpunkte als diagonales Element der **Matrix** **XXX** $^{\ddot{y}1}$ **X** ist größer als 0,2, wobei die kritische Grenze weder von der Anzahl der Parameter *p* noch von der Anzahl der Beobachtungen n abhängt.

BEISPIELE

Betrachten Sie die folgende Tabelle, in der *Y* eine **abhängige Variable** ist , die durch die **unabhängige Variable** X erklärt wird:

X	Y
50	6
52	8
55	9
75	7
57	8
58	10

Das lineare Regressionsmodell in Matrixschreibweise lautet:

$$\mathbf{Y} = \mathbf{X} \cdot \ddot{\mathbf{y}} + \ddot{\mathbf{y}} \quad ,$$

Wo

$$\mathbf{Y} = \begin{matrix} \ddot{y} \\ \\ \\ \\ \\ \ddot{y} \end{matrix} \begin{matrix} 6 \\ 8 \\ 9 \\ 7 \\ 8 \\ 10 \end{matrix} \begin{matrix} \ddot{y} \\ \\ \\ \\ \\ \ddot{y} \end{matrix} , \quad \mathbf{X} = \begin{matrix} \ddot{y} \\ \\ \\ \\ \\ \ddot{y} \end{matrix} \begin{matrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 75 \\ 1 & 57 \\ 1 & 58 \end{matrix} \begin{matrix} \ddot{y} \\ \\ \\ \\ \\ \ddot{y} \end{matrix} ,$$

ÿ ist der (6 × 1)-Residuenvektor **und** ÿ **ist** der (2 × 1)-Parametervektor.

Wir erhalten **Matrix H** durch:

$$\mathbf{H} = \mathbf{XXX}^{\ddot{y}1}\mathbf{X} \quad .$$

Wir finden:

$$\mathbf{H} = \begin{matrix} \ddot{y} \\ \\ \\ \\ \\ \\ \ddot{y} \end{matrix} \begin{matrix} 0{,}32 & 0{,}28 & 0{,}22 & \ddot{y}0{,}17 & 0{,}18 & 0{,}16 \\ 0{,}28 & 0{,}25 & 0{,}21 & \ddot{y}0{,}08 & 0{,}18 & 0{,}16 \\ 0{,}22 & 0{,}21 & 0{,}19 & 0{,}04 & 0{,}17 & 0{,}17 \\ \ddot{y}0{,}17 & \ddot{y}0{,}08 & 0{,}04 & 0{,}90 & 0{,}13 & 0{,}17 \\ 0{,}18 & 0{,}18 & 0{,}17 & 0{,}13 & 0{,}17 & 0{,}17 \\ 0{,}16 & 0{,}16 & 0{,}17 & 0{,}17 & 0{,}17 & 0{,}17 \end{matrix} \begin{matrix} \\ \ddot{y} \\ \\ \\ \\ \\ \ddot{y} \end{matrix} .$$

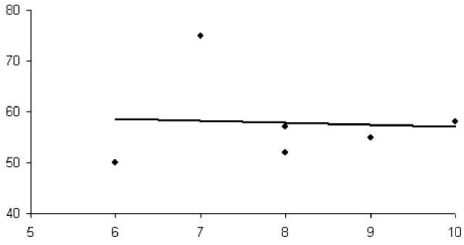
Machine Translated by Google

Wir schreiben die Diagonale der Matrix Hon auf den Vektor h:

$$\mathbf{h} = \begin{pmatrix} 0{,}32 \\ 0{,}25 \\ 0{,}19 \\ 0{,}90 \\ 0{,}17 \\ 0{,}17 \end{pmatrix}.$$

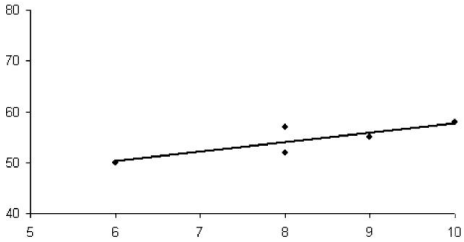
In diesem Fall gibt es zwei Parameter *(p* = 2) und sechs Beobachtungen *(n* = 6), d. h. der **kritische Wert** ist gleich $\frac{4}{6} = 0{,}67$.

Wenn wir die Komponenten von **h** mit diesem **Wert vergleichen,** stellen wir fest, dass mit Ausnahme der Beobachtung *i* = 4 alle Beobachtungen *hii* < 0,67 haben. Die vierte Beobachtung ist ein Ansatzpunkt für das untersuchte Problem. Die folgende Abbildung, in der die Variablen X *und Ya* auf den beiden Achsen dargestellt sind, veranschaulicht die Tatsache, dass der Punkt (75, 7) ein Punkt mit hoher Hebelwirkung ist.



Bei der Betrachtung dieser Abbildung stellen wir fest, dass die Regressionslinie stark von diesem Hebelpunkt beeinflusst wird und den untersuchten Datensatz nicht korrekt darstellt.

In der folgenden Abbildung haben wir diesen Punkt aus dem Datensatz gelöscht. Wir können sehen, dass die neue Regressionsgerade eine viel bessere Darstellung des Datensatzes liefert.



WEITERLESEN

- Hutmatrix
- Matrix
- Multiple lineare Regression
- Regressionsanalyse
- Einfache lineare Regression

VERWEISE

Huber, PJ: Robuste Statistik. Wiley, New York (1981)

Mosteller, F., Tukey, JW: Datenanalyse und Regression: Ein zweiter Kurs in Statistik. Addison-Wesley, Reading, MA (1977)

L

Likelihood-Ratio-Test

Der Likelihood-Ratio-Test ist ein **Hypothesentest.** Es ermöglicht das Testen allgemeiner Hypothesen bezüglich der interessierenden Parameter einer Parameterfamilie sowie das Testen zweier verschiedener Modelle, die auf denselben Daten basieren. Das Wichtigste

Die Idee dieses Tests ist folgende: Berechnen Sie die **Wahrscheinlichkeit** der Beobachtung der Daten unter der Nullhypothese H0 und unter der **Alternativhypothese** mithilfe der Likelihood-Funktion.

GESCHICHTE

Es sind **Neyman, Jerzy** und **Pearson, Egon Sharpe** (1928), der die Idee hatte, die Likelihood-Ratio-**Statistik** zum Testen zu verwenden

Machine Translated by Google

Hypothesen.Wald, Abraham (1941) verallgemeinerte den Likelihood-Ratio-Test auf mehr Kom

komplizierte Hypothesen. Die asymptotischen Ergebnisse über die Verteilung dieser Statistik (vorbehaltlich bestimmte Regularitätsbedingungen) wurden erstmals von Wilks, Samuel Stanley (1962) vorgestellt.

MATHEMATISCHE ASPEKTE

Beim Likelihood-Ratio-Test wird die **Nullhypothese** verworfen, wenn die Wahrscheinlichkeit unter der liegt **Alternativhypothese** deutlich größer als die Wahrscheinlichkeit unter der Nullhypothese. Daher besteht das Problem nicht darin, das zu finden Der am besten geeignete Parameter, aber zu wissen, ob Unter der Alternativhypothese erhalten wir eine deutlich bessere Schätzung.

Das Neyman-Pearson-Theorem besagt, dass für alle **signifikanten Ebenen** ÿ die Wahrscheinlichkeit Verhältnistest (zum Testen der einfachen Hypothese) hat mehr Leistung als jeder andere Test.

Seien x1, *x2,...,xn* n *unabhängige* Beobachtungen *(n* > 1) einer **Zufallsvariablen** *X* folgt einer Wahrscheinlichkeitsverteilung mit **Parameter** ÿ (oder ein Parametervektor ÿ). Wir bezeichne mit *L (ÿ;x1, x2,...,xn)* die Wahrscheinlichkeit Funktion, die für festes *xi* nur von abhängt ÿ, und wir bezeichnen damit den Raum aller möglichen Werte von ÿ und 0, wobei 1 Unterraum sind von .

Wir unterscheiden zwischen:

Einfache Nullhypothese vs. eine einfache Alternativhypothese

$$H0 : \ddot{y} = \ddot{y}0$$

vs. H1 : ÿ = ÿ1 .

In diesem Fall beträgt das Wahrscheinlichkeitsverhältnis:

$$\ddot{y} = \ddot{y}\,(x1, x2,...,xn) = \frac{L\,(\ddot{y}0;x1, x2,...,xn)}{L\,(\ddot{y}1;x1, x2,...,xn)} .$$

Das Wahrscheinlichkeitsverhältnis für ein **Signifikanzniveau** ÿ wird durch die Entscheidungsregel definiert:

H0 ablehnen, wenn ÿ>ÿ0 ,

wobei ÿ0 der durch definierte kritische Wert ist Signifikanzniveau (ÿ) und die Verteilung von ÿ unter der Nullhypothese:

$$\ddot{y} = P\,(\ddot{y}\,(x1, x2,...,xn)\ \ddot{y}\ \ddot{y}0|H0) .$$

Das Problem besteht hier darin, das **Kritische** zu finden **Region ,** da es keine allgemein akzeptierte gibt Verfahren dazu. Die Idee ist zu suchen eine Statistik, deren Verteilung bekannt ist.

Einfache Nullhypothese vs. eine allgemeine Alternative

$$H0 : \ddot{y} = \ddot{y}0$$

vs. H1 : ÿ ÿ 1 (zum Beispiel ÿ<ÿ0) .

Wir stellen:

$$\ddot{y} = \frac{\sup\ddot{y}1\ L\,(\ddot{y}0;x1, x2,...,xn)}{L\,(\ddot{y}1;x1, x2,...,xn)} ,$$

wobei supÿÿ1 L(ÿ; x1, *x2,...,xn)* repre sendet den Wert der Likelihood-Funktion für den Parameter, der diese Like-Lihood-Funktion im eingeschränkten Parameter maximiert Raum (ÿ ÿ 1), und wir suchen erneut nach dem Verteilung von ÿ unter H0.

Wenn die Alternativhypothese H1 ÿ = ÿ0 ist, wir können ÿ schreiben als:

$$\frac{L\,\ddot{y}\hat{};x1, x2,...,xn}{L\,(\ddot{y}0;x1, x2,...,xn)} ,$$

wobei ÿˆ die Maximum-Likelihood-Schätzung ist für ÿ.

Allgemeine Hypothese H_0 : ÿ ÿ 0 vs. Alternativhypothese H_1 : ÿ ÿ \0 = {ÿ ÿ ; ÿ /ÿ 0}

Betrachten Sie die folgende **Statistik:**

$$\ddot{y} = \frac{\sup\ddot{y}0\ L\ (\ddot{y};x1, x2,...,xn)}{\sup\ddot{y}\ddot{y}\ L\ (\ddot{y};x1, x2,...,xn)}$$

Machine Translated by Google

und durchsuchen Sie den Ablehnungsbereich wie in den vorherigen Fällen.

Das supÿÿ0 L(ÿ;x1, *x2,...,xn)* repräsentiert der Wert der Wahrscheinlichkeitsfunktion für der Parameter, der die Likelihood-Funktion (themaximumlikelihoodesti mate) auf dem begrenzten Raum der Parameter maximiert. Die supÿÿ *L (ÿ;x1, x2,...,xn)* repräsentiert dasselbe, außer für den gesamten Parameterraum. Je größer der Unterschied zwischen diesen beiden Werten ist, desto ungeeigneter ist das eingeschränkte Modell (und H0 wird abgelehnt).

Unter H0 folgt die Verteilung von ÿ2 log (ÿ) asymptotisch einer **Chi-Quadrat** -Verteilung mit *u* Freiheitsgraden, mit *u* = dim ÿ dim 0 (d. h. der Anzahl der durch H0 angegebenen Parameter). Mit anderen Worten:

$$\lim_{n\ddot{y}\ddot{y}\log} \frac{\sup_{\ddot{y}\ddot{y}} L\,(\ddot{y};x1,\,x2,...,xn)\;\ddot{y}\;2}{\sup_{\ddot{y}\ddot{y}0} L\,(\ddot{y};x1,\,x2,...,xn)} = \ddot{y}2\,(u)\,.$$

Für die zuvor behandelten Einzelfälle ergibt sich folgendes Ergebnis:

ÿ 2 log ÿÿÿ2 auf 1 Freiheitsgrad

wenn *n* ÿ ÿ .

BEREICHE UND EINSCHRÄNKUNGEN

Wenn die Datenverteilung normal ist, entspricht der Like-Lihood-Ratio-Test dem **Fisher-Test.** Der Vorteil des Letzteren (wenn er explizit konstruiert werden kann) besteht darin, dass er die beste Leistung und Kraft [gemäß dem Neyman-Pearson-Theorem] zum Testen der einfachen Hypothese garantiert.

BEISPIELE

Beispiel 1

Sei X eine normale Zufallsvariable mit Mittelwert ÿ und Varianz ÿ2. Wir wollen folgendes testen Lowing-Hypothese:

$$H0 : \ddot{y} = 0 \quad , \quad \ddot{y}2 > 0$$

$$H1 : \ddot{y} = 0 \quad , \quad \ddot{y}2 > 0\,.$$

Unter Berücksichtigung des Parameterraums = {(ÿ, ÿ2); ÿÿ <ÿ< *ÿ,* 0 < ÿ2 < ÿ} und ein Unterraum 1 = {(ÿ, ÿ2) ÿ ; ÿ = 0} kann die Hypothese nun wie folgt geschrieben werden:

$$H0 : \ddot{\mathbf{y}} = \ddot{y}, \qquad \ddot{y}2\ \ddot{y}\ 0$$

$$H1 : \ddot{\mathbf{y}} = \ddot{y}, \qquad \ddot{y}2\ \ddot{y}\ \backslash 0\,.$$

Betrachten Sie nun *x1,...,xn n* Beobachtungen der Variablen *X (n* > 1). Die Wahrscheinlichkeitsfunktion ist (siehe Beispiel unter **Maximum Like Lihood):**

$$L\,\mathbf{(\ddot{y};}\,x1,...,xn) = L\,\ddot{y},\,\ddot{y}2;\,x1,...,xn$$

$$= \frac{1}{2\ddot{y}\ddot{y}2}^{n/2} \ddot{y} \cdot \exp \frac{\sum_{i=1}^{N} 2\,(xi\;\ddot{y}\;\ddot{y})}{2\ddot{y}2} \ddot{y}\,.$$

Wir schätzen den Stichprobenmittelwert und die Varianz mithilfe der Maximum-Likelihood-Schätzungen:

$$\ddot{y} = \frac{1}{N} \cdot \sum_{i=1}^{N} xi = \bar{x}\,,$$

$$\ddot{y}2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (xi\;\ddot{y}\;\ddot{y})^{2}\,.$$

Daher lauten die Wahrscheinlichkeitsfunktionen für die Nullhypothese:

$$L\ 0,\ddot{y}2;\,x1,...,xn$$

$$= \frac{\ddot{y}}{\ddot{y}} \frac{N}{2\ddot{y}\ \sum_{i=1}^{N} x2} \frac{\ddot{y}}{\ddot{y}}^{n/2} \ddot{y} \cdot \exp \frac{\sum_{i=1}^{N} x2}{2} \frac{\ddot{y}}{\ddot{y}}$$

$$= \frac{\ddot{y}}{\ddot{y}} \frac{N}{2\ddot{y}\ \sum_{i=1}^{N} x2} \frac{\ddot{y}}{\ddot{y}}^{n/2} \exp \frac{\ddot{y}n}{2}$$

Machine Translated by Google

$$= \left(\frac{n \exp(\ddot{y}1)}{2\ddot{y} \sum_{i=1}^{N} x_{\bar{}}^2}\right)^{n/2} \ddot{y},$$

und für die Alternativhypothese:

$$L(\ddot{y}, \ddot{y}2; x1, \ldots, xn)$$

$$= \left(\frac{N}{2\ddot{y} \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}\right)^{n/2} \ddot{y}$$

$$\ddot{y} \cdot \exp\left(\frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}{2 \ddot{y} \ddot{y}}\right)$$

$$= \left(\frac{1}{2\ddot{y} \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}\right)^{n/2} \cdot \exp\left(\frac{\ddot{y}n}{2}\right)$$

$$= \left(\frac{n \exp(\ddot{y}1)}{2\ddot{y} \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}\right)^{n/2},$$

Das Wahrscheinlichkeitsverhältnis ist also:

$$\ddot{y} = \ddot{y}(x1, x2, \ldots, xn)$$

$$= \frac{L(0, \ddot{y}2; x1, x2, \ldots, xn)}{L(\ddot{y}, \ddot{y}2; x1, x2, \ldots, xn)}$$

$$= \left(\frac{n \exp(\ddot{y}1)}{2\ddot{y} \sum_{i=1}^{N} x_{\bar{}}^2}\right)^{n/2} \ddot{y}$$

$$\left(\frac{n \exp(\ddot{y}1)}{2\ddot{y} \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}\right)^{n/2} \ddot{y}$$

$$= \left(\frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}{n \sum_{i=1}^{N} x_{\bar{}}^2}\right)^{n/2} \ddot{y}.$$

An dieser Stelle muss man die Verteilung finden von ÿ. Dazu erinnern wir uns daran $\sum_{i=1}^{N} x_{\bar{}}^2 = \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2 + n\bar{x}2$, und wir erhalten:

$$\ddot{y} = \left(\frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2 + n\bar{x}2}\right)^{n/2} \ddot{y}$$

$$= 1 \ddot{y} \left(1 + \frac{n\bar{x}2}{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}\right)^{n/2} \ddot{y}.$$

Wir suchen nach ÿ0 , für das die Nullhypothese nicht abgelehnt wird, wenn ÿ ÿ ÿ0, was wir wie folgt schreiben können:

$$\ddot{y} \ddot{y} \ddot{y}0$$

$$\text{oder } 1 \ddot{y} \ddot{y} \left(1 + \frac{n\bar{x}2}{2 \sum_{i=1}^{N} (xi \ddot{y} \bar{x})}\right)^{n/2} \ddot{y} \ddot{y}0$$

$$\text{oder } \left(1 + \frac{n\bar{x}2}{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}\right)^{n/2} \ddot{y} \ddot{y}\ddot{y}1_0$$

$$\text{oder } \frac{n\bar{x}2}{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2} \ddot{y} \ddot{y}\ddot{y}2_{0n} \ddot{y} 1.$$

Indem wir jeden Term der Ungleichung mit $(n \ddot{y} 1)$ multiplizieren, erhalten wir:

$$\ddot{y} \ddot{y} \ddot{y}0 \text{ oder}$$

$$\frac{\ddot{y}n \bar{|}x\bar{|}}{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2 \quad (n \ddot{y} 1)}$$

$$\ddot{y} (n \ddot{y} 1) \ddot{y}\ddot{y}2_{0n} \ddot{y} 1.$$

Machine Translated by Google

Da die Zufallsvariable

$$T = \frac{\overline{X} \ddot{y} \ddot{y}}{\ddot{y} \, (n \ddot{y} \, 1)} ,$$

wobei ÿ2 die Stichprobenvarianz ist, folgt einer *t*- Verteilung mit *(n* ÿ 1) Freiheitsgraden; wir können äquivalent schreiben:

$$T = \frac{\ddot{y} n \overline{X} \ddot{y} \ddot{y}}{n\ddot{y}2 \, (n \ddot{y} \, 1)} ,$$

was für einen auf Null gesetzten Mittelwert Folgendes ergibt:

$$T = \frac{\ddot{y} n \overline{X}}{n\ddot{y}2 \, (n \ddot{y} \, 1)} = \frac{\ddot{y} n \overline{X}}{\sqrt{\frac{\sum_{i=1}^{N} x2}{(n \ddot{y} \, 1)}}} .$$

Man muss nun nach dem Wert ÿ0 suchen , für den *c := (n* ÿ 1) ÿÿ2/n gilt $_{0}$ ÿ 1 ist die Wert einer Student-Variablen für ein gewähltes Signifikanzniveau ÿ:

$$\ddot{y}0 = \left(\frac{c2}{n \ddot{y} \, 1} + 1 \right)^{\ddot{y}n/2} .$$

Wenn wir beispielsweise eine Stichprobe mit der Größe *n* = 6 und für ÿ = 0,05 angenommen haben, finden wir *z5,0,05/2* = 2,751, sodass es ausreicht, Folgendes zu sehen:

$$\ddot{y} \ddot{y} \quad \frac{(2.751)^{2}}{5} + 1 \quad ^{\ddot{y}3} = 0{,}063$$

damit die Nullhypothese wahr ist.

Beispiel 2

Zwei unabhängige Zufallsstichproben *X1,...,Xn* und *Y1,...,Ym (m + n* > 2) folgen Normalverteilungen mit Parametern ÿ1, ÿ2 bzw. ÿ2, ÿ2 . Wir möchten die folgende Hypothese testen:

$$H0 : \ddot{y}1 = \ddot{y}2$$
$$H1 : \ddot{y}1 = \ddot{y}2 .$$

Wir definieren die Mengen:

$$= (\ddot{y}1, \ddot{y}2, \ddot{y}2); \ddot{y}\ddot{y} < \ddot{y}1, \ddot{y}2 < \ddot{y}, 0 < \ddot{y}2 < \ddot{y} \quad ,$$

$$1 = (\ddot{y}1, \ddot{y}2, \ddot{y}2) \ddot{y} \ ; \ddot{y}1 = \ddot{y}2 .$$

Wir können die Hypothese nun wie folgt formulieren:

$$H0 : \ddot{y}1, \ddot{y}2, \quad \ddot{y}2 \ddot{y} \, 1$$
$$H1 : \ddot{y}1, \ddot{y}2, \quad \ddot{y}2 \ddot{y} \, \backslash 1 .$$

Die Wahrscheinlichkeitsfunktionen sind also:

$$L \ddot{y}1, \ddot{y}1, \ddot{y}2; x1,...,xn, y1,...,ym = \frac{1}{\ddot{y}2\sqrt{\ddot{y}\ddot{y}}}^{n+m} \ddot{y} \cdot \exp \ddot{y} \frac{\sum_{i=1}^{N} (xi \ddot{y} \ddot{y}1)^{2} + \sum_{i=1}^{M} (yi \ddot{y} \ddot{y}1)^{2}}{2\ddot{y}2} \ddot{y}, \ddot{y} \ddot{y}\ddot{y} \ddot{y},$$

$$L \ddot{y}1, \ddot{y}2, \ddot{y}2; x1,...,xn, y1,...,ym = \frac{1}{\ddot{y} \, 2\ddot{y}\sqrt{\ddot{y}2}}^{n+m} \ddot{y} \cdot \exp \ddot{y} \frac{\sum_{i=1}^{N} (xi \ddot{y} \ddot{y}1)^{2} + \sum_{i=1}^{M} (yi \ddot{y} \ddot{y}2)^{2}}{2\ddot{y}2} \ddot{y}, \ddot{y} \ddot{y}\ddot{y} \ddot{y}.$$

Mithilfe der Maximum-Likelihood-Schätzungen, die für die Nullhypothese gelten:

$$\ddot{y}1 = \frac{\sum_{i=1}^{N} xi + \sum_{i=1}^{M} yi}{n + m} = \frac{n\overline{x} + me\overline{in}}{n + m},$$

$$\ddot{y}2 = \frac{\sum_{i=1}^{N} (xi \ddot{y} \ddot{y}1)^{2} + \sum_{i=1}^{M} (yi \ddot{y} \ddot{y}1)^{2}}{n + m}$$

Machine Translated by Google

und für die Alternativhypothese:

$$\ddot{y}\,1 = \frac{\sum_{i=1}^{N} xi}{N} = \bar{x},$$

$$\ddot{y}2 = \frac{\sum_{i=1}^{M} yi}{M} = \bar{y},$$

$$\ddot{y}^{2} = \frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^{2} + \sum_{i=1}^{M} (yi \ddot{y} \bar{y})^{2}}{n + m},$$

Wir erhalten das Wahrscheinlichkeitsverhältnis:

$$\ddot{y} = \frac{L\,\ddot{y}1, \ddot{y}1, \ddot{y}2; x1,...,xn, y1,...,ym}{L\,\ddot{y}\,1, \ddot{y}\,2, \quad \ddot{y}\quad 2\,; x1,...,xn, y1,...,ym}$$

$$= \frac{2\ddot{y}}{n + m}^{\ddot{y}(n+m)/2} \exp \frac{\ddot{y}n + m}{2} \cdot \sum_{i=1}^{N} \left(mein\, xi \ddot{y} \frac{n\bar{x} +}{n + m} \right)^{2} + m \sum_{i=1}^{M} \left(mein\, yi \ddot{y} \frac{n\bar{x} +}{n + m} \right)^{2 \ddot{y}(n+m)/2} \cdot \frac{2\ddot{y}}{n + m}^{(n+m)/2} \text{expn} \frac{+ m}{2}$$

$$= \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^{2 + m} \sum_{i=1}^{M} (yi \ddot{y} \bar{y})^{2}{}^{(n+m)/2}$$

$$= \sum_{i=1}^{N} \left(mein\, xi \ddot{y} \frac{n\bar{x} +}{n + m} \right)^{2} + m \sum_{i=1}^{M} \left(mein\, yi \ddot{y} \frac{n\bar{x} +}{n + m} \right)^{2 \ddot{y}(n+m)/2} \cdot \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^{2 + m} \sum_{i=1}^{M} (yi \ddot{y} \bar{y})^{2}{}^{(n+m)/2}$$

$$= \sum_{i=1}^{N} (xi \ddot{y} \bar{x})^{2 + m} \sum_{i=1}^{M} (yi \ddot{y} \bar{y})^{2}{}^{(n+m)/2}$$

$$\sum_{i=1}^{N} \left(xi \ddot{y} \frac{n\bar{x} + mein}{n + m} \right)^{2} + \sum_{i=1}^{M} \left(mein\, yi \ddot{y} \frac{n\bar{x} +}{n + m} \right)^{2 (n+m)/2}.$$

Die Statistik *ÿ2/ (n+m)* ist gegeben durch:

$$\frac{\sum_{i=1}^{N} Xi \ddot{y} X^{2} + \sum_{i=1}^{M} Yi \ddot{y} Y^{2}}{\sum_{i=1}^{N} Xi \ddot{y} \frac{n\bar{X}+m\bar{Y}}{n+m}^{2} + \sum_{i=1}^{M} Yi \ddot{y} \frac{n\bar{X}+m\bar{Y}}{n+m}^{2}}.$$

Um die Verteilung dieser Statistik zu ermitteln,

verwenden wir die folgenden Argumente (beachten Sie, dass $\sum_{i=1}^{N} Xi \ddot{y} X = 0$):

$$\sum_{i=1}^{N} Xi \ddot{y} \frac{n\bar{X} + m\bar{Y}}{n + m}^{2}$$

$$= \sum_{i=1}^{N} Xi \ddot{y} \bar{X} + X \ddot{y} \quad \frac{n\bar{X} + m\bar{Y}}{n + m}^{2}$$

$$= \sum_{i=1}^{N} Xi \ddot{y} \bar{X}^{2} + n\,X \ddot{y} \frac{n\bar{X} + m\bar{Y}}{n + m}^{2}$$

$$= \sum_{i=1}^{N} Xi \ddot{y} X^{2} + \frac{nm2}{(n + m)^{2}} \bar{X} \ddot{y} Y^{2}.$$

Dasselbe gilt für Y:

$$\sum_{i=1}^{M} Yi \ddot{y} \frac{n\bar{X} + m\bar{Y}}{n + m}^{2}$$

$$= \sum_{i=1}^{M} Yi \ddot{y} Y^{2} + \frac{mn2}{2\,(n + m)} \bar{X} \ddot{y} Y^{2}.$$

Machine Translated by Google

Wir können jetzt schreiben:

$$\ddot{y}2/(n+m) = \frac{\sum_{i=1}^{N} Xi \ddot{y} \bar{X}^{2} + m \sum_{i=1}^{} Yi \ddot{y} \bar{Y}^{2}}{\sum_{i=1}^{N} Xi \ddot{y} \bar{X}^{2} + m \sum_{i=1}^{} Yi \ddot{y} \bar{Y}^{2}} + \frac{nm}{n+m} \bar{X} \ddot{y} \bar{Y}^{2} \ddot{y}1$$

$$= \ddot{y} 1 + \frac{\frac{nm}{n+m} \bar{X} \ddot{y} \bar{Y}^{2}}{\sum_{i=1}^{N} Xi \ddot{y} \bar{X}^{2} + \sum_{i=1}^{M} Yi \ddot{y} \bar{Y}^{2}} \ddot{y}1 \ddot{y} .$$

Wenn die Nullhypothese wahr ist, dann die Zufallsvariable

$$T = \frac{\frac{nm}{n+m} \bar{X} \ddot{y} \bar{Y}}{\sqrt{\frac{\sum_{i=1}^{N} (Xi \ddot{y} \bar{X})^{2} + m \sum_{i=1}^{} (Yi \ddot{y} \bar{Y})^{2}}{n+m \ddot{y} 2}}}$$

folgt der Student-Verteilung mit $n+m\ddot{y}2$ Freiheitsgraden. Daher erhalten wir:

$$\ddot{y}2/(n+m) = \frac{1}{1 + (n + m \ddot{y} 2) T2 n +} = \frac{m \ddot{y} 2 n +}{m \ddot{y} 2 + T2} .$$

Wir suchen nach $\ddot{y}0$, für das die Nullhypothese abgelehnt wird, vorausgesetzt $\ddot{y} \ddot{y} \ddot{y}0 \ddot{y} 1$:

$$\ddot{y} = \frac{n + m \ddot{y} 2 n}{+ m \ddot{y} 2 + T2}^{2/(n+m)} \ddot{y} \ddot{y}0$$

$$\ddot{y} /T/ \ddot{y} \qquad \frac{\ddot{y}n+m}{2 \ddot{y} 0} \ddot{y} 1 \qquad := c .$$

$$(n + m \ddot{y} 2)$$

Es genügt nun, den kritischen Wert c anzunehmen aus der Tabelle einer Student-Zufallsvariablen mit festem ÿ (dem gewünschten Wert) und n der Stichprobengröße. Zum Beispiel mit $n = 10$, $m = 6$, und ÿ = 0,05, finden wir:

$$c = z14,0,025 = 2,145 ,$$

$$\ddot{y}0 = \frac{c2}{n + m \ddot{y} 2} \ddot{y} 1 \ddot{y}2/(n+m) = \frac{(2.145)^{2}}{14} \ddot{y} 1 \ddot{y}1/8 = \ddot{y}1,671 .$$

WEITERLESEN

Alternative Hypothese
Fisher-Test
Hypothese
Hypothesentest
Maximale Wahrscheinlichkeit
Nullhypothese

L

REFERENZEN

Cox, DR, Hinkley, DV: Theoretische Statistik. Chapman & Hall, London (1973)

Edwards, AWF: Wahrscheinlichkeit. Ein Bericht über das statistische Konzept der Wahrscheinlichkeit und seine Anwendung auf wissenschaftliche Schlussfolgerungen. Cambridge University Press, Cambridge (1972)

Kendall, MG, Steward, A.: The Advanced Theory of Statistics, vol. 2. Griffin, London (1967)

Neyman, J., Pearson, ES: Zur Verwendung und Interpretation bestimmter Testkriterien für statistische Schlussfolgerungen, Teile I und II. Biometrie 20A, 175–240, 263–294 (1928)

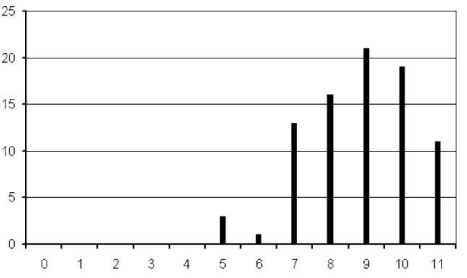
Wald, A.: Asymptotisch leistungsfähigste Tests statistischer Hypothesen. Ann. Mathematik. Stat. **12,** 1–19 (1941)

Wald, A.: Einige Beispiele für asymptotisch leistungsstärkste Tests. Ann. Mathematik. Stat. **12,** 396–408 (1941)

Machine Translated by Google

Wilks, SS: Mathematische Statistik. Wiley, New York (1962)

Liniendiagramm



Das Liniendiagramm ist eine besondere Art von Häufigkeitsdiagrammen. Es ist nützlich, um eine **Häufigkeitsverteilung** einer diskreten **Zufallsvariablen darzustellen.** Auf der Abszisse sind die **Werte** der **Variablen** und auf der Ordinate die diesen Werten entsprechenden Häufigkeiten dargestellt.

WEITERLESEN

Punktdiagramm

Häufigkeitsverteilung

Grafische Darstellung

Histogramm

GESCHICHTE Siehe **grafische Darstellung.**

MATHEMATISCHE ASPEKTE

Das Liniendiagramm hat zwei **Achsen.** Auf der Abszisse sind die **Werte** *xi* der diskreten **Zufallsvariablen** angegeben. Über jeden *xi*- Wert wird eine Linie gezogen, deren Länge der diesem Wert entsprechenden **Häufigkeit** *ni* (oder relativen Häufigkeit *fi)* entspricht .

BEISPIELE

Die Verteilung der Schuhgrößen einer **Stichprobe** von 84 Männern ist in der folgenden Tabelle angegeben.

Schuhgröße	Häufigkeit
5	3
6	1
7	13
8	16
9	21
10	19
11	11
Gesamt	84

Hier ist das daraus erstellte Liniendiagramm Tisch:

Lineares Programmieren

Bei der mathematischen Programmierung geht es um die optimale Zuteilung begrenzter Ressourcen, um die gewünschten Ziele zu erreichen. Inlineare ProgrammierungDas untersuchte thematische Modell wird durch die Verwendung linearer Beziehungen ausgedrückt. Ein lineares Programmierproblem tritt in Form einer Reihe linearer Gleichungen (oder linearer Qualitäten), sogenannter Einschränkungen, und einer linearen Funktion auf, die das Ziel angibt, die als Zielfunktion bezeichnet wird.

GESCHICHTE

Mathematische Programmierung tauchte um 1930 im Bereich der Wirtschaftswissenschaften auf. Einige Bemerkenswerte Arbeiten zur linearen Programmierung wurden von John Von Neumann (1935–1936) veröffentlicht. 1947 stellte George B. Dantzig die Simplex-Methode zur Lösung linearer Programmierprobleme vor.

MATHEMATISCHE ASPEKTE

Ein lineares Programmierproblem kann wie folgt definiert werden.

Machine Translated by Google

Finden Sie eine Reihe von Werten für die Variablen x1, *x2,...,xn* , das die Zielfunktion linear maximiert (oder minimiert):

$$z = c1 \cdot x1 + c2 \cdot x2 + ... + cn \cdot xn$$

unterliegen den linearen Einschränkungen:

$$a11 \cdot x1 + a12 \cdot x2 + ... + a1n \cdot xn = b1,$$

$$a21 \cdot x1 + a22 \cdot x2 + ... + a2n \cdot xn = b2,$$

$$\vdots$$

$$am1 \cdot x1 + am2 \cdot x2 + ... + amn \cdot xn = bm,$$

und zu den nichtnegativen Einschränkungen:

$$xj \geq 0 , j = 1, 2,...,n,$$

wobei *aij, bi* und *cj* bekannte Konstanten sind.
Mit anderen Worten bedeutet dies, dass wir bei einem linearen Programmierproblem nach einer nichtnegativen Lösung suchen, die die gestellten Einschränkungen erfüllt und die Zielfunktion optimiert.
Wir können ein lineares Programmierproblem mit der folgenden Matrixnotation schreiben:

Maximieren (oder minimieren) Sie $\mathbf{z} = \mathbf{c} \cdot \mathbf{x}$

unterliegt $\mathbf{A} \cdot \mathbf{x} = \mathbf{B}$

und $\mathbf{x} \geq \mathbf{0}$,

Dabei ist **c** ein (1×n)Linienvektor, **x** ein (n×1) Spaltenvektor, **A** eine *(m* × n) **-Matrix** und **b** ein *(m* × 1) Spaltenvektor und **0** ist der Nullvektor mit *n* Komponenten.
Das wichtigste Werkzeug zur Lösung eines linearen Programmierproblems ist die Simplex-Methode.
Dies besteht aus einer Reihe von Regeln, die sein sollten befolgt werden, um die Lösung für ein bestimmtes Problem zu erhalten. Es handelt sich um eine iterative Methode, die in endlich vielen Fällen eine exakte Lösung liefert Iterationen.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die lineare Programmierung strebt eine optimale Allokation dieser Ressourcen an, die ein oder mehrere Güter produzieren können. Dies geschieht durch Maximierung (oder Minimierung) von Funktionen wie dem Gewinn (oder den Kosten). Daher werden die Koeffizienten der Zielfunktion oft als bezeichnet „Preise“, die den Variablen zugeordnet sind.

BEISPIELE

Ein Unternehmen stellt Tische und Stühle mit zwei Maschinen A und B her. Jedes Produkt wird auf beiden Maschinen hergestellt. Um einen Stuhl herzustellen, werden 2 Stunden Maschine A und 1 Stunde Maschine B benötigt Hrsg. Um einen Tisch herzustellen, werden 1 Stunde Maschine A und 2 Stunden Maschine B benötigt. Die Firma macht einen Gewinn von 3,00 CHF pro Stuhl und 4,00 CHF pro Tisch. Die beiden Maschinen stehen maximal 12 Stunden täglich zur Verfügung.
Auf der Grundlage der oben genannten Daten suchen wir nach dem täglichen Produktionsplan, der den maximalen Gewinn garantiert. Mit anderen Worten: Das Unternehmen möchte das optimale Produktionsprogramm kennen, also die Menge an Stühlen und Tischen, die pro Tag produziert werden muss, um den maximalen Gewinn zu erzielen.
Wir bezeichnen mit x1 die Anzahl der Stühle und mit x2 die Anzahl der Tische, die pro Tag produziert werden müssen. Der Tagesgewinn ergibt sich aus der folgenden Funktion:

$$z = 3 \cdot x1 + 4 \cdot x2 .$$

L

Machine Translated by Google

Diese Zielfunktion muss maximiert werden.
Wir wissen, dass das Produkt x1 2 Stunden benötigt
Maschine A, während Produkt x2 1 Stunde benötigt
Maschine A. Daher die gesamten täglichen Stunden für
Maschine A ist:

$$2 \cdot x1 + 1 \cdot x2$$

da x1 Stühle und x2 Tische produziert werden.
Die Gesamtzeit für Maschine A kann jedoch nicht berechnet werden
12 Stunden überschreiten. Mathematisch gesehen ist das so
bedeutet:

$$2x1 + x2 \ddot{y} 12 \quad .$$

Auf die gleiche Weise finden wir für Maschine B:

$$x1 + 2x2 \ddot{y} 12 \quad .$$

Wir haben zwei Einschränkungen. Darüber hinaus haben wir
kann keine negativen Stuhlwerte erzeugen
und Tabellen, also zwei zusätzliche Einschränkungen
hinzugefügt:

$$x1 \ddot{y} 0 \text{ und } x2 \ddot{y} 0 .$$

Wir möchten die Werte für die Variablen x1 finden
und x2 , die die Einschränkungen erfüllen und den
Gewinn maximieren. Endlich können wir das vertreten
mathematisches Problem auf folgende Weise:

$$\max z = 3x1 + 4x2 ,$$

$$\text{M } 2x1 + x2 \ddot{y} 12 \quad ,$$

$$x1 + 2x2 \ddot{y} 12 \quad ,$$

$$\text{und } x1, x2 \ddot{y} 0$$

(st bedeutet unterliegen). Der maximale Profit ergibt
sich aus dem Paar (x1, x2) = (4, 4);
28.00 CHF Gewinn durch die Produktion von
4 Stühle und 4 Tische. Für dieses Problem ist die
Das Ergebnis kann durch Berechnung des Gewinns ermittelt werden
für jedes mögliche Paar (es gibt weniger als
36 davon.)

WEITERLESEN

Unternehmensforschung
Optimierung

VERWEISE

Arthanari, TS, Dodge, Y.: Mathematik
Programmierung in der Statistik. Wiley, Neu
York (1981)

Dantzig, GB: Anwendungen und Verlängerungen der
linearen Programmierung. Dun od, Paris (1966)

Hadley, G.: Lineare Programmierung. Addison Wesley,
Reading, MA (1962)

Von Neumann, J.: Über ein ökonomisches
Gleichungssystem und eine Verallgemeinerung
meines Brouwerschen Fixpunksatzes. Ergebn.
Mathematik. Kolloqu. Wien 8,
73–83 (1937)

Logarithmische Transformation

Eines der nützlichsten und gebräuchlichsten **Trans**
Formationen ist die logarithmische Transformation.
In der Tat, bevor eine lineare ausgeführt wird
Regression, es könnte sinnvoll sein, zu ersetzen
die **abhängige Variable** *Y* durch ihren Logarithmus
log *(Y)* oder durch eine Linearkombination
von letzterem $a + b \cdot \log (Y)$ (*a* und *b* sind
Konstanten angepasst durch die **kleinsten Quadrate).**
Eine solche Operation würde die **Varianz** von *Y*
stabilisieren und die Verteilung vornehmen
der transformierten Variablen näher an nor
mal.

Eine logarithmische Transformation wäre ebenfalls möglich
Erlauben Sie einem, von einem Multiplikativ zu einem zu wechseln
additives **Modell** mit einfacheren Schätzungen
zu definieren und einfacher zu interpretieren.
Andererseits ist die logarithmische Transformation von
Eine **Gamma-** Zufallsvariable *Y* ergibt eine neue
Zufallsvariable, die näher am **Normalen** liegt
Verteilung. Letzteres könnte in einer
Regressionsanalyse leicht angepasst werden.

Machine Translated by Google

WEITERLESEN

Varianzanalyse
Abhängige Variable
Regressionsanalyse

Logistische Regression

Bei einer binär abhängigen Variablen bietet die logistische Regression die Möglichkeit, ein Regressionsmodell anzupassen. Es ermöglicht die Untersuchung der **Beziehung** zwischen einem Verhältnis und einer Reihe erklärender Variablen, entweder quantitativ oder qualitativ.

MATHEMATISCHE ASPEKTE

Das logistische Regressionsmodell mit der abhängigen Variablen *Y* und den unabhängigen Variablen X1, *X2,..., _p* wird ausgedrückt durch:

$$\text{logit } \ddot{y}\ Xij\ i=1,...,p \quad = \text{Protokoll} \quad \frac{P\ Yj = 1}{P\ Yj = 0}$$

$$= \ddot{y}0 + \sum_{i=1}^{P} \ddot{y}iXij\ ,$$

$$j = 1,...,n,$$

wobei *Yj (j = 1,...,n)* den Beobachtungen der Variablen *Y* und *Xij* den Beobachtungen der unabhängigen Variablen entspricht und ÿ die Wahrscheinlichkeitsfunktion (des Erfolgs) der logistischen Regression ist:

$$\ddot{y}\ Xij \quad _{i=1,...,p} = \frac{\exp \ddot{y}0 + 1 \quad \sum_{i=1}^{p} \ddot{y}iXij}{+ \exp \ddot{y}0 + \quad \sum_{i=1}^{p} \ddot{y}iXij}\ .$$

Äquivalent können wir schreiben:

$$P\ Yj = 1 = 1 + \frac{\exp \ddot{y}0 + \quad \sum_{i=1}^{P} \ddot{y}iXij}{\exp \ddot{y}0 + \quad \sum_{i=1}^{P} \ddot{y}iXij}\ ,$$

Dies ermöglicht die Berechnung der Wahrscheinlichkeit, dass die Variable *Y* den Wert 1 annimmt, abhängig von der Realisierung der unabhängigen Variablen X1, *X2,...,Xp.*

Eine interessante Eigenschaft ist, dass die Modellparameter (ÿ0, *ÿ1,..., ÿp)* beliebige Werte annehmen können, da alle Werte der Logit-Transformation ÿ *(Y) logit(ÿ* (Y)) = log ÿ *(Y)* entsprechen. $\frac{1\ \ddot{y}}{}$

eine Zahl zwischen 0 und 1 für *P (Y* = 1).

Die Modellparameter können mithilfe der **Maximum-Likelihood-Methode geschätzt werden,** wodurch die folgende Funktion maximiert wird:

$$L\ (\ddot{y}0,\ \ddot{y}1,...,\ \ddot{y}n) = \quad \prod_{j=1}^{N} \ddot{y}\ Xij\ i=1,...,pYi \cdot$$

$$\cdot\ 1\ \ddot{y}\ \ddot{y}\ Xij\ i=1,...,p1\ddot{y}Yj$$

oder auf einfachere Weise durch Maximieren der Log-Likelihood-Funktion:

$$\log L\ (\ddot{y}0,\ \ddot{y}1,...,\ \ddot{y}n)$$

$$= \sum_{j=1}^{N} Yj \log \ddot{y}\ Xij \quad _{i=1,...,p}$$

$$+ 1\ \ddot{y}\ Yj \log 1\ \ddot{y}\ \ddot{y}\ Xij\ i=1,...,p \quad .$$

Wir können daher Hypothesentests für die Parameterwerte (z. B. ÿ0 = 0, ÿ3 = ÿ2 usw.) durchführen, indem wir den **Likelihood-Ratio-Test verwend**

Die Interpretation der Regressionskoeffizienten, die binären unabhängigen Variablen entsprechen, wird im nächsten Absatz beschrieben.

Wenn eine unabhängige Variable *X* stetig ist, spiegelt der Regressionskoeffizient das relative Risiko eines Anstiegs von X um eine Einheit wider

Da die unabhängige Variable X mehr als eine Klasse hat, müssen wir sie in binäre Variablen umwandeln. Die Interpretation des Koeffizienten ist dann klar und einfach.

L

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Die logistische Regression wird in **der Epidemiologie** häufig verwendet, da:

- Seine sigmoidale Form (des zuvor angegebenen Ausdrucks $P\ Yj = 1$) passt sehr gut die normalerweise beobachtete Beziehung zwischen eine Dosis X und die Häufigkeit einer Erkrankung Y: $\frac{\exp(x)}{1+\exp(x).}$
- Es ist leicht zu interpretieren: das Assoziationsmaß zwischen einer Krankheit und einem Risikofaktor M (entsprechend einer binären Variablen Xi) wird durch das **Odds Ratio ausgedrückt.** Das ist eine gute Annäherung an das **relative Risiko** wenn die WahrscheinlichkeitenP $(Y = 1 \mid Xi = 1)$ und $P(Y = 1 \mid Xi = 0)$ sind klein und das ist es auch einfach als Exponentialfunktion berechnet Parameter, der der Variablen Xi *zugeordnet ist;* daher:

  odds ÿ *Verhältnis* des Faktors $M = \exp(ÿi)$.

BEISPIELE

Wir nehmen ein Beispiel von Hosmer und Lemeshow. Die abhängige Variable gibt das Vorliegen oder Fehlen einer Herzerkrankung bei 100 Probanden an, die unabhängige Variable

Variable ist das Alter.

Das resultierende logistische Regressionsmodell ist:

$$ÿ(X) = ÿ5{,}3 + 0{,}11 \cdot X$$

mit einer Standardabweichung für die Konstante und theÿ (Koeffizient der unabhängigen Variablen X) gleich 1,1337 bzw. 0,241.

Die beiden geschätzten Parameter sind statistisch signifikant.

Das Risiko einer Herzerkrankung steigt um exp (0,11) = 1,12 mit jedem Jahr.

Es ist auch möglich, die **Konfidenz** zu berechnen **Intervall** durch Nehmen der Exponentialfunktion des Intervall für ÿ:

$$(\exp(1{,}11 - 1{,}96 \cdot 0{,}241),\ \exp(1{,}11 + 1{,}96 \cdot 0{,}241)) = (1{,}065,\ 1{,}171).$$

Die folgende Tabelle gibt die Werte für an Log-Likelihood sowie die Parameterwerte, die durch Maximieren der Log-Likelihood-Funktion in einem iterativen Algorithmus geschätzt werden:

Log-Likelihood-Konstante			ÿ
1	*ÿ54,24*	*ÿ4,15*	0,872
2	*ÿ53,68*	*ÿ5,18*	1.083
3	*ÿ53,67*	*ÿ5,31*	1.109
4	*ÿ53,67*	*ÿ5,31*	1.109

WEITERLESEN

Binärdaten

Kontingenztabelle

Likelihood-Ratio-Test

Maximale Wahrscheinlichkeit

Modell

Quoten und Quotenverhältnis

Regressionsanalyse

VERWEISE

Bishop, YMM, Fienberg, SE, Holland, PW: Diskrete multivariate Analyse: Theorie und Praxis. MIT Press, Cambridge, MA (1975)

Cox, DR, Snell, EJ: Die Analyse binärer Daten. Chapman & Hall (1989)

Hosmer, DW, Lemeshow S.: Appliedlogistic Regression. Wiley-Reihe in der Wahrscheinlichkeit und Statistik. Wiley, New York (1989)

Altman, DG: Praktische Statistik für Med ische Forschung. Chapman & Hall, London (1991)

Machine Translated by Google

Lognormalverteilung

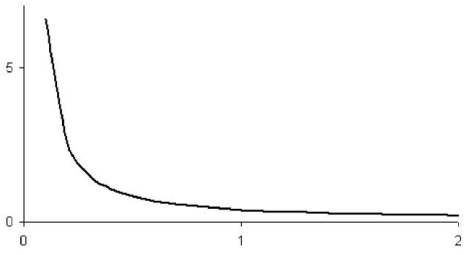
Eine **Zufallsvariable** *Y* folgt einer Lognormalverteilung, wenn die **Werte** von *Y* eine Funktion sind der Werte von *X* gemäß der Gleichung:

$$y = \exp(x),$$

wobei **die Zufallsvariable X** auf ein oder folgt **schlechte Verteilung.**

Die **Dichtefunktion** der Lognormalverteilung ist gegeben durch:

$$f(y) = \frac{1}{\ddot{y}y\,\ddot{y}2\ddot{y}} \exp\ddot{y}\,2\ddot{y}2 \frac{(\log y\,\ddot{y}\,\ddot{y})^{2}}{\cdot}.$$



Lognormalverteilung, ÿ = 0, ÿ = 1

Die Lognormalverteilung ist eine **kontinuierliche Wahrscheinlichkeitsverteilung.**

GESCHICHTE

Die logarithmische Normalverteilung geht auf die Arbeiten von **Galton, F.** (1879) und McAlister, D. (1879) zurück, die Ausdrücke für den **Mittelwert, den Median, den Modus, die Varianz** und bestimmte **Quantile** der resultierenden Verteilung erhielten. Galton, F. ging von *n* unabhängigen positiven **Zufallsvariablen** aus und konstruierte das Produkt und übergab es mit Hilfe des Logarithmus an eine Summe neuer Zufallsvariablen. Kapteyn, JC (1903) untersuchte mit van Uven, MJ die Konstruktion einer Zufallsvariablen nach einer logarithmischen Normalverteilung (1916). Sie entwickelten eine grafische Methode zur Schätzung von **Parametern.** Seitdem wurden viele Arbeiten zur Lognormalverteilung veröffentlicht. Eine erschöpfende Liste wird von Johnson, NL und Kotz, S. (1970) gegeben.

MATHEMATISCHE ASPEKTE

Der **Erwartungswert** der Lognormalverteilung ergibt sich aus:

$$E[Y] = \exp\ddot{y} + 2\ddot{y}2 \frac{1}{\cdot}.$$

Die **Varianz** ist gleich:

$$\mathrm{Var}(X) = \exp(2\ddot{y}) \cdot \exp 2\ddot{y}2\ \ddot{y}\ \exp\ddot{y}2.$$

Wenn die **Zufallsvariable** *Y* einer Lognormalverteilung folgt, ist die Zufallsvariable

$$X = \ln Y$$

folgt einer **Normalverteilung.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Die logarithmische Normalverteilung wird in der allgemeinen statistischen Praxis weitgehend angewendet. Beispielsweise folgen die kritische Proportionalität bei pharmazeutischen Anwendungen, die Dauer eines Arztbesuchs und die Verteilung der wirtschaftlichen Kräfte alle einer logarithmischen Normalverteilung. Die Lognormalverteilung kann als Näherung der **Normalverteilung verwendet werden.** Wenn ÿ kleiner wird, tendiert die Lognormalverteilung zur **Normalverteilung.**

WEITERLESEN

Kontinuierliche Wahrscheinlichkeitsverteilung
Normalverteilung

L

Machine Translated by Google

VERWEISE

Galton, F.: Das geometrische Mittel in der Vital- und Sozialstatistik. Proz. Roy. Soc. London. **29,** 365–367 (1879)

Johnson, NL, Kotz, S.: Verteilungen in der Statistik: Kontinuierliche univariate Verteilung

Butionen, Bde. 1 und 2. Wiley, New York (1970)

Kapteyn, J.: Skew-Häufigkeitskurven in Biologie und Statistik. Astronomisches Laboratorium Noordhoff, Groningen (1903)

Kapteyn, J., van Uven, MJ: Skew-Häufigkeitskurven in Biologie und Statistik. Gebrüder Hoitsema, Groningen (1916)

McAlister, D.: Das Gesetz des geometrischen Mittels. Proz. Roy. Soc. London. **29,** 367–376 (1879)

Längsdaten

Längsschnittdaten sind Daten, die über einen bestimmten Zeitraum hinweg zu Themen erhoben werden. Dies entspricht mehreren Beobachtungen pro Proband, da dieselben Einheiten bei mehr als einer Gelegenheit wiederholt gemessen werden. Wiederholte Messungen in der Biologie und Medizin sowie **Panelstudien** in den Sozial- und Wirtschaftswissenschaften sind charakteristische Beispiele für Längsschnittdatenstudien.

VERLAUF

Sehen Sie sich den Verlauf der **Paneldaten an.**

MATHEMATISCHE ASPEKTE

Eine sehr allgemeine Art, das lineare Modell in der Längsschnittdatenanalyse zu schreiben, ist die folgende:

$$yi = Xiÿ + \mathbf{Wiÿi} + ÿi ,$$

Dabei ist *yi* der *ni* × 1-Spaltenantwortvektor für das Subjekt i, *Xi* ist ein *ni* × *b* -Design

Die Matrix besteht normalerweise aus dummy-codierten Variablen, **ÿ** ist der *ni* × *b* -Regressionskoeffizientenvektor, Wi ist eine b×1-Designmatrix für Zufallseffekte, *ÿi* und *ÿi* sind der Subjektfehler

Bedingungen.

Das obige allgemeine Modell ist ein gemischtes Modell, da es sowohl feste **(ÿ)** als auch zufällige Effekte *(ÿi) umfasst.*

Die Parameterschätzung erfolgt häufig mittels **Maximum-Likelihood-Schätzung** mit

Die Wahrscheinlichkeitsfunktion hängt stark von den Annahmen zur Kovarianzstruktur von *ÿi* und *ÿi ab.* Es ist einfach, statistische Tests für die Schätzung zu erstellen

Parameter, um Konfidenzintervalle zu erhalten und die Vorhersage zu messen. Modellauswahltechniken werden auch zum Testen verschachtelter Modelle verwendet.

BEISPIEL

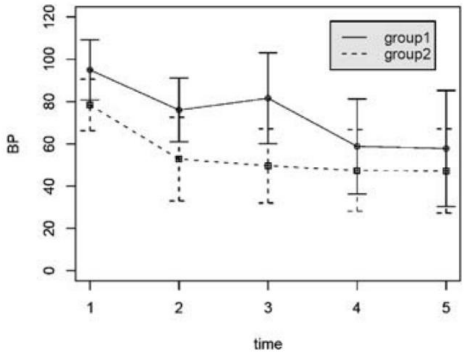
Zwei Gruppen von Menschen erhalten zwei unterschiedliche Blutdruckbehandlungen und wurden über einen Zeitraum von 5 Monaten beobachtet. Jede Gruppe besteht aus fünf Personen. Die interessierenden Maße sind in der folgenden Tabelle aufgeführt:

	Monat (Januar bis Mai)
	105 68 95 86 93
	100 92 98 74 35
1. Gruppe	102 85 95 62 82
	70 81 72 35 39
	98 34 34 37 40
	83 25 31 42 38
	69 44 41 30 32
2. Gruppe	84 64 77 80 82
	63 54 55 47 44
	93 77 44 38 40

Jede Zeile entspricht Beobachtungen, die im Laufe der Zeit für dieselbe Person gemacht wurden.

Unten sehen Sie die Abbildung, in der die Gruppenmittelkurven ±1 Standardfehler dargestellt sind.

Machine Translated by Google



Der einfachste Ansatz wäre in diesem Fall sein, die zufälligen Effekte zu ignorieren (**ÿi** = 0) und ein einfaches Modell mit festen zu konstruieren nur Wirkung, das heißt, die Gruppe, in der ein Subjekt existiert gehört. In einem solchen Fall reduziert sich das Modell auf

$$yij = \ddot{y}j + \ddot{y}ij ,$$

wobei *j* die Gruppe angibt und eine einfache Dummy-Variable ist. Wenn es einen erheblichen gibt Unterschied zwischen Individuen, dann zufällig Der Effekt *ÿi* wird dem obigen Modell hinzugefügt.

DOMÄNEN UND EINSCHRÄNKUNGEN

Längsschnittdaten werden hauptsächlich in der Epidemiologie und medizinischen Statistik sowie in verwendet sozioökonomische Studien. Im ersteren Fall Die Längsschnittdaten stellen wiederholte Messungen für verschiedene Behandlungsgruppen dar **klinische Versuche.** In letzterem beachten wir die **Tafel** Studien, die normalerweise soziale und ökologische Aspekte messen nomische Indizes wie Einkommen, Konsum und Arbeitszufriedenheit in verschiedenen Zeiträumen Perioden.

Es gibt viele mögliche Komplikationen Längsschnittdatenanalyse. Erstens sind es Daten korreliert, da die gleichen Themen im Zeitverlauf gemessen werden. Dies muss berücksichtigt werden letztlich um eine ordnungsgemäße Bereitstellung zu gewährleisten Ergebnisse. Darüber hinaus treten häufig fehlende Daten auf bei einer solchen Datenanalyse, da nicht alle Probanden während der gesamten Zeit gemessen werden Zeitraum. Letzteres wird beispielsweise bei Überlebensanalysen und Folgestudien problematischer, bei denen die Probanden (in der Regel die …) Patienten) sterben aus oder brechen den Prozess ab und brechen ihn ab die Behandlungen.

WEITERLESEN

Panel

VERWEISE

Twisk, JWR: Angewandte Längsschnittdaten Analyse für Epidemiologie. Ein praktischer Leitfaden. Cambridge University Press, Cambridge (2003)

Fitzmaurice, G., Laird, N., Ware, J.: Applied Längsschnittanalyse. Wiley, New York (2004)

Molenberghs, G., Geert, V.: Modelle für diskrete Längsschnittdatenreihen. Springer Reihe in Statistik. Springer, Berlin Heidelberg New York (2005)

Machine Translated by Google

Mahalanobis-Entfernung

Die Mahalanobis-Distanz ist die Distanz von *X* zur Größe **ÿ,** definiert als:

$$d2\ M(X,\ddot{y}) = (\mathbf{X}\ \ddot{y}\ \ddot{y})^{T}\ \ddot{y}1\ (\mathbf{X}\ \ddot{y}\ \ddot{y})\ .$$

Dieser Abstand basiert auf der Korrela zwischen Variablen oder der Varianz–Kovarianzmatrix.

Darin unterscheidet sie sich von der euklidischen Distanz es berücksichtigt die Korrelation der

Datensatz und hängt nicht von der Messskala ab. Die Mahalanobis-Entfernung beträgt

Wird häufig in **der Clusteranalyse** und anderen **Klassifizierungsmethoden** verwendet .

GESCHICHTE

Die Mahalanobis-Distanz wurde eingeführt von Mahalanobis, PC im Jahr 1936.

MATHEMATISCHE ASPEKTE

Wenn *ÿi E(Xi)* bezeichnet , dann ist per Definition der erwartete Wert von **X** = *(X1,...,Xp)* Vektor:

$$\mathbf{E[X]} = \ddot{\mathbf{y}} = \begin{pmatrix} \ddot{y}1 \\ \ddot{y}2 \\ . \\ . \\ . \\ \ddot{y}\ \ddot{y}p\ \ddot{y} \end{pmatrix}.$$

Die Varianz-Kovarianz-Matrix **ÿ** von **X** ist definiert von:

$$\ddot{\mathbf{y}} = \begin{pmatrix} \ddot{y}2 & \mathrm{cov}(X1,\ X2) & \ldots & \mathrm{cov}(X1,\ Xp) & \ddot{y}2 \\ \ddot{y} & & & & \ddot{y} \\ & 2 & & & \\ \vdots & & \ldots & \vdots & \\ \ddot{y} & & & \ddot{y}2_{P} & \ddot{y} \end{pmatrix}$$

$$= \mathbf{E[XX\ ]}\ \ddot{y}\ \ddot{y}\ddot{y}\ .$$

Es handelt sich um eine quadratisch symmetrische Matrix der Ordnung p.

Wenn die *Xi*- Variablen standardisiert sind, dann **ÿ** ist identisch mit der Korrelationsmatrix:

$$\begin{pmatrix} 1 & \ddot{y}12 & \ddot{y}13 & \ldots & \ddot{y}1p \\ \ddot{y} & & & & \ddot{y} \\ & \ddot{y}23 & \ldots & \ddot{y}2p & \\ & 1 & & & \\ & & \ldots & \ddot{y}3p & \\ & & \ldots & \vdots & \\ & & & 1 & \\ \ddot{y} & & & & \ddot{y} \end{pmatrix}.$$

Dann ist ÿÿ1/2. **Y** = ÿÿ1/2(X ÿ ÿ) ist ein Ran dom standardisierter Vektor mit Nichtkorrelat ed-Komponenten. Die Zufallsvariable **(X** ÿ ÿ) **ÿÿ1(Xÿÿ)** = D2 hat den Erwartungswert p. Tatsächlich ist D2 = wo die Y_{i}^{2} sind i=1 von Zufallsvariablen mit Mittelwert 0 und 1. *D* ist wird Mahalanobis-Abstand von **X** nach ÿ genannt.

Machine Translated by Google

WEITERLESEN

Einstufung

Clusteranalyse

Korrespondenzanalyse

Kovarianzanalyse

Distanz

VERWEISE

Mahalanobis, PC: Über Tests und Messungen der Gruppendivergenz. Teil I. Theoretische Formeln. J. Asiatic Soc. Bengalen **26,** 541–588 (1930)

Mahalanobis, PC: Zur verallgemeinerten Distanz in der Statistik, Proc. natn. Inst. Wissenschaft. Indien 2, 49–55 (1936)

Mahalanobis, Prasanta Chandra

Prasanta Chandra Mahalanobis (1893–1972) studierte bis 1908 in Kalkutta. 1912 erhielt er seinen B.Sc. in Physik am Presidential College von Cal Cutta. 1915 ging er nach England, um am King's College of Cambridge einen Tripos (B.Sc. in Cambridge) in Mathematik und Physik zu erwerben. Bevor er mit seinen Forschungen begann, kehrte er im Urlaub nach Kalkutta zurück, kehrte aber wegen des Krieges nicht nach England zurück.

Als Statistiker nutzte Mahalanobis Statistiken nicht als Selbstzweck, sondern vor allem als Instrument zum Verständnis und zur Interpretation wissenschaftlicher Daten sowie zur Entscheidungsfindung hinsichtlich des Wohlergehens der Gesellschaft. Er verwendete Statistiken zunächst in seinem Studium der Anthropologie (um 1917) und dann über Überschwemmungen. 1930 veröffentlichte er einen Artikel über die D2 -Statistik mit dem Titel „Test and Measures of Group Divergence". Sein größter Beitrag zur Statistik ist sicherlich die groß angelegte Volkszählung. Er hat in diesem Bereich viele Beiträge geleistet, beispielsweise die Opti Malauswahl der Stichprobe unter Verwendung von Varianz- und Kostenfunktionen. Er wurde zum Präsidenten der *Unterkommission der Vereinten Nationen für Sta* 1947 übernahm er *die technische Probenahme* , eine Position, die er bis 1951 innehatte.

Bemerkenswert ist auch seine Freundschaft mit **Ronald Aylmer Fisher,** die von 1926 bis zu Fishers Tod andauerte.

Hauptwerke und Artikel von Prasanta Chandra Mahalanobis:

1930 Über Tests und Messungen der Gruppendivergenz. Teil I. Theoretische Formeln. Zeitschrift der Asiatic Society of Bengal, 26 S. 541–588.

1936 Über die verallgemeinerte Distanz in der Statistik, Proc. Natn. Inst. Wissenschaft. Indien 2, 49–55.

1940 Eine Stichprobenerhebung der Juteanbaufläche in Bengalen. Sankhya 4, 511–530.

1944 Über groß angelegte Stichprobenerhebungen. Phi los. Trans. Roy. Soc. London. Ser. B 231329-451.

1963 Der Ansatz der Betriebsforschung zur Planung in Indien. Asia Publishing House, Bombay.

WEITERLESEN

Mahalanobis-Entfernung

Haupteffekt

In einem **faktoriellen Experiment** ist ein Haupteffekt der Beitrag einer **Faktorstufe ,** die für die Antwortvariable spezifisch ist .

MATHEMATISCHE ASPEKTE

Das lineare Modell , das nur die Haupteffekte von Faktoren berücksichtigt und die **Wechselwirkungen** zwischen Faktoren ignoriert, wird als Modell bezeichnet

Machine Translated by Google

additives Modell. Betrachten Sie das additive **Modell** eines Doppelklassifizierungsdesigns:

$$Yij = \ddot{y} + \ddot{y}i + \ddot{y}j + \ddot{y}ij ,$$
$$i = 1, 2,...,a \text{ und } j = 1, 2,...,b,$$

Dabei ist *Yij* die Antwort, die die **Behandlung** *ij erhält,* und ÿ ist der **Mittelwert, der** allen gemeinsam ist Behandlungen, ÿi ist die Hauptwirkung der i- **ten Stufe** von **Faktor** A, *ÿj* ist die Hauptwirkung der j-ten Stufe Niveau des Faktors B und *ij* ist das experimentelle Niveau **Fehler** der **Beobachtung** *Yij.*

Wenn $a = b = 2$, dann sind die Haupteffekte von FaktorenAundB, bezeichnetdaundb, bzw. sind gegeben durch:

$$a = \frac{(\ddot{y}1 \ddot{y} \ddot{y}0) (\ddot{y}1 + \ddot{y}0)}{} ,$$

$$b = 2 \frac{2 (\ddot{y}1 + \ddot{y}0) (\ddot{y}1 \ddot{y} \ddot{y}0)}{} .$$

WEITERLESEN

- **Varianzanalyse**
- **Einschätzung**
- **Faktor**
- **Faktorielles Experiment**
- **Hypothesentest**
- **Interaktion**

VERWEISE

Box, GEP, Hunter, WG, Hunter, JS: Statistiken für Experimentatoren. Eine Einführung in Design, Datenanalyse und Modellbau. Wiley, New York (1978)

Mann-Whitney-Test

Der Mann-Whitney-Test **ist nichtparametrisch** Test, der darauf abzielt, die Gleichheit zweier Bevölkerungsgruppen zu testen. Es wird verwendet, wenn wir zwei Proben haben, die stammen aus zwei Populationen.

GESCHICHTE

Der Mann-Whitney-Test wurde erstmals eingeführt für den Fall $m = n$ von **Wilcoxon, Frank** (1945). **Anschließend wurde der Wilcoxon-Test** von White, C. (1952) auf Proben mit unterschiedlichen Abmessungen ausgeweitet.

Beachten Sie, dass ein Test gleichwertig mit denen des Der Wilcoxon-Test wurde unabhängig entwickelt und von Festinger, L. (1946) eingeführt.

Mann, HB und Whitney, DR (1947) waren der erste, der die Proben verschiedener berücksichtigt Abmessungen und die Tabellen angepasst bereitzustellen bis hin zu kleinen Proben.

MATHEMATISCHE ASPEKTE

Sei (X1, *X2,...,Xn)* eine **Stichprobe** der Dimension *n ,* die aus einer **Grundgesamtheit** 1 stammt, und sei (Y1, *Y2,...,Ym)* sei eine Stichprobe der Dimension *m* aus einer Bevölkerung stammend 2.

Wir erhalten N = n+m Beobachtungen, die wir klassifizieren in aufsteigender Reihenfolge, wobei jedes Mal die Stichprobe notiert wird, zu der eine bestimmte Beobachtung gehört.

Die Mann- und Whitney **-Statistik,** bezeichnet mit U ist definiert als die Gesamtzahl, wie oft dies der Fall ist Ein *Xi* steht in der **Klassifikation** in vor einem *Yj* Aufsteigende Reihenfolge der Noobservations.When die Größe der Probe groß genug ist, Die Entschlossenheit von U wird lang, aber wir kann die folgende Beziehung verwenden, die identische Ergebnisse liefert:

$$U = mn + \frac{n (n + 1)}{2} \ddot{y} T ,$$

wobei *T* die Summe der dem zugeschriebenen Ränge ist *X* aus dem **Wilcoxon-Test.**

Hypothesen

Die entsprechenden Hypothesen Der Mann-Whitney-Test kann wie folgt formuliert werden: Folgt, je nach Ein-Schwanz oder Zwei-Schwanz Fall:

Machine Translated by Google

A: Zweiseitiger Fall:

$$H0 : P(X < Y) = \frac{1}{2}$$
$$H1 : P(X < Y) = \frac{1}{2}.$$

B: Einseitiger Fall:

$$H0 : P(X < Y) \ddot{y} \frac{1}{2}$$
$$H1 : P(X < Y) > \frac{1}{2}.$$

C: Einseitiger Fall:

$$H0 : P(X < Y) \ddot{y} \frac{1}{2}$$
$$H1 : P(X < Y) < \frac{1}{2}.$$

Im Fall A entspricht die **Nullhypothese** zu der Situation, in der es keinen Unterschied gibt zwischen den Populationen. Im Fall B die Null Hypothese bedeutet, dass Bevölkerung 1 größer ist als Grundgesamtheit 2. Im Fall C deutet die Nullhypothese darauf hin, dass Grundgesamtheit 1 kleiner ist als Bevölkerung 2.

Entscheidungsregeln

Wenn m, n < 12, vergleichen wir die Statistik U mit der **Wert** gefunden in der Mann-Whitney-Tabelle zu H0 testen . Wenn andererseits *m, n* ÿ 12, dann Die **Stichprobenverteilung** von *U* nähert sich sehr schnell der Normalverteilung an

Mittelwert $\ddot{y} = \frac{m \cdot n}{2}$ und **Standardabweichung**

$\ddot{y} = \frac{mn\,(m + n + 1)}{12}$.

So können wir die *Bedeutung* des bestimmen beobachteter Wert von *U* durch:

$$Z = \frac{U \ddot{y} \ddot{y}}{\ddot{y}},$$

wobei *Z* eine zufällige standardisierte Variable ist.

Mit anderen Worten, wir können den Wert von vergleichen *Z* ergibt sich so mit der **Normaltabelle.**

Darüber hinaus sind die Entscheidungsregeln unterschiedlich und hängen von den aufgestellten Hypothesen ab. In diesem So haben wir die Regeln A, B und C relativ zum frühere Fälle A, B und C.

Entscheidungsregel A

Wir lehnen die Nullhypothese H0 auf dem **Signifikanzniveau** ÿ ab, wenn *U* kleiner als der Wert ist in der Mann-Whitney-Tabelle mit Parametern n, m, und $\frac{\ddot{y}}{2}$ bezeichnet mit $tn,m,1\ddot{y}\frac{\ddot{y}}{2}$, oder wenn *U* ist größer als der Wert im Mann-Whitney-Modell Tabelle für n, m und 1ÿ ist, $\frac{\ddot{y}}{2}$, bezeichnet $tn,m,\frac{\ddot{y}}{2}$, Das wenn

$$tn,m,1\ddot{y}\frac{\ddot{y}}{2} < U < tn,m,\ddot{y}\frac{\ddot{y}}{2}.$$

Entscheidungsregel B

Wir lehnen die Nullhypothese H0 auf dem Signifikanzniveau ÿ ab, wenn *U* kleiner als der Wert ist in der Mann-Whitney-Tabelle mit Parametern n, m und ÿ, bezeichnet mit *tn,m,ÿ,* das heißt, wenn

$$U < tn,m,\ddot{y}.$$

Entscheidungsregel C

Wir lehnen die Nullhypothese H0 auf dem Signifikanzniveau ÿ ab, wenn *U* größer als der Wert ist in der Mann-Whitney-Tabelle mit Parametern n, m und 1 ÿ ÿ, bezeichnet mit *tn,m,1ÿÿ,* d. h.

Wenn

$$U > tn,m,1\ddot{y}\ddot{y}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Da zwischen den eine lineare Beziehung besteht **Wilcoxon-Test** und Mann-Whitney-Test *(U* = mn+n(n+1)/2ÿT), diese beiden Tests sind gleichwertig.

Daher haben sie zwei gemeinsame Eigenschaften, und beide müssen respektiert werden:

1. Bei beiden Stichproben handelt es sich um Stichproben aus ihren jeweiligen **Populationen.**
2. Zusätzlich zur **Unabhängigkeit** im Inneren Bei jeder Probe besteht eine gegenseitige Unabhängigkeit zwischen den Proben.
3. Die Messskala beträgt mindestens ordi nal.

Machine Translated by Google

BEISPIELE

In einer Klasse zählen wir neun Schüler: fünf Jungen und vier Mädchen. Die Klasse macht einen mentalen Test Berechnung, um zu sehen, ob die Jungen im Allgemeinen besser als die Mädchen in diesem Bereich.
Die Daten sind in der folgenden Tabelle aufgeführt, wobei die höchste Punktzahl dem besten Testergebnis entspricht.

Jungen (Xi)	Mädchen (Yi)
9.2	11.1
11.3	15.6
19.8	20.3
21.1	23.6
24.3	

Wir ordnen die Punktzahl in aufsteigender Reihenfolge und notieren sie jedes Mal **die Probe ,** zu der eine **bestimmte Beobachtung** gehört.

Bewertung	Stichprobenrang	Anzahl der vorangegangenen Male X Y
9,2 *X*	1	4
11,1 *J*	2	–
11,3 *X*	3	3
15,6 *J*	4	–
19,8 *X*	5	2
20,3 *J*	6	–
21,1 *X*	7	1
23,6 *J*	8	–
24,3 *X*	9	0

Wir finden $U = 4 + 3 + 2 + 1 = 10$. Im

auf die gleiche Weise, wenn wir die Beziehung $U = mn + \frac{n(n+1)}{2} - T$ verwenden, dann gilt:

$$T = \sum_{i=1}^{N} R(Xi) = 1 + 3 + 5 + 7 + 9 = 25$$

und $U = 5 \cdot 4 + \frac{5 \cdot 6}{2} - 25 = 10$.

Die zu prüfenden Hypothesen lauten wie folgt:

H0: Jungen sind im Allgemeinen nicht besser als Mädchen im Kopfrechnen.

H1: Jungen sind im Allgemeinen besser als Mädchen mentale Berechnung.

Die Nullhypothese kann wie folgt umformuliert werden:
H0: P(X < Y) ≥ 1/2 und entspricht
zu Fall C. Die Entscheidungsregel lautet dann wie folgt:

Lehnen Sie H0 auf dem **Signifikanzniveau** α ab, wenn

$$U > tn,m,1-\alpha ,$$

wobei $tn,m,1-\alpha$ der **Wert** im Mann– ist
Whitney-Tabelle mit den Parametern n, m und 1α
α. Denken Sie daran, dass in unserem Fall $m = 4$ und $n = 5$ ist.
Da wir eine **Beziehung** zwischen den **Wilcox** haben
stattdessen den **On-Test** und den Mann-Whitney-Test
Um die *U*- Tabelle zu verwenden, verwenden wir die **Wilcoxon-Methode**
Tabelle für T. Wenn wir α = 0,05 wählen, dann ist die
Der Wert der Wilcoxon-Tabelle beträgt 32. As
$T = 25 < 32$, wir lehnen H0 nicht ab und wir
kommen zu dem Schluss, dass Jungen im Allgemeinen nicht besser sind
als Mädchen im Kopfrechnen.

WEITERLESEN

- **Hypothesentest**
- **Nichtparametrischer Test**
- **Wilcoxon-Test**

REFERENZ

Festinger, L.: Die Bedeutung der Differenz zwischen Mitteln ohne Bezug auf
die Häufigkeitsverteilungsfunktion. Psy Chometrika **11,** 97–105 (1946)

Mann, HB, Whitney, DR: Auf einem Test ob eine von zwei Zufallsvariablen ist
stochastisch größer als die anderen. Ann. Mathematik. Stat. **18,** 50–60 (1947)

White, C.: Die Verwendung von Rängen ist ein Signifikanztest für den Vergleich zweier Behandlungen. Biometrik 8, 33–41 (1952)

Wilcoxon, F.: Einzelvergleiche von Ranking-Methoden. Biometrie 1, 80–83 (1945)

Machine Translated by Google

Randdichtefunktion

Im Fall eines **Paars von Zufallsvariablen** *(X,* Y), wenn **die Zufallsvariable** *X* (oder Y) für sich betrachtet wird, wird ihre **Dichtefunktion** als Grenzdichtefunktion bezeichnet.

GESCHICHTE

Siehe **Wahrscheinlichkeit.**

MATHEMATISCHE ASPEKTE

Im Fall eines **Paares kontinuierlicher Zufallsvariablen** *(X,* Y) mit einer **gemeinsamen Dichtefunktion** *f(x,* y) erhält man die **Randdichtefunktion** von *X* durch Integration:

$$P(X \le a, y_1 < Y < y_2) = \int_{y_1}^{y_2} \int_{-\infty}^{a} f(x, y)\,dy\,dx = \int_{-\infty}^{a} f_X(x)\,dx ,$$

wobei $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$ ist die Dichte von X.

Die Randdichtefunktion **von** Y *wird* auf die gleiche Weise erhalten:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx .$$

WEITERLESEN

- **Dichtefunktion**
- **Funktion der Gelenkdichte**
- **Paar Zufallsvariablen**

VERWEISE

Hogg, RV, Craig, AT: Einführung in die Mathematische Statistik, 2. Aufl. Macmillan, New York (1965)

Ross, SM: Einführung in Wahrscheinlichkeitsmodelle, 8. Aufl. John Wiley, New York (2006)

Randverteilungsfunktion

Im Fall eines **Paars von Zufallsvariablen** *(X,* Y) wird **die Verteilungsfunktion ,** wenn **die Zufallsvariable** *X* (oder Y) allein betrachtet wird, als Randverteilungsfunktion bezeichnet.

MATHEMATISCHE ASPEKTE

Betrachten Sie **ein Paar Zufallsvariablen (X,** Y) und ihre **gemeinsame Verteilungsfunktion:**

$$F(a, b) = P(X \le a, Y \le b) .$$

Die Randverteilungsfunktion von *X* ist gleich:

$$\begin{aligned} F_X(a) &= P(X \le a) \\ &= P(X \le a, Y < \infty) \\ &= P[\lim_{b \to \infty} (X \le a, Y \le b)] \\ &= \lim_{b \to \infty} P(X \le a, Y \le b) \\ &= \lim_{b \to \infty} F(a, b) \\ &= F(a, \infty) . \end{aligned}$$

In ähnlicher Weise ist die Randverteilungsfunktion von *Y* gleich:

$$F_Y(b) = F(\infty, b) .$$

WEITERLESEN

- **Verteilungsfunktion**
- **Gemeinsame Verteilungsfunktion**
- **Paar Zufallsvariablen**

Markow, Andrei Andrejewitsch

Markov, Andrei Andreevich wurde 1856 in Rjasan geboren und starb 1922 in St. Petersburg. Dieser Schüler **Tschebytschews, Pafnutii** Lwowitsch, leistete wichtige Beiträge dazu

Machine Translated by Google

Bereich der Wahrscheinlichkeitsrechnung. Er lieferte einen überzeugenden Beweis des **zentralen Grenzwertsatzes.** Durch seine Arbeiten zu Markov-Ketten entwickelte sich das Konzept der Markov-Abhängigkeit zur modernen Theorie und Anwendung zufälliger Prozesse. Der Globalismus seiner Arbeit hat die Entwicklung der Wahrscheinlichkeits- und internationalen Statistik beeinflusst. Er studierte Mathematik an der Universität St. Petersburg, wo er seinen Bachelor-Abschluss und 1878 seinen Doktortitel erhielt. Anschließend lehrte er und setzte die Kurse von Tchebychev, Pafnutii Lvovich fort, nachdem Tcheby chev, PL die Universität verlassen hatte. 1886 wurde er zum Mitglied der School of Mathe gewählt matics, gegründet von Tchebychev, PL an der St. Petersburger Akademie der Wissenschaften. Er wurde 1886 Vollmitglied der Akademie und schied 1905 aus der Universität aus, wo er weiterhin lehrte.

Markov und Liapunov waren berühmte Schüler von Tchebychev, PL und strebten stets danach, die Wahrscheinlichkeitsrechnung als exakte und praktische mathematische Wissenschaft zu etablieren.

Das erste Erscheinen von Markov-Ketten in Markovs Werk erfolgte in der Izvestiia (Bulletin) der Physikalisch-Mathematischen Gesellschaft der Kasaner Universität. Markov hat den Beweis des zentralen Grenzwertsatzes abgeschlossen, den **Tchebychev** begonnen, aber nicht abgeschlossen hat. Er ging dabei mit der Methode der Momente vor und der Beweis wurde in der dritten Auflage *von Ischislenie Veroiatnostei (Wahrscheinlichkeitsrechnung)* veröffentlicht.

Hauptwerk von Markov, Andrei Andree vich:

1899 Die Anwendung der Funktionen wird in der Wahrscheinlichkeitsrechnung fortgesetzt, Kasan. Stier. 9(2), 29–34.

WEITERLESEN

Zentraler Grenzwertsatz
Gauß-Markov-Theorem

VERWEISE

Markov, AA: Die Anwendung von Funktionen setzt sich in der Wahrscheinlichkeitsrechnung fort. Kasan. Stier. 9(2), 29–34 (1899)

Matrix

Eine rechteckige Tabelle mit *m* Zeilen und *n* Spalten wird als (m×n) -Matrix bezeichnet . Die in der Tabelle vorkommenden Elemente werden Koeffizienten oder Komponenten der Matrix genannt.

MATHEMATISCHE ASPEKTE

Eine Matrix kann wie folgt dargestellt werden:

$$\mathbf{A} = \begin{pmatrix} a11 & a12 & \dots & a1n \\ a21 & a22 & \dots & a2n \\ \vdots & \vdots & \dots & \vdots \\ am1 & am2 & \dots & amn \end{pmatrix}.$$

Ein Element wird durch seinen **Wert** und seine Position charakterisiert. Daher ist *aij* ein Element der Matrix *A* , wobei *i* die Zeilennummer angibt und *j* die Spaltennummer. Eine Matrix mit *m* Zeilen und *n* Spalten hat die Ordnung (m×n). Es gibt verschiedene Arten von Matrizen: • Quadratische Matrix der Ordnung *n* Eine Matrix heißt quadratische Matrix der Ordnung *n* , wenn sie so viele Zeilen wie Spalten hat (bezeichnet mit n). • Nullmatrix Die Matrix, auf der sich alle Elemente befinden null wird als Nullmatrix bezeichnet. Es wird sein bezeichnet mit O. • Identitätsmatrix Für jede quadratische Matrix der Ordnung n die Elemente, die den gleichen Index für die Linie haben

M

Machine Translated by Google

Was die Säule betrifft [d. h. die Elemente in Position *(i,* i)] bilden die Hauptdiagonale.

Eine **Identitätsmatrix** ist eine quadratische Matrix, die auf der Hauptdiagonale „1“ und überall sonst „0“ hat. Es wird mit *I* oder manchmal mit *In bezeichnet,* wobei *n* die Ordnung der Identitätsmatrix ist. •

Symmetrische Matrix Eine quadratische Matrix , A ist symmetrisch, wenn A = *A* , wobei *A* die **Transponierte** von Matrix A ist. • Antisymmetrische Matrix Eine quadratische *ÿA* , Matrix A ist antisymmetrisch , wenn A = wobei *A* die **Transponierte** von Matrix A ist ;

Diese Bedingung führt dazu, dass die Elemente auf der Hauptdiagonale *aij* für *i* = *j* null sind. • Skalarmatrix

Eine quadratische MatrixA wird als Skalarmatrix bezeichnet, wenn

$$aij = ÿ \text{ für } i = j \text{ , } aij = 0$$

$$\text{für } i = j \text{ ,}$$

wobei ÿ ein Skalar ist.

• Diagonalmatrix Eine quadratische Matrix *A* heißt Diagonalmatrix, wenn *aij* = 0 für *i* = *j* und *aij* eine beliebige Zahl ist. •

Dreiecksmatrix

Eine quadratische Matrix *A* heißt obere Dreiecksmatrix, wenn *aij* = 0 für *i* > j, oder untere Dreiecksmatrix, wenn *aij* = 0 für *i* < j.

Operationen auf Matrizen

• Addition von Matrizen

Wenn A = (aij) und B = (bij) zwei Matrizen der Ordnung *(m* × n) sind, dann ist ihre Summe *A* + *B* durch die Matrix *C* = *(cij)* der Ordnung (m×n) definiert , deren Summe jedes Element ist entsprechende Elemente von *A* und B:

$$\mathbf{C} = cij = aij + bij \text{ .}$$

• Multiplikation einer Matrix mit einem Skalar

Wenn ÿ ein Skalar und *A* eine Matrix ist, ist das Produkt *ÿA* die Matrix derselben Ordnung wie *A* erhalten durch Multiplikation jedes Elements *aij* von *A* mit dem Skalar ÿ:

$$\mathbf{ÿA} = Aÿ = ÿ \cdot aij \text{ .}$$

• Multiplikation von Matrizen

Seien *A* und *B* zwei Matrizen; Das Produkt *AB* ist genau dann definiert, wenn die Anzahl der Spalten von *A* gleich der Anzahl der Zeilen von B ist. Wenn *A* die Ordnung *(m* × n) und *B* die Ordnung (n × q) hat, dann ist das Produkt *AB* definiert durch die Matrix *C* der Ordnung *(m* × q), deren Elemente durch Berechnung erhalten werden:

$$cij = \sum_{k=1}^{N} aik \cdot bkj$$

für *i* = *1,...,m* und *j* = *1,...,q.*

Das bedeutet, dass die Elemente der ith-Linie von *A* werden mit den entsprechenden Elementen der j-ten Spalte von B multipliziert und die Ergebnisse addiert.

• Ganzzahlige Potenz quadratischer Matrizen

Sei *A* eine quadratische Matrix; Wenn *p* eine positive ganze Zahl ist, ist die p-te Potenz von *A* definiert durch:

$$Ap = A \cdot A \cdot \quad ... \quad \cdot A \; p \text{ mal} \qquad .$$

Normalerweise setzen wir für *p* = 0 A0 = I, wobei *I* die **Identitätsmatrix** derselben Ordnung wie A ist .

• Quadratwurzel einer Diagonalmatrix

Sei *A* eine Diagonalmatrix nichtnegativer Koeffizienten. Wir bezeichnen mit $\overline{ÿA}$ die Quadratwurzel von A, definiert durch die Diagonalmatrix derselben Ordnung wie A, in der wir jedes Element durch seine Quadratwurzel ersetzen:

$$ÿ\overline{\mathbf{A}} = \begin{pmatrix} ÿ\overline{a11} & 0 & \dots & 0 \\ 0 & ÿ\overline{a22} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & ÿ\overline{ann} \end{pmatrix} .$$

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Verwendung von Matrizen oder allgemein von Matrizenrechnungen wird häufig in der Regressionsanalyse, der **Varianzanalyse** und in vielen anderen Bereichen **der Statistik verwendet.** Es ermöglicht die Beschreibung Es vereinfacht ein Gleichungssystem auf prägnanteste Weise und erleichtert die Notation erheblich. Aus den gleichen Gründen tauchen die Matrizen in **Datenanalysen** auf , bei denen wir im Allgemeinen auf große Datentabellen stoßen.

BEISPIELE

Seien **A** und **B** zwei quadratische Matrizen der Ordnung 2:

$$\mathbf{A} = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \quad \text{und } \mathbf{B} = \begin{matrix} 6 & 2 \\ -2 & 1 \end{matrix} .$$

Die Summe zweier Matrizen ist definiert durch:

$$\mathbf{A} + \mathbf{B} = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} + \begin{matrix} 6 & 2 \\ -2 & 1 \end{matrix}$$

$$= \begin{matrix} 1+6 & 2+2 \\ 1-2 & 3+1 \end{matrix}$$

$$= \begin{matrix} 7 & 4 \\ -1 & 4 \end{matrix} .$$

Die Multiplikation der Matrix **A** mit 3 ergibt:

$$3 \cdot \mathbf{A} = 3 \cdot \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} = \begin{matrix} 3 \cdot 1 & 3 \cdot 2 \\ 3 \cdot 1 & 3 \cdot 3 \end{matrix}$$

$$= \begin{matrix} 3 & 6 \\ 3 & 9 \end{matrix} .$$

Das Matrixprodukt von **A** und **B** ist gleich:

$$\mathbf{A} \cdot \mathbf{B} = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \cdot \begin{matrix} 6 & 2 \\ -2 & 1 \end{matrix}$$

$$= \begin{matrix} 1 \cdot 6 + 2 \cdot (-2) & 1 \cdot 2 + 2 \cdot 1 \\ 1 \cdot 6 + 3 \cdot (-2) & 1 \cdot 2 + 3 \cdot 1 \end{matrix}$$

$$\mathbf{A} \cdot \mathbf{B} = \begin{matrix} 2 & 4 \\ 0 & 5 \end{matrix} .$$

Anmerkung: **A** · **B** und **B** · **A** sind im Allgemeinen nicht gleich, was die Nichtkommutativität der Matrixmultiplikation verdeutlicht.

Das Quadrat der Matrix *A* erhält man durch Multiplikation von *A* mit A. Somit:

$$A2 = \mathbf{A} \cdot \mathbf{A} = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \cdot \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix}$$

$$= \begin{matrix} 1 \cdot 2 + 2 \cdot & 1 \cdot 2 + 2 \cdot 3 \\ 1 \cdot 2 + 3 \cdot 1 & 1 \cdot 2 + 3 \cdot 3 \end{matrix}$$

$$A2 = \begin{matrix} 3 & 8 \\ 4 & 11 \end{matrix} .$$

Sei **D** die folgende Diagonalmatrix der Ordnung 2:

$$\mathbf{D} = \begin{matrix} 9 & 0 \\ 0 & 2 \end{matrix} .$$

Wir definieren die Quadratwurzel der Quadratmatrix D, bezeichnet als $\sqrt{\mathbf{D}}$, durch:

$$\sqrt{\mathbf{D}} = \begin{matrix} \sqrt{9} & 0 \\ 0 & \sqrt{2} \end{matrix} = \begin{matrix} 3 & 0 \\ 0 & \sqrt{2} . \end{matrix}$$

M

WEITERLESEN

Korrespondenzanalyse
Datenanalyse
Bestimmend
Hutmatrix
Trägheitsmatrix
Inverse Matrix
Multiple lineare Regression
Regressionsanalyse
Verfolgen
Transponieren

VERWEISE

Balestra, P.: Calcul matriciel pour économistes. Editions Castella, Albeuve, Schweiz (1972)

Machine Translated by Google

Graybill, FA: Theorie und Anwendungen von
das lineare Modell. Duxbury, North Scit uate, MA
(Waldsworth und Brooks/Cole,
Pacific Grove, CA) (1976)

Searle, SR: Matrixalgebra Nützlich für
Statistiken. Wiley, New York (1982)

Maximale Wahrscheinlichkeit

Der Begriff maximale Wahrscheinlichkeit bezieht sich auf eine Methode zur Schätzung von Parametern einer **Population** anhand einer Zufallsstichprobe . Es wird angewendet wenn wir die allgemeine Form der Verteilung der Bevölkerung kennen, aber wenn ein oder Weitere Parameter dieser Verteilung sind Unbekannt. Die Methode besteht im Wählen ein Schätzer unbekannter Parameter, dessen Werte maximieren die **Wahrscheinlichkeit ,** die beobachtete Probe zu erhalten.

GESCHICHTE

Im Allgemeinen die Maximum-Likelihood-Methode wird Fisher, Ronald Aylmer zugeschrieben . Die Methode lässt sich jedoch zurückverfolgen zu den Werken der Wissenschaftler des 18. Jahrhunderts Lambert, JH und Bernoulli, D. Fisher, RA führte die Methode der Maximum Likeli ein Hood in seinen ersten statistischen Veröffentlichungen in 1912.

MATHEMATISCHE ASPEKTE

Betrachten Sie eine **Stichprobe** von *n* Zufallsvariablen X1, *X2,...,Xn,* wobei jede Zufallsvariable gemäß einer **Wahrscheinlichkeitsverteilung** verteilt wird, die eine **Dichtefunktion** ergibt *f (x,* ÿ), mit ÿ, wobei der Parameterraum ist. Die **Gelenkdichtefunktion** von X1, *X2,...,Xn* ist

$$f\,(\text{x1},\ \ddot{y})\cdot f\,(\text{x2},\ \ddot{y})\cdot \quad \cdots \quad \cdot f\,(xn,\ \ddot{y})\ .$$

Für festes x ist diese Funktion eine Funktion von ÿ und heißt Wahrscheinlichkeitsfunktion L:

$$L\ (\ddot{y};\ \text{x1},\ x2,\ldots,xn) = f\,(\text{x1},\ \ddot{y})\cdot f\,(\text{x2},\ \ddot{y}) \cdot \cdots \cdot f\,(xn,\ \ddot{y}).$$

Ein **Schätzer** ist die maximale Wahrscheinlichkeit Schätzer von ÿ, wenn er die Wahrscheinlichkeit maximiert Haubenfunktion. Beachten Sie, dass diese Funktion *L* kann maximiert werden, indem der erste Wert auf Null gesetzt wird partielle Ableitung nach ÿ und Auflösen der so erhaltene Gleichungen. Außerdem, wann Jede Funktion *L* und *lnL* wird maximiert Für jeden Wert von ÿ ist dies häufig möglich beschließen:

$$\frac{\ddot{y}\ \ln L\ (\ddot{y};\ \text{x1},\ x2,\ldots,xn)}{\ddot{y}\ddot{y}} = 0\ .$$

Schließlich sind die Maximum-Likelihood-Schätzer dieser Methode nicht immer eindeutig liefert nicht immer unvoreingenommene Schätzungen tors.

BEISPIELE

Sei X1, X2,...,Xn eine **Zufallsstichprobe ,** die aus der Normalverteilung mit dem stammt **Mittelwert** ÿ und die **Varianz** ÿ2. Die Methode der **Schätzung** der maximalen Wahrscheinlichkeit con besteht darin, als Schätzer des Unbekannten auszuwählen Parameter ÿ und ÿ2 sind die Werte, die die **Wahrscheinlichkeit** maximieren , dass das beobachtete Ereignis auftritt Probe.

Dadurch wird die Wahrscheinlichkeitsfunktion maximiert beinhaltet Folgendes:

$$L\ \ddot{y},\ \ddot{y}2;\ \text{x1}\ \ldots,\text{xn} = \frac{1}{\ddot{y}\ \ddot{y}2\ddot{y}^{\frac{N}{N}}}\ \exp\ \ddot{y}^{\frac{\ddot{y}}{\ddot{y}}}\ \frac{\sum_{i=1}^{N}(xi\ \ddot{y}\ \ddot{y})^2}{2\ddot{y}2}\ \ddot{y}^{\frac{\ddot{y}}{\ddot{y}}},$$

Machine Translated by Google

oder, was dasselbe ist, sein natürlicher Logarithmus:

$$\ln L = \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}^{N}} \cdot \exp\left(-\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{2\sigma^2} \right) \right)$$

$$= \ln \left(2\pi\sigma^2 \right)^{-\frac{n}{2}} + \ln \left(\exp\left(-\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{2\sigma^2} \right) \right)$$

$$= -\frac{n}{2} \cdot \ln (2\pi\sigma^2) - n \cdot \ln(\sigma) - \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{2\sigma^2} .$$

Wir setzen die partiellen Ableitungen gleich Null $\frac{\partial \ln L}{\partial \mu}$ Und $\frac{\partial \ln L}{\partial \sigma}$:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{N} (x_i - \mu) = 0 \quad , \qquad (1)$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \cdot \sum_{i=1}^{N} (x_i - \mu)^2 \qquad (2)$$

$$= 0 .$$

Aus (1) leiten wir ab:

$$\frac{1}{\sigma^2} \cdot \sum_{i=1}^{N} (x_i - \mu) = 0 \quad ,$$

$$\sum_{i=1}^{N} (x_i - \mu) = 0 \quad ,$$

$$\sum_{i=1}^{N} x_i - n \cdot \mu = 0 \quad ,$$

$$\hat{\mu} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i = \bar{x} .$$

Somit ist **für die Normalverteilung das arithmetische Mittel** der Maximum-Likelihood-Schätzwert mator.

Aus (2) erhalten wir:

$$-\frac{N}{\sigma} + \frac{1}{\sigma^3} \cdot \sum_{i=1}^{N} (x_i - \mu)^2 = 0 \quad ,$$

$$\frac{1}{\sigma^3} \cdot \sum_{i=1}^{N} (x_i - \mu)^2 = \frac{N}{\sigma} ,$$

$$2\hat{\sigma} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^2 .$$

Wir finden in diesem Fall die maximale Wahrscheinlichkeit Schätzer die empirische Varianz der Stichprobe.

Beachten Sie, dass die Maximum-Likelihood-Schätzung tion **für die Bedeutung** ist unvoreingenommen und diejenigen von Die Varianz ist verzerrt (stattdessen durch *n* geteilt). von *n* ÿ 1). Wir stellen fest, dass für die Normalverteilung der Mittelwert und die Varianz der Stichprobe die Maximum-Likelihood-Schätzer sind des Mittelwerts und der Varianz der **Grundgesamtheit.** Dies gilt nicht unbedingt für andere Distributionen: Betrachten Sie zum Beispiel den Fall der **Gleichverteilung** im Intervall *[a,* b], wobei *a* und *b* unbekannt sind und wo Der zu schätzende Parameter ist der Mittelwert von die Verteilung ÿ = *(a* + b) /2. Wir können überprüfen, dass der Schätzer der maximalen Likelihood von ÿ gleich ist:

$$\hat{\mu} = \frac{x1 + xn}{2} ,$$

wobei x1 und *xn* die kleinsten sind und die größten **Beobachtungen** der Stichprobe (x1, *x2,...,xn)* . Dieser Schätzer ist der Mittelpunkt des Intervalls mit dem Grenzen x1 und *xn.* Wir sehen, dass in diesem Fall die

M

Machine Translated by Google

Der Maximum-Likelihood-Schätzer ist unterschiedlich aus dem Mittelwert der Stichprobe:

$$\ddot{y}\hat{} = \bar{x} .$$

WEITERLESEN

- Einschätzung
- Schätzer
- Likelihood-Ratio-Test
- Robuste Schätzung

REFERENZEN

Fisher, RA: Über ein absolutes Kriterium für die Anpassung von Frequenzkurven. Durcheinander. Mathematik. **41,** 155–160 (1912)

DOMÄNEN UND EINSCHRÄNKUNGEN

Empirisch liefert uns die Bestimmung des **arithmetischen Mittels** einer Menge von Beobachtungen ein Maß für den zentralen **Wert** dieser Menge.

Dieses Maß gibt uns eine ***Schätzung*** des Mittelwerts ÿ der **Zufallsvariablen** *X* , aus der die Beobachtungen stammen.

WEITERLESEN

- Arithmetisches Mittel
- Erwarteter Wert
- Maß für die zentrale Tendenz
- Maß der Kurtosis
- Maß für die Schiefe
- Gewichtetes arithmetisches Mittel

Bedeuten

Der Mittelwert einer **Zufallsvariablen** ist ein **zentraler Wert** **Tendenzmaß** dieser Variablen. Er wird auch **Erwartungswert genannt.** In der Praxis wird der Begriff Mittelwert häufig im Sinne des **arithmetischen Mittels verwendet.**

GESCHICHTE

Siehe **arithmetisches Mittel.**

MATHEMATISCHE ASPEKTE

Sei *X* eine kontinuierliche **Zufallsvariable** deren **Dichtefunktion** *f* (x) ist . Der Mittelwert ÿ von X, falls vorhanden, ist gegeben durch:

$$\ddot{y} = x \cdot f(x)\, dx .$$

Der Mittelwert einer diskreten Zufallsvariablen *X* von **Wahrscheinlichkeitsfunktion** *P* (*X* = x), falls vorhanden, ist gegeben durch:

$$\ddot{y} = x \cdot P(X = x) .$$

Mittlere absolute Abweichung

Die mittlere Abweichung einer Reihe quantitativer **Beobachtungen** ist ein **Maß für die Streuung,** das den **Mittelwert** der absoluten Werte der **Abweichungen** jeder Beobachtung angibt

in Bezug auf ein **Maß für die Position** der Verteilung.

GESCHICHTE

Siehe L1- **Schätzung.**

MATHEMATISCHE ASPEKTE

Betrachten Sie x1, *x2,...,xn* als eine Menge von *n* Größen oder von *n* **Beobachtungen,** die sich auf eine **quantitative Variable** X *beziehen* . Gemäß der Definition wird die mittlere Abweichung wie folgt berechnet:

$$EM = \frac{\sum_{i=1}^{N} |xi \ddot{y} \bar{x}|}{N} ,$$

wobei x¯ das **arithmetische** Mittel des Beobachters ist vationen. Wenn die **Beobachtungen** geordnet sind

Machine Translated by Google

In einer **Häufigkeitsverteilung** wird die mittlere Abweichung wie folgt berechnet:

$$EM = \frac{\sum_{i=1}^{k} fi \cdot |xi \ddot{y} \bar{x}|}{\sum_{i=1}^{k} fi},$$

Wo

xi repräsentiert die verschiedenen **Werte** der **Variable,**

fi sind die damit verbundenen **Frequenzen** Werte und

k ist die Anzahl der verschiedenen **Werte.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Die mittlere absolute Abweichung ist ein Maß der Streuung robuster als der Standard Abweichung.

BEISPIELE

Fünf Studenten haben nacheinander zwei Prüfungen bestanden Dafür erhielten sie Folgendes Noten:

Prüfung 1:

3,544,5 *3,5 4,5* $\bar{x} =$ $\frac{20}{5} = 4$

Prüfung 2:

2,5 5,5 3,5 4,5 4 $\bar{x} = = 4$ $\frac{20}{5}$

Das **arithmetische Mittel** der Noten ist bei beiden Prüfungen identisch. Die Variabilität der Noten (bzw. die Streuung) ist jedoch nicht gegeben das gleiche.

Für Prüfung 1 beträgt die mittlere Abweichung:

Bedeuten Abweichung $= \frac{|3,5 \ddot{y} 4|}{5} + \frac{|4 \ddot{y} 4|}{5} + \frac{|4,5 \ddot{y} 4|}{5} + \frac{|3,5 \ddot{y} 4|}{5} + \frac{|4,5 \ddot{y} 4|}{5}$

$= \frac{2}{5} = 0,4 .$

Bei Prüfung 2 ist die mittlere Abweichung sehr hoch ähm:

Bedeuten Abweichung $= \frac{|2,5 \ddot{y} 4|}{5} + \frac{|5,5 \ddot{y} 4|}{5} + \frac{5\,|3,5 \ddot{y} 4|}{5} + \frac{|4,5 \ddot{y} 4|}{5} + \frac{5\,|4 \ddot{y} 4|}{5}$

$= \frac{4}{5} = 0,8 .$

Die Noten der zweiten Prüfung sind daher stärker **um das arithmetische Mittel** verteilt als die Noten der ersten Prüfung.

WEITERLESEN

L1 -Schätzung
Maß der Streuung
Standardabweichung
Varianz

M

Mittlere quadratische Fehler

Sei $\ddot{y}\hat{}$ ein **Schätzer** des **Parameters** $\ddot{y}$.

Der quadratische **Fehler** bei der Schätzung von $\ddot{y}$ durch $\ddot{y}\hat{}$ beträgt $\ddot{y} \ddot{y} \ddot{y}\hat{}^{2}$.

Der mittlere quadratische Fehler *(MSE) ist* der **mathematische Erwartungswert** dieses Werts:

$$MSE(\ddot{y})\hat{} = E\, \ddot{y} \ddot{y} \ddot{y}\hat{}^{2} .$$

Wir können beweisen, dass der mittlere quadratische **Fehler** entspricht der Summe der **Varianz** des Schätzers und des Quadrats des **Bias:**

$$MSE(\ddot{y})\hat{} = \mathrm{Var}\, \ddot{y}\hat{} + E\, \ddot{y}\hat{} \ddot{y} \ddot{y}^{2} .$$

Wenn also der betrachtete Schätzer mit ist Aus Bias entspricht der mittlere quadratische Fehler genau auf die Varianz des Schätzers.

Machine Translated by Google

GESCHICHTE

Im frühen 19. Jahrhundert verglichen **de Laplace, Pierre Simon** (1820) und **Gauss, Carl Friedrich** (1821) die **Schätzung** einer unbekannten Größe, basierend auf Beobachtungen mit zufälligen Fehlern, mit einem Zufallsspiel. Sie zogen eine Parallele zwischen dem Fehler des geschätzten **Werts** und dem dadurch verursachten Verlust so ein Spiel. Gauß und Carl Friedrich schlugen ein Maß für die Ungenauigkeit vor, das Fehlerquadrat. Er erkannte, dass seine Wahl zufällig war, begründete sie jedoch mit der mathematischen Einfachheit der Funktion „Quadrat bilden“.

MATHEMATISCHE

ASPEKTE Seien X1 , _ _ _ _ _ _ _ _

ÿ ÿ ÿˆ 2 .

Der mittlere quadratische **Fehler** ist definiert durch:

$$MSE\ \ddot{y}\hat{} = E\ \ddot{y}\ \ddot{y}\ \ddot{y}\hat{}^{2} .$$

Wir können Folgendes beweisen:

$$MSE\ \ddot{y}\hat{} = \text{Var}\ \ddot{y}\hat{} + E\ \ddot{y}\hat{}\ \ddot{y}\ \ddot{y}^{2} ,$$

Das heißt, der mittlere quadratische Fehler entspricht der Summe der **Varianz** des Schätzers und des Quadrat der **Vorspannung.**

In der Tat:

$$MSE(\ddot{y}\ )\hat{} = E\ \ddot{y}\ \ddot{y}\ \ddot{y}\hat{}^{2}$$

$$= E\ \ddot{y}\ \ddot{y}\ E\ \ddot{y}\hat{} + E\ \ddot{y}\hat{}\ \ddot{y}\ \ddot{y}\hat{}^{2}$$

$$= \mathrm{E}[\ \ddot{y}\ \ddot{y}\ E\ \ddot{y}\hat{}^{2} + E\ \ddot{y}\hat{}\ \ddot{y}\ \ddot{y}\hat{}^{2}$$

$$+ 2 \cdot \ddot{y}\ \ddot{y}\ E\ \ddot{y}\hat{} \cdot E\ \ddot{y}\hat{}\ \ddot{y}\ \ddot{y}\hat{}\ ]$$

$$= E\ \ddot{y}\ \ddot{y}\ E\ \ddot{y}\hat{}^{2} + EE\ \ddot{y}\hat{}\ \ddot{y}\ddot{y}\hat{}^{2}$$

$$+ 2 \cdot \ddot{y}\ \ddot{y}\ E\ \ddot{y}\hat{} \cdot (E\ \ddot{y}\hat{}\ \ddot{y}\ E\ \ddot{y}\hat{}\ )$$

$$= EE\ \ddot{y}\hat{}\ \ddot{y}\ \ddot{y}\hat{}^{2} + \ddot{y}\ \ddot{y}\ E\ \ddot{y}\hat{}^{2}$$

$$= \text{Var}\ \ddot{y}\hat{} + (Bias)^{2} ,$$

wobei die Vorspannung E[ÿˆ] ÿ ÿ ist.

Wenn also der Schätzer ÿˆ ohne Bias ist, also E[ÿˆ] = ÿ, erhalten wir:

$$MSE\ \ddot{y}\hat{} = \text{Var}\ \ddot{y}\hat{} .$$

Die Effizienz eines Schätzers ist umgekehrt proportional zu seinem mittleren quadratischen Fehler. Wir sagen, dass der Schätzer ÿˆ konsistent ist, wenn

$$\lim_{n\ddot{y}\ddot{y}} \mathrm{MSE}(\ddot{y})\hat{} = 0 .$$

In diesem Fall sind die Varianz und der Bias der Schätzer tendiert zu 0, wenn die Anzahl der Beobachter vations *n* tendiert gegen Unendlich.

DOMÄNEN UND EINSCHRÄNKUNGEN

Bestimmte Autoren bevorzugen die Verwendung der durch E[|ÿ ÿÿˆ|] definierten **mittleren** Absolutwertabweichung , wobei ÿˆ der **Schätzer** des **Parameters** ÿ Der Nachteil der letzten Definition liegt auf der Ebene der mathematischen Eigenschaften, da die Funktion des Absolutwerts nicht differenziert werden kann überall gefeiert.

BEISPIELE

Wir betrachten **eine multiple lineare Regression in** Matrixform:

$$\mathbf{Y} = \mathbf{X} \cdot \ddot{\mathbf{y}} + \ddot{\mathbf{y}} \quad ,$$

Wo

Y ist ein **Vektor** der Ordnung *n (n* Beobachtungen der **unabhängigen Variablen),**

X ist eine gegebene **Matrix** *n* × p, **ÿ** ist der Vektor der Ordnung *p* der Koeffizienten zu bestimmen, und

Machine Translated by Google

ÿ ist der Fehlervektor, von dem wir annehmen, dass er unabhängig ist und jeweils einer **Normalverteilung** mit **Mittelwert** 0 und **Varianz** ÿ2 folgt.

Der Vektor ÿˆ = **(X** · X)ÿ1 · **X** · **Y** ergibt einen **Schätzer** ohne **Bias** von ÿ.
Wir erhalten dann:

$$MSE\ \ddot{y}\hat{} = \mathrm{Var}\ \ddot{y}\hat{} \quad \cdot$$

WEITERLESEN

Voreingenommenheit

Schätzer

Erwarteter Wert
Varianz

VERWEISE

Gauss, CF: In: Gauss' Werk (1803–1826) über die Theorie der kleinsten Quadrate. Trans. Trotter, HF Statistical Techniques Research Group, Technischer Bericht Nr. 5. Princeton University Press, Princeton, NJ (1821)

Laplace, PS de: Théorie analytique des probabilités, 3. Aufl. Courcier, Paris (1820)

Maß der zentralen Tendenz

Ein Maß für die zentrale Tendenz ist eine Statistik, die einen Satz relativer Daten zusammenfasst auf eine **quantitative Variable.** Genauer gesagt ermöglicht es die Bestimmung eines festen **Werts,** eines sogenannten Zentralwerts, um den sich die Datenmenge tendenziell gruppieren kann. Die Hauptmaße zentraler Tendenzen sind:

- **Arithmetisches Mittel**
- **Median**
- **Modus**

MATHEMATISCHE ASPEKTE

Der **Mittelwert** ÿ, **der Median** Md und **der Modus** Mo sind Maße für das Verteilungszentrum.

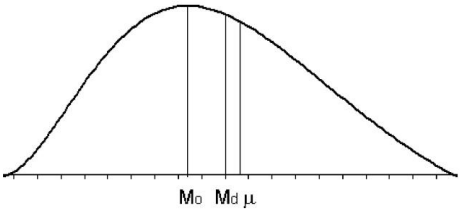
Bei einer symmetrischen Verteilung fallen Mittelwert, Median und Modus in einem gemeinsamen **Wert zusammen.** Dieser Wert gibt das Symmetriezentrum der Verteilung an. Für eine symmetrische Unimodalverteilung gilt:

$$\ddot{y} = \mathrm{Md} = \mathrm{Mo}\ .$$

In diesem Fall kann eine **Schätzung** des Verteilungszentrums durch den Mittelwert, den Median oder den Modus erfolgen.
Bei einer asymmetrischen Verteilung haben die verschiedenen Maße der zentralen Tendenz unterschiedliche Werte. (Ihr Vergleich kann als **Maß für die Schiefe verwendet werden.)**
Für eine unimodale Verteilung gestreckt zum

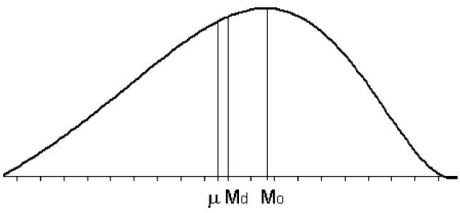
Richtig, wir haben im Allgemeinen:

ÿ ÿ Md ÿ Mo .



Diese Beziehung wird umgekehrt, wenn die Verteilung erfolgt Bution ist unimodal und nach links gestreckt:

ÿ ÿ Md ÿ Mo .



Für eine unimodale, mäßig asymmetrische Verteilung wurde dies empirisch bewiesen

M

Machine Translated by Google

Der Modus Mo, der Median Md und der Mittelwert ȳ erfüllen häufig näherungsweise die folgende Beziehung:

$$\text{Mo} \ddot{y} \ddot{y} = 3 \cdot (Md \ddot{y} \ddot{y}) .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

In der Praxis hängt die Wahl des Maßes, das es ermöglicht, das Zentrum einer Reihe von Beobachtungen am besten zu charakterisieren, von den spezifischen Merkmalen der statistischen Studien und der Ergebnisse ab Informationen, die wir erhalten möchten.

Vergleichen wir das **arithmetische Mittel,** den **Modus** und den **Median** nach verschiedenen Kriterien:

1. Das *arithmetische Mittel:*
   - Hängt vom **Wert** aller Beobachtungen ab tionen.
   - Ist einfach zu interpretieren. • Ist die bekannteste und am häufigsten verwendete messen.
   - Wird häufig als **Schätzer** für den **Mittelwert** der **Grundgesamtheit verwendet.** • Hat einen Wert, der durch die Ausreißer verfälscht werden kann.
   - Darüber hinaus ist die Summe der quadratischen Abweichungen jeder Beobachtung *xi* eines Datensatzes und ein Wert ȳ minimal, wenn ȳ entspricht dem arithmetischen Mittel:

$$\text{Mindest}_{\ddot{y}} \sum_{i=1}^{N} (xi \ddot{y} \ddot{y})2$$

ȳ ȳ = Arithmetik. bedeuten

2. Der *Median:*
   - Ist leicht zu bestimmen, da nur eine Datenklassifizierung erforderlich ist.
   - Ist leicht zu verstehen, wird aber seltener verwendet das arithmetische Mittel.
   - Wird nicht durch Ausreißer beeinflusst, was ihm einen Vorteil gegenüber der Arithmetik verschafft metrisches Mittel, wenn die Reihen wirklich Ausreißer haben.
   - Wird als **Schätzer** des Zentralwerts verwendet Merkmale einer Verteilung, insbesondere wenn sie asymmetrisch ist oder Ausreißer aufweist.
   - Die Summe der quadratischen Abweichungen des Absolutwerts zwischen jeder Beobachtung *xi* eines Datensatzes und einem Wert ȳ ist minimal wenn ȳ gleich dem Median ist:

$$\text{Mindest}_{\ddot{y}} \sum_{i=1}^{N} /xi \ddot{y} \ddot{y}| \ddot{y} \ddot{y} = \text{Md} .$$

3. Der *Modus:*
   - Hat praktisches Interesse, da es sich um den am häufigsten dargestellten Wert einer Menge handelt. • Ist auf jeden Fall eine selten verwendete Maßnahme. • Hat einen Wert, der kaum von Ausreißern beeinflusst wird.
   - Hat einen Wert, der stark von den Schwankungen einer **Stichprobe beeinflusst wird.** Es kann von **Probe** zu Probe stark variieren .

Darüber hinaus kann ein Datensatz viele (oder keine) Modi enthalten.

WEITERLESEN

Arithmetisches Mittel

Erwarteter Wert

Bedeuten

Maß für die Schiefe

Median

Modus

Gewichtetes arithmetisches Mittel

REFERENZEN

Stuart, A., Ord, JK: Kendall'sadvancedthe ory of Statistics.Vol. I. Verteilungstheorie. Wiley, New York (1943)

Dodge, Y.: Premiers pas en statistique. Springer, Paris (1999)

Machine Translated by Google

Maß der Streuung

Ein Maß für die Streuung ermöglicht die Beschreibung ein Datensatz zu einer bestimmten **Variablen,** der einen Hinweis auf die Variabilität gibt der **Werte** innerhalb des Datensatzes. Die Mea Streusicherheit rundet die Beschreibung ab gegeben durch das **Maß der zentralen Tendenz** einer Verteilung.

Wenn wir unterschiedliche Verteilungen beobachten, können wir das Sagen Sie, dass für einige von ihnen alle Daten vorliegen gruppiert in einer mehr oder weniger **kurzen Entfernung von** der zentrale Wert; für andere ist die Entfernung viel besser.

Wir können die Streuungsmaße in einteilen zwei Gruppen:

1. Maße, die durch den Abstand zwischen definiert werden zwei repräsentative Werte der Verteilung:
   - **Bereich,** auch Vari- **Intervall** genannt ation
   - **Interquartilbereich**
2. Maße, die in Abhängigkeit von den Abweichungen jedes Datums von einem zentralen Wert berechnet werden:
   - **Geometrische Abweichung**
   - **Mittlere Abweichung**
   - **Mittlere absolute Abweichung**
   - **Standardabweichung**

Unter den Streuungsmaßen am meisten Wichtig und am häufigsten verwendet wird die Standardabweichung.

DOMÄNEN UND EINSCHRÄNKUNGEN

Mit Ausnahme des beschreibenden Aspekts, der es einem ermöglicht, eine Verteilung zu charakterisieren oder Um viele Verteilungen zu vergleichen, kann das Wissen über die Streuung oder die Variabilität einer Verteilung von erheblichem praktischem Interesse sein. Was würde passieren wenn unsere Entscheidungen auf dem **Mittelwert** basieren würden allein.

- Straßen würden gebaut, um nur zu gelangen Der durchschnittliche Verkehr und die Staus an Feiertagen wären unermesslich.
- Häuser würden in solchen gebaut werden Weise, der mittleren Kraft des Windes zu widerstehen mit den daraus resultierenden Konsequenzen im Falle eines starken Sturms.

Diese Beispiele zeigen, dass in bestimmten Fällen ein **Maß an zentraler Tendenz** nicht ausreicht, um eine verlässliche Entscheidung treffen zu können. Darüber hinaus ist das Maß von Die Streuung ist ein Hinweis auf den Charakter des zentralen Tendenzmaßes. Der kleiner die Variabilität einer gegebenen Menge, die mehr der Wert des Maßes der Zentrale Die Tendenz wird für die Daten repräsentativ sein Satz.

WEITERLESEN

Variationskoeffizient

Interquartilbereich
Mittlere absolute Abweichung

Maß für die zentrale Tendenz

Bereich
Standardabweichung
Varianz

VERWEISE

Dodge, Y.: Premiers pas en statistique. Springer, Paris (1999)

Maß der Unähnlichkeit

Das Maß der Unähnlichkeit ist eine Distanz Maß, das einem Objektpaar einen reellen Wert zuweist positive (oder null) Zahl.

Nehmen wir zum Beispiel an, dass jedes Objekt wird durch *n* Variablen charakterisiert, ein Maß der Unähnlichkeit zwischen zwei Objekten würde bestehen darin, die Anzahl der verschiedenen anzugeben Punkte, die zwei betrachtete Objekte darstellen.

Machine Translated by Google

Diese Zahl ist eine ganze Zahl zwischen 0

und *n* und ist gleich *n* , wenn zwei Objekte keinen gemeinsamen Punkt haben.

MATHEMATISCHE ASPEKTE

Sei eine **Population ,** für die wir ein Maß für die Unähnlichkeit *s* definieren möchten .
Wir nennen ein Maß der Unähnlichkeit *s* jede Abbildung von × auf die reellen positiven (oder Null-)Zahlen, die die folgenden zwei Eigenschaften erfüllt: • Für jedesw,w in ,s *(w, w)*
= *s (w ,* w);

das heißt, *s* ist symmetrisch.

• Für jedes *w* in *s (w,* w) ≠ 0 ist das Maß der Unähnlichkeit zwischen einem Objekt und sich selbst Null.

DOMÄNEN UND EINSCHRÄNKUNGEN

Als Maß für die Unähnlichkeit *einer* endlichen **Grundgesamtheit** ist eine begrenzte Anwendung möglich, d. h.:

$$\max sw, w ; w, w \ddot{y} = m < \ddot{y} \quad .$$

Wir können ein **Maß für die Ähnlichkeit** der Population definieren durch:

$$sw, w = m \ddot{y} sw, w .$$

Diese beiden Begriffe, Maß der Ähnlichkeit und Maß der Unähnlichkeit, können als komplementär betrachtet werden.
Wir verwenden das Maß der Unähnlichkeit, um die Probleme der **Klassifikation zu lösen,** anstelle einer **Distanz ,** die ein Sonderfall davon ist und die sogenannte Eigenschaft der Dreiecksungleichung aufweist.

BEISPIELE

Wir betrachten die **Population** als bestehend aus drei Blumen

$$= \{w1, w2, w3\}$$

an dem die folgenden drei Variablen gemessen wurden:

• v1: Farbe der Blüte • v2:
Anzahl der Blütenblätter • v3:
Anzahl der Blätter wo

v1 wird hier als **dichotome Variable** zusammengefasst , da wir nur Rot berücksichtigen (R) und gelbe (Y) Blüten und

v2 und v3 sind diskrete Zufallsvariablen, deren Werte in Kardinalzahlen angegeben sind.

Wir haben beobachtet:

Variablen

	v_1	v_2	v_3
w_1	R	5	2
w_2	und	4	2
w_3		R 5	1

Wir ermitteln für jedes Blumenpaar die Anzahl der Punkte, die es unterscheiden. Wir sehen, dass wir ein gewisses Maß an Unähnlichkeit für die Gruppe der drei Blumen haben:

$$s (w1, w1) = s (w2, w2) = s (w3, w3) = 0$$

weil eine Blume sich nicht von sich selbst unterscheidet;

$$s (w1, w2) = s (w2, w1) = 2$$

weil die Farbe und die Anzahl der Blütenblätter bei w1 und w2 unterschiedlich sind;

$$s (w1, w3) = s (w3, w1) = 1$$

da nur die Anzahl der Blätter es erlaubt, w1 von w3 zu unterscheiden ;

$$s (w2, w3) = s (w3, w2) = 3$$

asw2 undw3 unterscheiden sich in jedem der drei untersuchten Merkmale.

Mithilfe dieser möglichen Werte können wir ableiten, dass *s* die beiden Eigenschaften eines Maßes für Unähnlichkeit erfüllt.

Machine Translated by Google

WEITERLESEN

- **Einstufung**
- **Distanz**
- **Maß der Ähnlichkeit**
- **Bevölkerung**

VERWEISE

Celeux, G., Diday, E., Govaert, G., Lecheval lier, Y., Ralambondrainy, H.: Classifi cation automatique des données – Aspekte Statistik und Informatik. Dunod, Paris (1989)

Hand, DJ: Diskriminierung und Klassifizierung. Wiley, New York (1981)

Maß der Kurtosis

Die Maße der Kurtosis sind Teil davon Formmaße und charakterisieren einen Aspekt der Form einer gegebenen Verteilung. Genauer gesagt charakterisieren sie den Grad der Kurtosis der Verteilung in Richtung einer Normalverteilung Bution. Bestimmte Distributionen liegen in der Nähe von **Normalverteilung,** ohne vollständig zu sein identisch. Es ist sehr nützlich, testen zu können, ob Die Form der Verteilung stellt eine Abweichung von einer Kurtosis der Normalverteilung dar Bution.

Wir sprechen von einer platikurtischen Verteilung wenn Die Kurve ist flacher als normal Kurve und einer leptokurtikalen Verteilung, wenn es ist schärfer als die normale Kurve.

Um die Kurtosis einer Kurve zu testen, verwenden wir die **Kurtosis-Koeffizient.**

MATHEMATISCHE ASPEKTE

Gegeben sei eine Zufallsvariable *X* mit Wahrscheinlichkeit Verteilung *fx* ist der Kurtosis-Koeffizient definiert als:

$$ÿ2 = \frac{ÿ4}{ÿ4} \; ÿ\,3 \quad ,$$

wobei ÿ4 für das vierte Moment ungefähr steht Der Mittelwert und ÿ4 ist die quadrierte Varianz Probenkurtosis wird durch Verwendung von Probenanaloga erhalten, d. h.:

$$ÿ\hat{} = ÿ\,\frac{ÿ\hat{}4}{\hat{}ÿ4} - 3$$

WEITERLESEN

- **Kurtosis-Koeffizient**
- **Formmaß**
- **Maß für die Schiefe**

VERWEISE

Joanes, DN & Gill, CA: Vergleich der Messungen von Probenschiefe und Wölbung, JRSS.D 47(1), 183–189 (1998)

Standortmaß

Ein Maß für den Standort (oder die Position) ist eine Maßnahme, die vorschlägt, eine Reihe von zu synthetisieren statistische Daten durch nur einen festen **Wert,** Hervorhebung eines bestimmten Punktes der untersuchten **Variablen.** Die am häufigsten verwendeten Maße vo Position sind:

- Maße der zentralen Tendenz **(Mittelwert, Modus, Median) ,** die tendenziell bestimmen das Zentrum eines Datensatzes.
- Die Quantile (Quartil, **Dezil, Zentil),** die dazu neigen, nicht das Zentrum zu bestimmen, sondern eine bestimmte Position einer Teilmenge von Daten.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Standortmaße sowie die von Dispersion und Form werden insbesondere verwendet um viele Datensätze oder viele Häufigkeitsverteilungen zu vergleichen.

WEITERLESEN

- **Maß für die zentrale Tendenz**
- **Quantil**

M

Machine Translated by Google

Formmaß

Wir können versuchen, die Form einer **Häufigkeitsverteilung** mit Hilfe geeigneter Koeffizienten zu charakterisieren. Bestimmte Häufigkeitsverteilungen liegen nahe an der **Normalverteilung** („Glockenkurve"). Diese Verteilungen können auch eine Asymmetrie oder eine Kurtosis gegenüber der Normalkurve darstellen. Die beiden Formmaße Sind:

• **Maß der Asymmetrie.** • **Maß der Kurtosis.**

GESCHICHTE

Pearson, Karl (1894, 1895) war der erste, der die Unterschiede zwischen bestimmten Distri testete Butionen und die **Normalverteilung.** Wir haben gezeigt, dass es zu Abweichungen vom Normalen kommt Die Kurve kann durch Momente der Ordnung 3 und 4 einer Verteilung charakterisiert werden.

Vor 1890 entwickelten Gram, Jorgen Pedersen und Thiele sowie Thorvald Nicolai aus Dänemark eine Theorie über die Symmetrie von Frequenzkurven. Nachdem Pearson und Karl sein ausgeklügeltes und äußerst interessantes System veröffentlicht hatten (1894, 1895), wurden viele Artikel zu diesem Thema veröffentlicht.

WEITERLESEN

Maß der Streuung
Maß der Kurtosis
Standortmaß
Maß für die Schiefe

VERWEISE

Pearson, K.: Beiträge zur mathematischen Evolutionstheorie. I. In: Karl Pearsons frühe statistische Arbeiten. Cam Bridge University Press, Cambridge, S. 1–40 (1948). Erstveröffentlichung 1894 as: Zur Zerlegung asymmetrischer Frequenzkurven. Philos. Trans. Roy. Soc. London. Ser. A **185,** 71–110

Pearson, K.: Beiträge zur mathematischen Evolutionstheorie. II: Skew-Variation in homogenem Material. In: Karl Pearsons frühe statistische Arbeiten. Cambridge University Press, Cambridge, S. 41–112 (1948). Erstveröffentlichung 1895 in Philos. Trans. Roy. Soc. London. Ser. A **186,** 343–414

Maß der Ähnlichkeit

Das Ähnlichkeitsmaß ist ein Abstandsmaß, das einem Objektpaar der untersuchten **Population eine reelle positive (oder Null-) Zahl zuordnet.** Wenn wir beispielsweise annehmen, dass jedes Objekt durch *n* Variablen gekennzeichnet ist, würde ein Maß für die Ähnlichkeit zwischen zwei Objekten darin bestehen, die Anzahl der gemeinsamen Punkte zweier betrachteter Objekte anzugeben. Diese Zahl ist eine ganze Eins zwischen 0 und *n* und ist gleich *n,* wenn die beiden Objekte identisch sind.

MATHEMATISCHE ASPEKTE

Sei eine **Population ,** für die wir ein Ähnlichkeitsmaß s definieren möchten. Wir nennen ein Ähnlichkeitsmaß *s* jede Abbildung von × auf die reellen positiven (oder Null-)Zahlen, die Folgendes erfüllt: • Für jedes w, *w* in ,sw, *w* = *sw* , das heißt, *s* ist , *w* , symmetrisch; • Für jedes w, *w* in , mit *w* = *w*

s(w, w) = *sw* , *w* > *sw, w* ,

Das heißt, ein Objekt ist sich selbst „ähnlicher" als einem anderen Objekt.

DOMÄNEN UND EINSCHRÄNKUNGEN

Als Maß für die Ähnlichkeit *einer* **Bevölkerungszahl** ist eine begrenzte Anwendung möglich, d. h .:

Machine Translated by Google

max *sw, w ; w, w* ÿ =

$$s(w, \mathrm{w}) < \ddot{y} \quad ,$$

wobei *w* ein beliebiges Objekt von ist; wenn wir *m* = stellen *s(w,* w), dann können wir **ein Maß für die Unähnlichkeit** der Population definieren durch:

$$sw, w = m \ddot{y} sw, w .$$

Diese beiden Begriffe, Maß der Ähnlichkeit und Maß der Unähnlichkeit berücksichtigt werden kann komplementär.

Wir verwenden auch die Ähnlichkeitsmaße als Maß der Unähnlichkeit für **Klassifikationsprobleme .**

BEISPIELE

Die **Bevölkerung** sei die Form von drei Blumen

$$= \{w1, w2, w3\}$$

an dem wir die drei Variablen messen:

- v1: Farbe der Blume
- v2: Anzahl der Blütenblätter
- v3: Anzahl der Blätter

Wo

v1 wird hier als **dichotom** zusammengefasst **variabel ,** da wir nur Rot berücksichtigen (R) und gelbe (Y) Blüten und

v2 und v3 sind diskrete Zufallsvariablen, deren Werte in Kardinalzahlen angegeben sind.

Wir haben beobachtet:

Variablen

	v_{12} v	v_3
w_1	R 5	2
w_2	Ja 4	2
w_3	R 5	1

Wir ermitteln für jedes Blumenpaar die Anzahl der Punkte, die ihnen gemeinsam sind.

Wir sehen, dass wir ein gewisses Maß an Ähnlichkeit haben aus dem Set aus drei Blumen:

$$s(w1, \mathrm{w1}) = s(w2, \mathrm{w2}) = s(w3, \mathrm{w3}) = 3 \quad ,$$

was der Anzahl der Variablen entspricht;

$$s(w1, \mathrm{w2}) = s(w2, \mathrm{w1}) = 1$$

weil die Farbe und Anzahl der Blütenblätter sind unterschiedlich bei w1 und w2;

$$s(w1, \mathrm{w3}) = s(w3, \mathrm{w1}) = 2$$

da nur die Anzahl der Blätter es uns erlaubt, w1 von w3 zu unterscheiden ;

$$s(w2, \mathrm{w3}) = s(w3, \mathrm{w2}) = 0$$

da w2 und w3 keine gemeinsamen Punkte haben.

Mit Hilfe dieser möglichen Werte können wir

Daraus lässt sich schließen, dass *s* die beiden Eigenschaften von erfüllt ein Maß für die Ähnlichkeit.

WEITERLESEN

Distanz

Maß der Unähnlichkeit

Bevölkerung

VERWEISE

Siehe **Maß der Unähnlichkeit.**

Maß für die Schiefe

In einer symmetrischen Verteilung ist der **Median, arithmetisches Mittel** und **Modus** sind gleich zentraler Punkt. Dies gilt nicht, wenn die Verteilung schief ist. In diesem Fall ist der Modus getrennt vom arithmetischen Mittel, und die

Der Median liegt zwischen zwei davon. Im Umkehrschluss Daher ist es notwendig, die Maße der Schiefe zu entwickeln, um den Grad der Abweichung der Form der Verteilung von der zu untersuchen symmetrische Verteilung. Der Hauptversatz

ness-Maßnahmen sind:

- **Yule- und Kendall-Koeffizient**
- **Schiefekoeffizient**

Machine Translated by Google

MATHEMATISCHER ASPEKT

Gegeben sei eine Zufallsvariable *X* mit Wahrscheinlichkeit. Verteilung *fx,* der Schiefekoeffizient ist definiert als:

$$\ddot{y}1 = \frac{\ddot{y}3}{\ddot{y}3}$$

wobei ÿ3 für das dritte Moment ungefähr steht der Mittelwert und ÿ3 ist die Standabweichung.

GESCHICHTE

Siehe **Schiefekoeffizient.**

WEITERLESEN

Schiefekoeffizient
Maß der Kurtosis
Formmaß
Yule- und Kendall-Koeffizient

Median

Der Median ist ein Maß **für die zentrale Tendenz**

cy ist als der **Wert** definiert , der in der Mitte liegt einer Reihe geordneter Beobachtungen, wenn es in ist in aufsteigender oder absteigender Reihenfolge.

Für eine ZufallsvariableX ,derMedian, bezeichnet von Md, hat die Eigenschaft, dass:

$$P(X \ddot{y} \text{Md}) \ddot{y} \frac{1}{2} \text{ und } P(X \ddot{y} \text{Md}) \ddot{y} \frac{1}{2}.$$

Wir finden dann jeweils 50 % der Beobachtungen Seite des Medians:



Wenn wir eine ungerade Anzahl von Beobachtungen haben, der Median entspricht den Werten der mittlere Beobachtung. Wenn wir eine gerade Zahl haben In der Anzahl der Beobachtungen gibt es keine eindeutige Beobachtung in der Mitte. Der Median ergibt sich aus dem **arithmetischen Mittel** der Werte von zwei Beobachtungen der Mitte der Ordnung Beobachtungen.

GESCHICHTE

1748 Euler, Leonhard und Mayer, Johann Tobias schlug unabhängig eine Methode vor das besteht darin, die Beobachtungen von zu teilen einen Datensatz in zwei gleiche Teile.

MATHEMATISCHE ASPEKTE

Wenn die Beobachtungen einzeln angegeben werden, erfolgt die Berechnung des Medians einfach:

1. Ordnen Sie die *n* Beobachtungen in aufsteigender Reihenfolge an oder absteigender Reihenfolge.
2. Wenn die Anzahl der Beobachtungen ungerade ist:
   - Finden Sie die Beobachtung der Mitte $\frac{n+1}{2}$.
   - Der Median entspricht dem **Wert** von mittlere Beobachtung.
3. Wenn die Anzahl der Beobachtungen gerade ist:
   - Finden Sie die beiden Beobachtungen der Mitte dle $\frac{N}{2}$ Und $\frac{N}{2} + 1$.
   - Der Median entspricht der **Arithmetik Mittelwert** der Werte dieser beiden Beobachter vationen.

Wenn wir keine individuellen Beobachtungen haben tionen, aber sie sind in Klassen gruppiert, die Der Median wird wie folgt ermittelt:

1. Bestimmen Sie die Medianklasse (diejenige, die enthält die $\frac{N}{2}$-te Beobachtung).
2. Berechnen Sie den Median wie folgt Formel:

$$\text{Md} = L + \frac{\frac{N}{2} \ddot{y} \text{finf}}{fMd} \cdot c,$$

Wo

L ist die untere Grenze des Medians Klasse,

N ist die Gesamtzahl der Beobachtungen tionen,

Machine Translated by Google

finf ist die Summe der Frequenzen kleiner als die Mittelklasse,

fMd ist die Häufigkeit des Medians Klasse, und

c ist die Länge des **Intervalls** von mittlere Klasse.

Diese Formel basiert auf der **Hypothese** wonach die Beobachtungen sind gleichmäßig innerhalb jeder Klasse verteilt. Es geht außerdem davon aus, dass die Unter- und Obergrenzen der Medianklasse definiert sind und bekannt.

Eigenschaften

Die Summe der Abweichungen in absoluten Werten zwischen jeder Beobachtung *xi* eines Datensatzes und Ein Wert ÿ ist minimal, wenn ÿ gleich dem Medi ist ein:

$$\underset{\ddot{y}}{\text{minimieren}} \sum_{i=1}^{N} |xi \ddot{y} \ddot{y}| \ddot{y} \ddot{y} = Md .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Bei der Berechnung des **arithmetischen Mittels** Wir berücksichtigen den Wert **jedes Beobachters vation.** Somit hat ein **Ausreißer** einen starken Einfluss das arithmetische Mittel. Andererseits in Die Berechnung des Medians nehmen wir in Anspruch Berücksichtigen Sie nur den Rang der Beobachtungen. Daher haben die Ausreißer keinen Einfluss auf den Median.

Wenn wir eine stark schiefe **Häufigkeitsverteilung haben,** sind wir daran interessiert, den Median als **Maß für die zentrale Tendenz** zu wählen und nicht das arithmetische Mittel der Reihenfolge um die Wirkung des extremen Wertes zu neutralisieren ues.

BEISPIELE

Der Median einer Menge von fünf Zahlen 42, 25, 75, 40 und 20 ist 40, weil 40 die Mitte ist Nummer, wenn die Nummern bestellt werden:

20 25 40 42 75 .

Median

Wenn wir dieser Menge die Zahl 62 hinzufügen, ist die Anzahl der Beobachtungen gerade und der Median entspricht dem **arithmetischen Mittel** der beiden mittleren Beobachtungen:

20 25 40 42 62 75 .

$\frac{40+42}{2}$ =41=Median

Wir betrachten nun den Median der gruppierten Daten in Klassen. Nehmen wir zum Beispiel die folgende **Häufigkeitstabelle ,** die das darstellt Tagesumsatz von 50 Lebensmittelgeschäften:

Klasse (Einnahmen in Euro)	Häufigkeit fi (Anzahl der Lebensmittelgeschäfte)	Kumuliert Frequenz
500–550	3	3
550–600	12	15
600–650	17	32
650–700	8	40
700–750	6	46
750–800	4	50
Gesamt	50	

Die mittlere Klasse ist die Klasse 600–650 as es enthält die mittlere Beobachtung (den 25.). Überwachung).

Unter der Annahme, dass die Beobachtungen in jeder Klasse gleichmäßig verteilt sind, ergibt sich der Median entspricht:

$$Md = 600 + \frac{\frac{50}{2} \ddot{y} 15}{17} \cdot (650 \ddot{y} 600)$$

$$= 629{,}41 .$$

Dieses Ergebnis bedeutet, dass 50 % der Lebensmittelgeschäfte einen Tagesumsatz von mehr als haben 629,41 Euro und die anderen 50 % haben rev

Die Gesamtsumme beträgt weniger als 629,41 Euro.

M

Machine Translated by Google

WEITERLESEN

Bedeuten

Maß für die zentrale Tendenz

Modus

Quantil

REFERENZEN

Boscovich, RJ: De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex exemplaria etiam sensorum impres sa. Bononiensi Scientiarium et Artium

Instituto Atque Academia Commentarii, vol. IV, S. 353–396 (1757)

Euler, L.: Recherche über die Frage nach den Galaxien der Saturn- und Jupiterbewegung, Teil des Jahresberichts 1748, von der Académie Royale des Sciences de Paris. Neuauflage 1960, von Leonhardi Euleri, Opera Omnia, 2. Serie. Turici, Basel, 25, S. 47–157 (1749)

Mayer, T.: Abhandlung über die Umwälzung des Monds um seine Achse und die schein

bloße Bewegung der Mondflecken. Kosmographische Nachrichten und Sammlungen auf das Jahr 17481, 52–183 (1750)

Mittlere absolute Abweichung

Die mittlere absolute Abweichung einer Menge Quantitative Beobachtungen sind ein **Maß für die Streuung.** Er entspricht dem Mittelwert der absoluten Abweichungswerte jeder Beobachtung relativ zum **Median.**

GESCHICHTE

Siehe L1- **Schätzung.**

MATHEMATISCHE ASPEKTE

Sei x1, *x2,...,xn* eine Menge von *n* Beobachtungen relativ zu einer **quantitativen Variablen** X.

Die mittlere absolute Abweichung, bezeichnet mit EMAD wird wie folgt berechnet:

$$\text{EMAD} = \frac{\sum_{i=1}^{N} |xi \ddot{y} Md|}{N},$$

wobei *Md* der Median der Beobachtungen ist.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die mittlere absolute Abweichung ist ein Funda mentale Vorstellung von der L1 **-Schätzung.** Die L1- **Schätzung** eines **Parameters** des **Maßes der zentralen Tendenz** ÿ ist eigentlich eine Schätzmethode, die **auf** dem minimalen Absolutwert basiert Abweichungen der Beobachtungen *xi* vom Parameter ÿ:

$$\sum_{i=1}^{N} |xi \ddot{y} \ddot{y}| .$$

Der Wert, der diesen Ausdruck minimiert, ist der **Median** der **Stichprobe.** Bei Daten, die Ausreißer enthalten, ist diese Methode effizienter als die Methode der **kleinsten Quadrate .**

WEITERLESEN

L1 -Schätzung

Maß der Streuung

Standardabweichung

Varianz

Methode der Momente

Für esti wird die Momentenmethode verwendet Anpassen der Parameter einer Verteilung aus einer **Stichprobe.** Die Idee ist wie folgt. Wenn die Anzahl der zu schätzenden Parameter gleich k ist, setzen wir die *k* ersten Momente der Verteilung (deren Ausdruck von unbekannten Parametern abhängt) gleich den entsprechenden empirischen Momenten, also dem

Machine Translated by Google

Schätzer der Momente der Ordnung *k* cal

anhand der Probe berechnet. Wir sollten das System von *k* Gleichungen mit *k* unbekannten Variablen lösen.

GESCHICHTE

Im späten 19. Jahrhundert nutzte **Pearson und Karl** die Momentenmethode zur Schätzung der

Parameter eines normalen gemischten Modells (mit Parametern p1, p2, ÿ1, ÿ2, ÿ2 ÿ2), gegeben durch

$$P(x) = \sum_{i=1}^{2} pi \cdot gi(x),$$

wobei die *GI* normalverteilt sind mit dem Mittelwert *ÿi* und der Varianz ÿ2 . Die Methode wurde unter anderem von Chuprov, AA (1874–1926), Thiele, Thor vald Nicolai, **Fisher, Ronald Aylmer** und **Pearson, Karl** weiterentwickelt und untersucht.

MATHEMATISCHE ASPEKTE

Seien *ÿ1,..., ÿk* die *k* Parameter, die für die Zufallsvariable X mit einer gegebenen Verteilung geschätzt werden sollen .

Wir bezeichnen mit ÿj((ÿ1,..., ÿk)) = *E Xj* das Moment der Ordnung *j* von *X ,* ausgedrückt als Funktion der zu schätzenden Parameter.
Die Momentenmethode besteht in der Lösung des folgenden Gleichungssystems:

$$\ddot{y}j(\ddot{y}1,\ldots,\ddot{y}k) = mj \text{ für } j = 1,\ldots,k.$$

Um zwischen realen Parametern und Schätzern zu unterscheiden, bezeichnen wir diese mit *ÿˆ1,...,* ÿˆ k.

BEISPIELE

1. Sei *X* eine Zufallsvariable, die einer **Exponentialverteilung** mit unbekanntem Parameter ÿ folgt. Wir haben *n* Beobachtungen *x1,...,xn* von X und wollen ÿ schätzen.

   Da das Moment der Ordnung 1 der Mittelwert von ist

   Um die Verteilung abzuschätzen, sollten wir den theoretischen und den empirischen Mittelwert angleichen. Wir wissen, dass (siehe Exponentialverteilung) der (theoretische) Mittelwert 1 ist gleich $\frac{1}{\ddot{y}}$. den gewünschten Schätzer ÿˆ erfüllt $\frac{1}{\ddot{y}} = \frac{1}{n}\sum_{i=1}^{N} xi.$
2. Nach der Methode der Momente werden der **Mittelwert ÿ** und die **Varianz** ÿ2 einer Zufallsvariablen *X ,* die einer **Normalverteilung** folgt, durch den empirischen Mittelwert x¯ und die Varianz geschätzt

   $$\frac{1}{N}\sum_{i=1}^{N} (xi - \bar{x})^2$$

   einer Stichprobe {x1, *x2,...,xn}* von X.

DOMÄNEN UND EINSCHRÄNKUNGEN

Zunächst sollten wir beobachten, ob das zuvor angegebene Gleichungssystem Lösungen hat.
Es kann keine Lösungen geben, wenn beispielsweise die zu schätzenden Parameter einer oder mehreren besonderen Einschränkungen unterliegen. Jedenfalls garantiert die Methode keine besonders effizienten Schätzer. Zum Beispiel der Schätzer für die Varianz einer Normalverteilung

Bution ist voreingenommen.

WEITERLESEN

- **Chi-Quadrat-Abstand**
- **Schätzer**
- **Kleinsten Quadrate**
- **Maximale Wahrscheinlichkeit**
- **Moment**
- **Parameter**
- **Probe**

VERWEISE

Fisher, RA: Momente und Produktmomente von Stichprobenverteilungen. Proz. London. Mathematik. Soc. **30,** 199–238 (1929)

M

Machine Translated by Google

Fisher, RA: Die Momente der Distribution
Bution für normale Stichproben von Maßen für die Abweichung von der Normalität. Proz. Roy. Soc. London. Ser. A **130,** 16–28 (1930)

Kendall, MG, Stuard, A.: Die fortgeschrittene Theorie der Statistik. Bd. 1. Verteilungstheorie, 4. Aufl. Griffin, London (1977)

Pearson, K.: Beiträge zur mathematischen Evolutionstheorie. I. In: Karl Pearsons frühe statistische Arbeiten. Cam Bridge University Press, Cambridge, S. 1–40 (1948). Erstveröffentlichung 1894 als: Zur Zerlegung asymmetrischer Frequenzkurven. Philos. Trans. Roy. Soc.
London. Ser. A **185,** 71–110

Thiele, TN: Theorie der Beobachtungen. C. und E. Layton, London; Ann. Mathematik. Stat. 2, 165–308 (1903)

Fehlende Daten

Fehlende Daten treten in Statistiken auf, wenn bestimmte Ergebnisse nicht vorliegen. Dies geschieht, wenn Daten nicht verfügbar sind (die Daten wurden durch einen Speicherfehler zerstört, ein Tier ist während eines **Experiments** gestorben usw.).

GESCHICHTE

Allan, FE und Wishart, J. (1930) waren offenbar die ersten, die sich mit dem Problem fehlender Daten bei der Analyse des Versuchsdesigns befassten. Ihren Werken folgten die von Yates, Frank (1933).

Allan, FE und Wishart, J. sowie Yates, F. versuchten, mithilfe der Iterationsmethode eine **Schätzung** fehlender Daten zu finden. In einer Situation, in der viele Beobachtungen fehlen, bietet die vorgeschlagene Methode keine zufriedenstellende Lösung. Bartlett, MS (1937) verwendete die **Kovarianzanalyse** zur Lösung desselben Problems.

Seitdem wurden viele andere Methoden vorgeschlagen, darunter die Methode von Rubin, DB (1972). Birkes, D. et al. (1976) stellten mithilfe der **Inzidenzmatrix** eine exakte Methode zur Lösung des Problems fehlender Daten vor . Der verallgemeinerte Matrixansatz wurde von Dodge, Y. und Majum dar, D. (1979) vorgeschlagen. Weitere Einzelheiten zu den historischen Aspekten fehlender Daten von 1933 bis 1985 finden sich in Dodge, Y. (1985).

MATHEMATISCHE ASPEKTE

Wir können das Problem fehlender Daten lösen, indem wir das Experiment unter den gleichen Bedingungen wiederholen und so die Werte für erhalten
fehlende Daten. Eine solche Lösung ist wirtschaftlich und materiell nicht immer möglich, auch wenn sie ideal ist.
Es gibt auch andere Ansätze, das Problem fehlender Daten im experimentellen Design zu lösen. Wir erwähnen den Ansatz durch die **verallgemeinerte Inverse** oder durch die Inzidenzmatrix , die angibt, welche Beobachtungen im Design fehlen.

WEITERLESEN

Daten

Versuchsplanung

Experiment

Überwachung

VERWEISE

Allan, FE, Wishart, J.: Eine Methode zur Schätzung des Ertrags einer fehlenden Parzelle in experimentellen Feldarbeiten. J. Landwirtschaft. Wissenschaft. **20,** 399–406 (1930)

Machine Translated by Google

Bartlett, MS: Einige Beispiele statistischer Forschungsmethoden in der Landwirtschaft und angewandten Biologie. J. Roy. Stat. Soc. (Suppl.) 4, 137–183 (1937)

Birkes, D., Dodge, Y., Seely, J.: Spannende Sets für schätzbare Kontraste in der Klassik fikationsmodelle. Ann. Stat. 4, 86–107 (1976)

Dodge, Y.: Analyse von Experimenten mit fehlenden Daten. Wiley, New York (1985)

Dodge, Y., Majumdar, D.: Ein Algorithmus zum Finden kleinster verallgemeinerter Quadratinversen für Klassifizierungsmodelle mit beliebigen Mustern. J. Stat. Berechnen. Simul. 9, 1–17 (1979)

Rubin, DB: Ein nicht iterativer Algorithmus für die Kleinste-Quadrate-Schätzung fehlender Werte in jeder Varianzanalyse. Appl. Stat. **21,** 136–141 (1972)

Yates, F.: Die Analyse wiederholter Experimente, wenn die Feldergebnisse vorliegen vollständig. Empire J. Exp. Landwirtschaft. 1, 129–142 (1933)

Fehlende Datenanalyse

Siehe **fehlende Daten.**

Modus

Der Modus ist ein **Maß für die zentrale Tendenz cy.** Der Modus einer Reihe von Beobachtungen ist der **Wert** der Beobachtung mit der höchsten **Häufigkeit.** Nach dieser Definition kann eine Verteilung einen eindeutigen Modus haben (unimodale Verteilung genannt). In manchen Situationen kann eine Verteilung viele Modi haben (sogenannte bimodale, trimodale, multimodale usw. Verteilung).

MATHEMATISCHE ASPEKTE

Betrachten Sie eine **Zufallsvariable** *X,* deren **Höhle Sity-Funktion** ist *f* (x). Mo ist ein Modus der Verteilung, wenn *f (x)* bei Mo ein lokales Maximum hat.

Empirisch wird der Modus einer Reihe von Beobachtungen wie folgt bestimmt:

1. Ermitteln Sie die **Häufigkeitsverteilung** von die Menge der Beobachtungen.
2. Der Modus oder die Modi sind die Werte, deren **Die Häufigkeit** ist größer oder gleich der Häufigkeit der anderen Beobachtungen.

Zur Verteilung von Beobachtungen Gruppiert in Klassen desselben Bereichs (= *Maximum* ÿ *Minimum) ,* deren Grenzen bekannt sind, wird der Modus wie folgt bestimmt: 1. Bestimmen Sie die Modalklassen – die Klassen mit einer Häufigkeit, die größer oder gleich der Häufigkeit anderer Klassen ist. (Wenn viele Klassen die gleiche Häufigkeit haben, gruppieren wir diese Klassen so, dass nur ein Modus dieser Klassenmenge berechnet wird).

2. Berechnen Sie den Modus Mo unter Berücksichtigung der Gelenkklassen:

$$Mo = L + \frac{d1}{d1 + d2} \cdot c ,$$

Wo

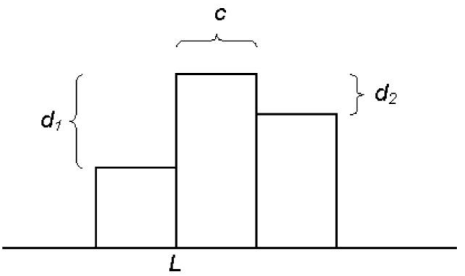
L ist die kleinere Grenze der Modalklasse,

d1 ist die Differenz zwischen der Häufigkeit der Modalklasse und der Häufigkeit der vorherigen Klasse,

d2 ist die Differenz zwischen der Häufigkeit der Modalklasse und der Häufigkeit der nächsten Klasse und

c ist die Länge der Modalklasse, die allen Klassen gemeinsam ist.

Machine Translated by Google



Wenn die Länge der Klassen nicht identisch ist,
Der Modus muss durch Ändern berechnet werden
die Einteilung der Klassen, um zu erhalten,
wenn möglich, sind die Klassen gleich lang.

DOMÄNEN UND EINSCHRÄNKUNGEN

Für eine diskrete **Variable** hat der Modus das
Vorteil der einfachen Bestimmung und
interpretieren. Es wird insbesondere dann verwendet, wenn
die Verteilung nicht symmetrisch ist.
Für eine kontinuierliche Variable die Bestimmung
des Modus ist nicht immer präzise, weil es
hängt von der Einteilung in Klassen ab. Im
Allgemeinen ist der Modus ein guter **Indikator** dafür
Der Mittelpunkt der Daten ist nur dann gegeben, wenn
es einen dominanten **Wert** im Datensatz gibt. Wenn es gibt
viele dominante Werte, die Verteilung ist
Plurimodal genannt. In diesem Fall die Modi
sind nicht das Maß der zentralen Tendenz.
Eine bimodale Verteilung weist im Allgemeinen darauf hin
dass die betrachtete **Population** in Wirklichkeit
heterogen ist und aus zwei Personen besteht
Subpopulationen mit unterschiedlichem Zentralwert
ues.

BEISPIELE

Die Menge sei die folgenden Zahlen:

1223334445555667789 .

Die aus diesem Satz abgeleitete **Häufigkeitstabelle**
Ist:

Werte	1	2	3	4	5	6	7	8	9	
Frequenz	1	2	3	3	4	2	2	1		1

Der Modus dieser Zahlenmenge ist 5, weil
es erscheint mit der größten **Häufigkeit** (4).
Betrachten Sie nun die folgende Häufigkeitstabelle:
was dem Tagesumsatz von 50 entspricht
Geschäfte, gruppiert nach Umsatzklasse:

Klasse (Umsatz in Euro)	Häufigkeit **fi** (Anzahl der Geschäfte)
500–550	3
550–600	12
600–650	17
650–700	8
700–750	6
750–800	4
Gesamt	50

Wir können sofort erkennen, dass die modale Klasse
ist 600–650, weil es die größte Frequenz hat
quenz.
Der Modus wird wie folgt berechnet:

$$Mo = L1 + \frac{d1}{d1 + d2} \cdot c$$

$$= 600 + \frac{5}{5 + 9} \cdot 50 = 617{,}86 \ .$$

Nehmen wir noch einmal das gleiche Beispiel, aber
Gruppierung der Zahlen in verschiedene Klassen:

Klasse (Umsatz in Euro)	Häufigkeit **fi** (Anzahl der Geschäfte)
500–600	15
600–650	17
650–700	8
700–800	10
Gesamt	50

Die Klassen haben nicht die gleiche Länge. Wir
wird eine neue Gruppierung durchführen, um Klassen zu erhalten
gleich lang.

Machine Translated by Google

Fall 1

Nehmen wir zum Beispiel als Referenz die Klasse 600–650, deren Länge 50 beträgt, und nehmen an, dass die Einnahmen gleichmäßig in einer Klasse verteilt sind.

Die Tabelle, die diese neue Gruppierung erzeugt, ist die Folgendes:

Klasse (Umsatz in Euro)	Häufigkeit f_i (Anzahl der Geschäfte)
500–550	7.5
550–600	7.5
600–650	17
650–700	8
700–750	5
750–800	5
Gesamt	50

Die Modalklasse ist die Klasse 600–650 und die Werte des Modus sind:

$$\text{Mo} = 600 + \frac{9.5}{9{,}5 + 9} \cdot 50 = 625{,}67\,.$$

Fall 2

Wir könnten unsere Berechnung auch auf die stützen Länge der ersten Klasse; Daraus ergibt sich folgende Tabelle:

Klasse (Umsatz in Euro)	Häufigkeit f_i (Anzahl der Geschäfte)
500–600	15
600–700	25
700–800	10
Gesamt	50

Die Modalklasse ist die Klasse 600–700 und die Werte des Modus sind:

$$\text{Mo} = 600 + \frac{10}{10 + 15} \cdot 100 = 640\,.$$

Diese Ergebnisse zeigen, dass wir vorsichtig sein sollten zur Verwendung des Modus, weil der erhaltene Die Ergebnisse sind unterschiedlich, je nachdem, wie die Daten werden in Klassen gruppiert.

WEITERLESEN

Bedeuten

Maß für die zentrale Tendenz
Maß für die Schiefe
Median

Modell

Ein Modell ist eine theoretische Darstellung von eine reale Situation. Es besteht aus mathematischen Symbolen. Ein mathematisches Modell, basierend auf einer bestimmten Anzahl von Beobachtungen und Hypothesen, versucht, das bestmögliche zu geben Beschreibung des untersuchten Phänomens. Ein mathematisches Modell enthält Wesentliches Es gibt zwei Arten von Elementen: die direkten oder indirekt beobachtbare Variablen, die betreffen das untersuchte Phänomen und die Parameter, bei denen es sich um feste, im Allgemeinen unbekannte Größen handelt, die die Variablen untereinander in Beziehung setzen sie im Modell.

M

GESCHICHTE

Der Begriff „Modell“ tauchte erstmals im auf 18. Jahrhundert, in drei Hauptproblemen:

1. Die beobachtete Ungleichheit verstehen zwischen den Bewegungen von Jupiter und Saturn.
2. Bestimmung und mathematische Beschreibung der Bewegungen des Mondes.
3. Bestimmung der Form der Erde.

Diese drei Probleme verwendeten Beobachtungen aus der Astronomie und etwas theoretischem Wissen über die Gravitation. Sie besetzten viele Wissenschaftler, insbesondere Euler, Leonhard (1749), der das Problem zuerst behandelte, und

Machine Translated by Google

Mayer, Tobias (1750), der interessiert war
in den Bewegungen des Mondes. Der dritte

Das Problem wurde von **Boscovich, Roger** behandelt
Joseph und Maire, Christopher (1755).

MATHEMATISCHE ASPEKTE

Es gibt viele Arten von Modellen. Unter
Für sie ist das deterministische Modell ein Modell, wo
die Relationen gelten als exakt. An

Andererseits geht das stochastische Modell davon
aus, dass die Beziehungen nicht exakt sind, sondern
durch einen Zufallsprozess ermittelt.
Ein stochastisches Modell wird hauptsächlich durch
die Natur unabhängiger Variablen bestimmt
(exogen oder erklärend). Wir sprechen über
Kontinuierliche Modelle, wenn die unabhängigen
Variablen *Xj* sind die quantitative Kontinuität
Variablen. Wir versuchen, das **Abhängige** auszudrücken
Variable *Y* (auch endogen oder genannt).
Antwort) durch eine Funktion des Typs:

$$Y = f\,X1,\ldots,Xp|\ddot{y}0,\ \ddot{y}1,\ldots,\ \ddot{y}p\ddot{y}1 + \ddot{y} \quad ,$$

wobei *ÿj* die zu schätzenden Parameter (oder
Koeffizienten) sind und ÿ ein nicht beobachtbarer
Zufallsfehler ist. Von dieser Art von Modell, wir
kann zwischen linearen und nichtlinearen Modellen unterscheiden
Modelle; die Linearität ist relativ zu den Koeffizienten.
Ein lineares Modell kann in geschrieben werden
Form eines Polynoms. Zum Beispiel:

$$Y = \ddot{y}0 + \ddot{y}1X + \ddot{y}2X2 + \ddot{y}$$

ist ein lineares Modell namens „zweiter Ordnung“
(Der Befehl kommt von der größten Macht von
Variable X). Das Model

$$\log V = \ddot{y}0 + \frac{\ddot{y}1}{W}$$

ist auch ein lineares Modell für die Parameter ÿ0 und
ÿ1. Dieses Modell wird auch für Variablen linear, die
Folgendes darstellen:

$$Y = \log V \text{ und } X = \frac{1}{W} \cdot$$

Es gibt:

$$Y = \ddot{y}0 + \ddot{y}1 \cdot X \quad .$$

Andererseits das Modell

$$Y = + \ddot{y}\frac{\ddot{y}0}{X + \ddot{y}1}$$

ist ein nichtlineares Modell für den Parameter ÿ1.
Es gibt viele Methoden zur Schätzung
Parameter eines Modells. Unter diesen Methoden können wir die
Methode der **kleinsten Quadrate** und die Methode der kleinsten Quadrate nennen
L1- **Schätzmethoden .**
Mit Ausnahme der kontinuierlichen Modelle reden wir
über diskrete Modelle, bei denen die untersuchte
Bevölkerung in Teilpopulationen unterteilt werden kann
auf der Grundlage qualitativer kategorialer Variablen
(oder Faktoren), wobei jede eine bestimmte Anzahl
von **Ebenen haben kann.** Der Typ des Modells
hängt vom etablierten **Experiment** ab
Design. Das Ziel eines solchen Designs ist es
Bestimmen Sie, wie die experimentellen Bedingungen
beeinflussen die beobachteten Werte der Antwort
Variable. Wir können verschiedene Typen unterscheiden
von diskreten Modellen:

- Ein Modell mit festen Effekten (Modell I) ist
  ein lineares Modell, für das die Faktoren gelten
  fest, das heißt, sie nehmen eine endliche Anzahl von Ebenen an
  die alle im etablierten Versuchsdesign abgebildet
  sind. In diesem Fall,
  Das Modell kann geschrieben werden:

$$\text{beobachteter Wert} = \begin{matrix} \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} \text{lineare Funktion} \\ \text{von unbekannt} \\ \text{Parameter} \end{matrix} \begin{matrix} \ddot{y} \\ \ddot{y} \end{matrix} + \text{Fehler ;}$$

  Die Fehler sind unabhängige Zufallsvariablen,
  deren Mittelwert gleich Null ist.
- Ein Modell mit variablen Effekten (Modell II)
  ist ein lineares Modell, für das der untersuchte Lev gilt
  Die anderen Faktoren jedes Faktors wurden zufällig ausgewählt

Machine Translated by Google

aus einer endlichen oder unendlichen Anzahl möglicher Ebenen. In diesem Fall können wir schreiben:

beobachteter Wert = konstant + lineare Funktion von Zufallsvariablen + Fehler ;

Die Zufallsvariablen sind unabhängig und haben **einen Mittelwert** von Null. • Ein Modell mit gemischten Effekten (Modell III) ist ein lineares Modell, das gleichzeitig enthält zeitlich feste und zufällige Faktoren. Also wir kann schreiben:

beobachteter Wert = ÿ + lineare Funktion von unbekannt Parameter ÿ + ÿ + lineare Funktion von zufällig Variablen ÿ + Fehler .

Wenn der Versuchsplan einen einzelnen Faktor enthält, der mit ÿ bezeichnet wird, kann das Modell wie folgt geschrieben werden:

$$Yij = ÿ + ÿi + ÿij ,$$
$$i = 1,...,I \text{ und } j = 1,...,ni ,$$

Dabei ist *I* die Anzahl der Ebenen eines Faktors ÿ, *ni* die Anzahl der Beobachtungen auf jeder Ebene, ÿ der **Mittelwert** der **Grundgesamtheit und** *ÿi* die mit jeder Ebene des Faktors verbundenen Effekte. ÿ und *ÿij* sind die unabhängigen Zufallsfehler, normalverteilt mit **Mittelwert** Null und homogenen **Varianzen** ÿ2.

Ein solches Modell kann je nach Art des Faktors vom Typ I oder Typ II sein: Es ist vom Typ I, wenn die Ebenen des Faktors festgelegt sind und alle im gesamten Design vorkommen. Es ist vom Typ II, wenn ÿ eine Zufallsvariable ist, die unabhängig von der Verteilung gemäß einer beliebigen Verteilung mit einem mathematischen Erwartungswert von Null und Vari ist ance ÿ2 (wenn die Niveaus von ÿ zufällig aus einer endlichen oder unendlichen Anzahl möglicher Niveaus ausgewählt werden).

Wenn das Versuchsdesign viele Faktoren enthält, zum Beispiel:

$$Yijk = ÿ + ÿi + ÿj + (ÿÿ)ij + ÿijk ,$$
$$i = 1,...,I,$$
$$j = 1,...,J \text{ und } k$$
$$= 1,...,nij ,$$

Das Modell kann vom Typ I (das Niveau aller Faktoren ist fest), II (alle Faktoren sind Zufallsvariablen mit den gleichen Einschränkungen wie) sein im vorherigen Modell) oder III (bestimmte Faktoren haben feste Ebenen, andere sind variabel). Der Begriff (ÿÿ)ij stellt hier den Effekt der **Wechselwirkung** zwischen den I-Ebenen des Faktors dar ÿ und die *J*- Niveaus des Faktors ÿ.

Ein Modell vom Typ I kann mithilfe einer **Varianzanalyse analysiert werden.**

WEITERLESEN

- **Varianzanalyse**
- **Abhängige Variable**
- **Versuchsplanung**
- **Unabhängige Variable**
- **Parameter**
- **Zufällige Variable**
- **Regressionsanalyse**
- **Variable**

VERWEISE

Anderson, VL, McLean, RA: Versuchsplanung: ein realistischer Ansatz. Marcel Dekker, New York (1974)

Boscovich, RJ und Maire, C.: De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus. Pal ladis, Rom (1755)

Machine Translated by Google

Euler, L.: Recherchiert über die Frage, die die Bewegung des Saturns und des Mondes bestimmt hat Jupiter, ein Teil, hat den Preis gemeldet Im Jahr 1748, von der Académie Royale des Wissenschaften von Paris. Republiziert im Jahr 1960, dat Leonhardi Euleri, Opera Omnia, 2ème Serie. Turici, Basel, 25, S. 47–157 (1749)

Mayer, T.: Abhandlung über die Umwälzung desMonds um seineAchse und die schein bloße Bewegung der Mondflecken. Kosmographische Nachrichten und Sammlungen auf das Jahr 17481, 52–183 (1750)

De Moivre, Abraham

Abraham de Moivre (1667–1754) studierte zunächst in Sedan und Saumur (beide in Frankreich) bevor er sein Mathematikstudium fortsetzte und Physik an der Sorbonne. Kommen von Als er aus einer protestantischen Familie stammte, musste er diese verlassen Frankreich im Jahr 1685 aufgrund der Aufhebung des Edikt von Nantes und wurde nach London verbannt.

Nach einem schwierigen Start in seiner neuen Heimat hat er konnte Mathematik studieren. Er wurde zum Mitglied der Royal Society of London gewählt im Jahr 1697.

Sein wichtigstes Werk war *The Doc trine of Chances,* das 1718 erschien und widmete sich Fragen, die auch heute noch eine grundlegende Rolle spielen Wahrscheinlichkeitstheorie und *Miscellanea Analytica de Seriebus et Quadraturis,* wo für Zum ersten Mal erschien die **Dichtefunktion** der **Normalverteilung** im Con

Text von Wahrscheinlichkeitsberechnungen für Spiele der Gefahr. **Carl Friedrich Gauß** übernahm die dieselbe Dichtefunktion und wandte sie auf Messprobleme in der Astronomie an. Er auch leistete Beiträge zur Versicherungsmathematik und mathematische Analyse.

Hauptwerke von Abraham De Moivre:

1718 Die Lehre vom Zufall: oder Eine Methode zur Berechnung der Wahrscheinlichkeit von Ereignissen im Spiel. Pearson, London.

1725 Renten auf Leben. Pearson, London.

1730 Miscellanea Analytica de Seriebus et Quadraturis. Tonson und Watts, London.

WEITERLESEN

De Moivre-Laplace-Theorem
Normalverteilung

De Moivre-Laplace-Theorem

Dieser Satz begründete die Beziehung zwischen der **Binomialverteilung** mit Parameter *n* und *p* und die **Normalverteilung** für *n* , die gegen Unendlich geht. Sein Nutzen besteht insbesondere in der Schätzung der **Verteilungsfunktion** *F (k)* des Binomials Verteilung durch die Verteilungsfunktion von die Normalverteilung, die gefunden wird sei einfacher.

GESCHICHTE

Dieser Satz stammt aus dem Jahr 1870, als **Pierre Simon de Laplace** studierte, zusammen mit anderen (darunter **de Moivre, Abraham),** Probleme im Zusammenhang mit der Annäherung an die Binomialverteilung und zur Theorie von Fehler.

MATHEMATISCHE ASPEKTE

Sei *X1 ,* _ _ _ _ _ _

p < 1.

Machine Translated by Google

Wir betrachten das entsprechende angepasst Variable *Zn, n = 1, 2,...* wie zum Beispiel:

$$Zn = \frac{Xn \ddot{y} np}{\ddot{y}npq},$$

wobei *np* den **Mittelwert** einer Binomialverteilung darstellt und $\ddot{y}npq$ deren **Varianz** (die (Variable *Zn* entspricht der **Normalisierung** von Xn).

Das De-Moivre-Laplace-Theorem wird wie folgt ausgedrückt:

$$P(Zn \ddot{y} x) \ddot{y}\ddot{y}n\ddot{y}\ddot{y}(x),$$

Wo ist die **Verteilungsfunktion** des Standardnormalverteilung .

Auch die Näherung kann verbessert werden unter Verwendung des **Korrekturfaktors** „eine Hälfte“. Daher folgende Näherung:

$$P(Xn \ddot{y} x) \ddot{y}= 1 \ddot{y} \frac{x \ddot{y} 2 \frac{1}{} \ddot{y} np}{\ddot{y}npq}$$

ist besser als diejenigen, die man ohne Einnahme erhält Berücksichtigung des Korrekturfaktors $_2 .^{\frac{1}{}}$

DOMÄNEN UND EINSCHRÄNKUNGEN

Dieser Satz ist ein Sonderfall des **zentralen Grenzwertsatzes.** Eine **Binomialverteilung**

Bution (wie zuvor definiert) kann sein geschrieben als Summe von *n* Zufallsvariablen dis

Tribut nach einer **Bernoulli-Verteilung** des **Parameters** p. Der **Mittelwert** einer Bernoulli-Zufallsvariablen ist gleich *p* und seine **Varianz** zu *p* (1 ÿ p). Das de Moivre–

Der Laplace-Satz lässt sich leicht daraus ableiten Zentraler Grenzwertsatz.

BEISPIELE

Wir wollen die Wahrscheinlichkeit von mehr berechnen mehr als 27 Erfolge in einem **Binomialexperiment**

von 100 Versuchen, wobei jeder Versuch eine Wahrscheinlichkeit hat des Erfolgs von *p* = 0,2. Das gewünschte Binomial Wahrscheinlichkeit kann durch die Summe ausgedrückt werden:

$$P(Xn > 27) = \sum_{k=27}^{100} \binom{100}{k} \cdot (0,2)^k (0,8)^{100.000}.$$

Die direkte Berechnung dieser Wahrscheinlichkeit erfordert eine Bewertung von jeweils 74 Begriffen des Formulars

$$\binom{100}{k} (0,2)^k (0,8)^{100.000}.$$

Unter Verwendung des De-Moivre-Laplace-Theorems ist dies Wahrscheinlichkeit kann näherungsweise abgeschätzt werden unter Berücksichtigung der Beziehung zwischen den Binomial- und Normalverteilungen. Daher:

$$P(Xn > 27) \ddot{y}= 1 \ddot{y} \frac{27 \ddot{y} np}{\ddot{y}npq}$$

$$= 1 \ddot{y} \frac{27 \ddot{y} 100 \cdot 0,2}{\ddot{y}100 \cdot 0,2 \cdot 0,8}$$

$$= 1 \ddot{y} (1,75).$$

Unter Bezugnahme auf die **normale Tabelle** erhalten wir (1,75) = 0,9599, also:

$$P(Xn > 27) \ddot{y}= 0,0401.$$

Der genaue **Wert** auf 4 Dezimalstellen genau Wahrscheinlichkeit ist: *P (Xn* > 27) = 0,0558. Ein Vergleich der beiden Werte 0,0401 und 0,0558 zeigt, dass die Näherung durch die Die Normalverteilung liefert ein sehr gutes Ergebnis nahe am genauen Wert.

Wir erhalten die folgende Näherung unter Verwendung des Korrekturfaktors:

$$P(Xn > x) \ddot{y}= 1 \ddot{y} (1,625) = 0,0521.$$

Das Ergebnis liegt näher am exakten Wert 0,0558 als diejenigen, die ohne Verwendung des Korrekturfaktors (0,0401) erhalten wurden.

M

Machine Translated by Google

WEITERLESEN

Binomialverteilung

Zentraler Grenzwertsatz

Normalverteilung

Normalisierung

REFERENZ

Laplace, PS: Merken Sie sich die Approximationen von Formeln, die in sehr großer Zahl funktionieren, und ihre Anwendung auf Wahrscheinlichkeiten. Mémoires de l'Académie Royale des Sciences de Paris, 10. Abgebildet in: Œuvres de Laplace **12,** 301–347 (1810)

Moivre, A. de: Approximatio ad summam terminorum binomii *(a* + b) n, in seriem expansi. Supplementum II zu Miscellanae Analytica, S. 1–7 (1733). Fotografisch nachgedruckt in einer seltenen Broschüre über Moivre und einige seiner Entdeckungen. Kneipe

gelistet von Archibald, RC Isis 8, 671–683 (1926)

Sheynin, OB: PSLaplaces Arbeit zur Wahrscheinlichkeit, Archive for History of Exact Sciences **16,** 137–187 (1976)

Sheynin, OB: PS Laplaces Fehlertheorie, Archive for History of Exact Sciences **17,** 1–61 (1977)

Moment

Wir nennen das Ordnungsmoment *k* des **Zufalls**

Variable *X* relativ zu einem beliebigen **Wert** x0 die mittlere Differenz zwischen der Zufallsvariablen und

x0, hoch k. Wenn x0 = 0, sprechen wir vom Moment (oder Anfangsmoment) der Ordnung k.

Wenn x0 = ÿ, wobei ÿ der **erwartete Wert** der Zufallsvariablen X ist, beziehen wir uns auf das zentrale Moment der Ordnung k.

GESCHICHTE

Der Begriff „Moment“ in der Statistik geht auf den Begriff „Moment“ in der Physik zurück

Dies leitet sich von der Entdeckung der Betätigung des Hebels durch Archimedes (287–212 v. Chr.) ab, wo wir vom Moment der Kräfte sprechen. Die korrigierten Formeln zur Berechnung der

Momente gruppierter Daten wurden von Sheppard, WF (1886) angepasst.

MATHEMATISCHE ASPEKTE

Das Anfangsmoment der Ordnung *k* eines diskreten

Die Zufallsvariable *X* wird ausgedrückt durch:

$$E\,Xk = \sum_{i=1}^{N} xk_{_}\, P\,(xi)\ ,$$

wenn es existiert, das heißt, wenn $E\,Xk < ÿ$, wobei *P (x)* ist die **Wahrscheinlichkeitsfunktion** der diskreten Zufallsvariablen *X,* die die *n* Werte annimmt *x1,...,xn.*

Im Fall einer kontinuierlichen Zufallsvariablen ist das Moment, sofern es existiert, definiert durch:

$$E\,Xk = \int_{ÿÿ}^{ÿ} xk\, f\,(x)\, dx \quad ,$$

wobei *f (x)* die **Dichtefunktion** der kontinuierlichen Zufallsvariablen ist.

Für *k* = 1 ist der Anfangszeitpunkt derselbe wie der **erwartete Wert** der Zufallsvariablen

X (falls dieses *E(X)* existiert):

$$(X \text{ diskret})\ E\,[X] = \sum_{i=1}^{N} xiP\,(xi),$$

$$(X \text{ kontinuierlich})\ [X] = \int_{ÿÿ}^{ÿ} xf\,(x)\, dx\ .$$

Das zentrale Moment der Ordnung *k* eines diskreten

Die Zufallsvariable X, falls vorhanden, wird ausgedrückt durch:

$$E\,(X ÿ E\,[X])^{k} = \sum_{i=1}^{N} (xi ÿ E\,[X])^{k}\, P\,(xi)\ .$$

Machine Translated by Google

Auf die gleiche Weise gilt im Fall einer kontinuierlichen Zufallsvariablen:

$$E\,(X \ddot{y}\ E\,[X])^{k} = \int_{\ddot{y}\ddot{y}}^{\ddot{y}} (x \ddot{y}\ E\,[X])^{k} f\,(x)\,dx$$

wenn das Integral wohldefiniert ist.

Es ist offensichtlich, dass für jede Zufallsvariable X mit einem endlichen mathematischen Erwartungswert das zentrierte Moment der Ordnung 1 Null ist:

$$E\,[X \ddot{y}\ \ddot{y}] = \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})\,P\,(xi) = \sum_{i=1}^{N} xiP\,(xi) \ddot{y}\ \ddot{y} \sum_{i=1}^{N} P\,(xi) = \ddot{y}\ \ddot{y}\ \ddot{y} = 0\ .$$

Wenn es andererseits existiert, ist das zentrale Moment der Ordnung 2 dasselbe wie das **Vari Wert** der Zufallsvariablen:

$$(X \text{ diskret})\ E\,(X \ddot{y}\ \ddot{y})^{2} = \sum_{i=1}^{N} (xi \ddot{y}\ \ddot{y})^{2}\,P\,(xi) = \mathrm{Var(X)}\ ,$$

$$(X \text{ kontinuierlich})\ E\,(X \ddot{y}\ \ddot{y})^{2} = \int_{\ddot{y}\ddot{y}}^{\ddot{y}} (x \ddot{y}\ \ddot{y})^{2} f\,(x)\,dx = \mathrm{Var(X)}\ .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Der **Schätzer** des Moments der Ordnung k cal berechnet für eine **Stichprobe** x1, *x2,...,xn* wird mit *mk bezeichnet*. Es entspricht:

$$mk = \frac{\sum_{i=1}^{N} (xi \ddot{y}\ x0)^{k}}{N}\ ,$$

wobei n die Gesamtzahl der Beobachtungen ist.

Wenn x0 = 0, dann haben wir das Moment (oder Anfangsmoment) der Ordnung k.

Wenn x0 = $\bar{x}$, wobei x¯ das **arithmetische Mittel ist,** haben wir vom zentralen Moment der Ordnung k.

Für den Fall, dass eine **Zufallsvariable** X nimmt seine Werte *xi* mit den Frequenzen fi, *i = 1,* 2,...,h an, das Moment der Ordnung k ist gegeben durch:

$$mk = \frac{\sum_{i=1}^{H} k\ fi\ (xi \ddot{y}\ x0)}{\sum_{i=1}^{H} fi}\ .$$

WEITERLESEN

- **Arithmetisches Mittel**
- **Dichtefunktion**
- **Erwarteter Wert**
- **Wahrscheinlichkeitsfunktion**
- **Zufällige Variable**
- **Varianz**
- **Gewichtetes arithmetisches Mittel**

VERWEISE

Sheppard, WF: Zur Berechnung der wahrscheinlichsten Werte von Frequenzkonstanten für Daten, die nach äquidistanten Teilungen einer Skala angeordnet sind. Londoner Mathematik. Soc. Proz. **29,** 353–380 (1898)

Monte-Carlo-Methode

Jede numerische Technik zur Lösung mathematischer Probleme, die Zufalls- oder Pseudozufallszahlen verwendet, wird Monte Car genannt lo Methode. Diese mathematisch oder statistisch

Für Probleme gibt es keine analytischen Lösungen.

GESCHICHTE

Der Name Monte Carlo stammt von der gleichnamigen Stadt in Monaco, die für ihre Casinos bekannt ist.

Das Roulette ist einer der einfachsten Mechanismen zur **Generierung von Zufallszahlen.**

Laut Sobol, IM (1975) verdankt die Monte-Carlo-Methode ihre Existenz dem

Amerikanische Mathematiker von Neumann,

Machine Translated by Google

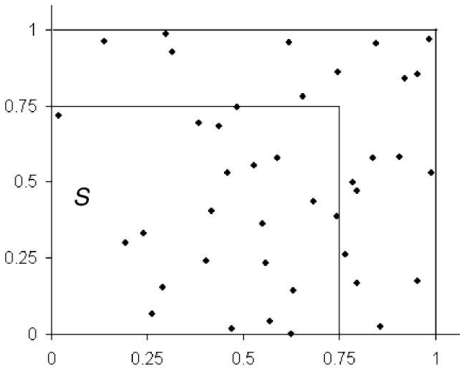
John (1951), Ulam, Stanislaw Marcin und Metropolis, Nicholas (1949). Es wurde um 1949 entwickelt, wurde aber praktisch mit der Erfindung des Computers.

BEISPIELE

Angenommen, wir möchten die Oberfläche berechnen einer Figur *S* , die sich innerhalb eines Quadrats befindet; seine Seite ist gleich 1. Wir werden N zufällig generieren Punkte innerhalb des Quadrats. Dazu notieren wir das Quadrat im System der Senkrechten Achsen, deren Ursprung der untere linke Winkel ist das Quadrat. Das bedeutet, dass alle Punkte, die Wir werden innerhalb des Quadrats erzeugen Koordinaten zwischen 0 und 1. Es reicht um einheitliche Zufallsvariablen zu erhalten ein Punkt. Die erste **Zufallszahl** wird sein die Abszisse und die Sekunde die Ordinate.

Im folgenden Beispiel wählen wir 40 zufällige Punkte. Wir benötigen zwei Proben davon 40 Zufallszahlen. Die Oberfläche *S* , die wir schätzen möchte, ist das Quadrat von 0,75 der Seite, also den genauen **Wert** entspricht 0,5625.

Nach dem Erzeugen der 40 zufälligen Punkte in Als Einheitsquadrat reicht es aus, die zu zählen Anzahl der Punkte innerhalb der Abbildung *S* (hier 21). Fall). Die **Schätzung** der gesuchten Oberfläche ergibt sich aus dem Verhältnis = 0,525. $\frac{21}{40}$



Es ist möglich, diese Schätzung zu verbessern, indem man das Experiment mehrmals **durchführt** und den **Mittelwert** der erhaltenen Oberflächen bildet.

WEITERLESEN

Generierung von Zufallszahlen

Jackknife-Methode

Zufallszahl

Simulation

VERWEISE

Metropolis, N., Ulam, S.: Das Monte Carlo Methode. Marmelade. Stat. Assoc. **44,** 335–341 (1949)

Sobol, IM: Die Monte-Carlo-Methode. Mir Verlag, Moskau (1975)

Von Neumann, J.: Verschiedene Techniken im Einsatz in Verbindung mit Zufallsziffern. Symposium des National Bureau of Standards, NBS, Reihe „Angewandte Mathematik“. 12, S. 36–38, National Bureau of Stan dards, Washington, D.C. (1951)

Gleitender Durchschnitt

Den gleitenden Durchschnitt der Ordnung nennen wir *k* **Arithmetisches Mittel** berechnet auf *k* aufeinanderfolgenden Werte **von Zeitreihen.** Der Rest dieser Arith

Metische Mittel geben die Reihe der gleitenden Durchschnittswerte der Ordnung k an. Der Prozess, der darin besteht Das Ersetzen der ursprünglichen Reihe wird als Glättung der Zeitreihe bezeichnet.

Bei den Zeitreihen, die aus jährlichen oder monatlichen Daten bestehen, werden die Durchschnittswerte als gleitende Durchschnitte über *k* Jahre bzw. *k* Monate bezeichnet. Wenn

Wir verfügen beispielsweise über monatliche Daten, die einen gleitenden Durchschnitt über 12 Monate berechnen

Machine Translated by Google

Die erhaltenen Ergebnisse entsprechen der Mitte des betrachteten Zeitraums, am Ende des sechsten Monat (oder am ersten Tag des siebten Monats), anstatt in der betroffen zu sein Mitte eines Monats wie in den Originaldaten.

Wir korrigieren dies, indem wir eine zentrierte Bewegung durchführen Durchschnitt über 12 Monate. Im Allgemeinen wird ein zentrierter gleitender Durchschnitt der Ordnung *k* berechnet Den gleitenden Durchschnitt der Ordnung 2 bilden gleitender Durchschnitt der Ordnung *k* der Anfangszeit Serie.

GESCHICHTE

Siehe **Zeitreihen.**

MATHEMATISCHE ASPEKTE

Wir haben die **Zeitreihen** Y1, Y2, Y3, ..., *YN,* und wir definieren die Reihe der gleitenden Durchschnitte der Ordnung *k* durch die Reihe arithmetischer Mittel:

$$\frac{Y1 + Y2 + \cdots + Yk}{k},$$

$$\frac{Y2 + Y3 + \cdots + Yk+1}{k}, \ldots,$$

$$\frac{YN\ddot{y}k+1 + YN\ddot{y}k+2 + \cdots + YN}{k}.$$

Die Summe der Zähler wird als Bewegung bezeichnet Summen der Ordnung k.

Wir können die gewichteten arithmetischen Mittel verwenden mit vorher angegebenen Gewichten; der Rest von was auf diese Weise erhalten wird, nennt man eine Reihe von gewichteten gleitenden Durchschnitten der Ordnung k.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die gleitenden Durchschnitte werden bei der Untersuchung von verwendet Zeitreihen zur **Abschätzung** der **säkularen Tendenz,** saisonaler Schwankungen und von zyklische Schwankungen.

Die Operation „gleitender Durchschnitt“ wird angewendet zu einer Zeitreihe ermöglicht:

- Bestimmen Sie die saisonalen Schwankungen in der Grenzen, wo sie streng periodisch sind,
- Glatte unregelmäßige Variationen und
- Etwaige Verlängerung der Saison einsparen al-Bewegung.

Lassen Sie uns die Nachteile der Methode des gleitenden Durchschnitts erwähnen:

- Die Daten vom Anfang und Ende von Serien fehlen.
- Gleitende Durchschnitte können Zyklen oder andere Bewegungen verursachen, die im Original nicht vorhanden sind Daten.
- Gleitende Durchschnitte werden stark beeinflusst durch Ausreißer und zufällige Werte.
- Wenn ein Diagramm der untersuchten Zeitreihe zeigt eine exponentielle säkulare Tendenz oder allgemeiner eine starke Kurve, die

Die Methode des gleitenden Durchschnitts liefert keine Ergebnisse sehr genaue Ergebnisse zur Abschätzung von außersaisonale Bewegung.

M

BEISPIELE

Mit den **Zeitreihen** 7, 2, 3, 4, 5, 0, 1 Daten, wir erhalten die gleitenden Durchschnitte der Ordnung 3:

$$\frac{7 + 2 + 3}{3}, \quad \frac{2 + 3 + 4}{3}, \quad \frac{3 + 4 + 5}{3},$$

$$\frac{3\,4 + 5 + 0}{3}, \quad \frac{3\,5 + 0 + 1}{3},$$

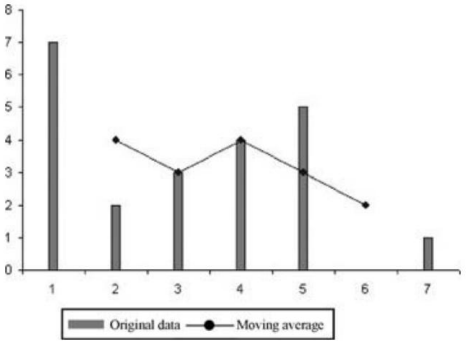
das heißt, 4, 3, 4, 3, 2.

In einer Reihe mit gleitenden Durchschnitten kommt dies häufig vor um jede Zahl anhand ihrer relativen Position zu den Originaldaten zu lokalisieren. In diesem Beispiel wir schreiben:

Originale Daten	7, 2, 3, 4, 5, 0, 1
Gleitender Durchschnitt der Ordnung 3	4, 3, 4, 3, 2

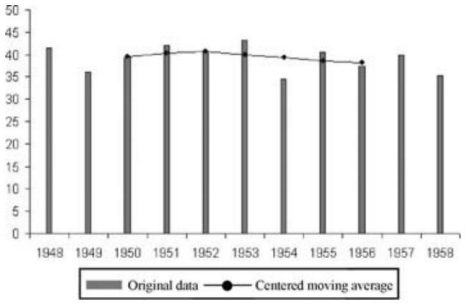
Machine Translated by Google

In der Grafik unten ist jede Zahl der gleitenden Durchschnittsreihe das **arithmetische Mittel** von die drei Zahlen direkt darunter.



Wir schlagen vor, den zentrierten gleitenden Durchschnitt über vier Jahre durchschnittlich monatlich zu berechnen Ölförderung in einem Land. Im ersten Schritt wir berechnen die Umzugssummen über 4 Jahre, Addition der Daten von 4 aufeinanderfolgenden Jahren. Dann, Wenn wir jede gleitende Summe durch 4 dividieren, erhalten wir gleitende Durchschnitte über 4 Jahre, wobei der letzte dem 1. Januar des dritten Jahres entspricht, in unterschiedlichen Beträgen betrachtet. Schließlich erhalten wir Gleitende Durchschnitte über 4 Jahre, berechnend das arithmetische Mittel zweier gleitender Durchschnitte über 4 aufeinanderfolgende Jahre.

Verschiebung des Jahresdatums	Summe über 4 Jahre	Ziehen um Durchschnitt über 4 Jahre	Zentriert ziehen um Durchschnitt über 4 Jahre
1948 41.4			
1949 36,0			
	158,8	39.7	
1950 39.3			39.6
	158,0	39,5	
1951 42.1			40.4
	165,2	41.3	
1952 40.6			40.7
	160,4	40.1	
1953 43,2			39.9
	158,8	39.7	
1954 34,5			39.3
	155,6	38.9	
1955 40,5			38,5
	152,4	38.1	
1956 37.4			38.2
	153,2	38.3	
1957 40,0			
1958 35.3			



Durchschnittliche monatliche Ölproduktion in einem Land in Millionen Tonnen

WEITERLESEN

- **Arithmetisches Mittel**
- **Zeitfolgen**
- **Gewichtetes arithmetisches Mittel**

Multikollinearität

Multikollinearität ist ein Begriff zur Beschreibung eine Situation, in der eine **Variable** nahezu gleich ist zu einer linearen Kombination anderer Variablen.

Manchmal liegt Multikollinearität vor Regressionsprobleme, wenn es eine starke gibt Korrelation zwischen den erklärenden Variablen.

Machine Translated by Google

Das Vorhandensein linearer oder nahezu linearer Rela

Unterschiede zwischen unabhängigen Variablen führen zu einer großen Ungenauigkeit bei der Schätzung der Regressionskoeffizienten . Wenn die Beziehungen streng linear wären, wären die Koeffizienten verzerrt.

MATHEMATISCHE ASPEKTE

Siehe **Ridge-Regression.**

WEITERLESEN

Korrelationskoeffizient

Multiple lineare Regression

VERWEISE

Belsley, D., Kuh, E., Welsch, RE: Regressionsdiagnostik Identifizierung einflussreicher Daten und Quellen der Kollinearität. Wiley, New York (1976)

Multimodale Verteilung

Siehe **Häufigkeitsverteilung.**

Multinomiale Verteilung

Betrachten Sie ein **Zufallsexperiment** mit *n* unabhängigen Versuchen. Angenommen, jeder Versuch kann nur ein **Ereignis** *Ei* mit einer **Wahrscheinlichkeit** *pi ergeben.*
Die Zufallsvariable *Xi* repräsentiert die Häufigkeit des Auftretens des **Ereignisses** *Ei.*
Die Multinomialverteilung mit **den Parametern** *n* und *pi* wird durch die folgende **gemeinsame Dichtefunktion** definiert :

$$P(X1 = x1, X2 = x2,...,Xs = xs) = \frac{N!}{\prod_{i=1}^{s} xi!} \prod_{i=1}^{s} pxi \ ,$$

mit den ganzen Zahlen *xi* ÿ 0, so dass das so ist $\sum_{i=1}^{} xi = n.$

Tatsächlich ist die **Wahrscheinlichkeit ,** das zu erhalten Reihenfolge

$$\underbrace{E1,...,E1}_{x1} ; \underbrace{E2,...,E2}_{x2} ;...; \underbrace{Es,...,Es}_{xs}$$

ist gleich $\prod_{i=1}^{s} pxi$.

Die Anzahl der vorhandenen Möglichkeiten ist die Anzahl der **Permutationen** von *n* Objekten, von denen x1 **Ereignis** E1 und x2 Ereignis E2 ist . . . , *xs* ist das Ereignis *Es,* was bedeutet, dass:

$$\frac{N!}{\prod_{i=1}^{s} xi!} .$$

Die Multinomialverteilung ist eine **diskrete Wahrscheinlichkeitsverteilung.**

MATHEMATISCHE ASPEKTE

Für die Multinomialverteilung

$$\sum_{i=1}^{s} pi = 1 \text{ und } \sum_{i=1}^{s} xi = n \ ,$$

mit $0 < pi < 1$.
Für jede **Zufallsvariable** Xi wird der **Erwartungswert** berechnet .
Wir haben daher:

$$E[Xi] = n \cdot pi .$$

Tatsächlich kann jede **Zufallsvariable** con sein sideredindividually als folgende **Abinomialverteilung.**

Wir können die gleiche Bemerkung für die **Vari** machen **ance.** Es ist dann gleich:

$$\text{Var(Xi)} = n \cdot pi \cdot qi .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Multinomialverteilung hat viele Anwendungen in der statistischen Analyse

Machine Translated by Google

Daten. Es wird insbesondere in Situationen verwendet, in denen die Daten in mehrere **Kategorien** klassifiziert werden müssen

Ereignisse (z . B. bei der Analyse kategorialer Daten).

BEISPIELE

Betrachten Sie das folgende **Zufallsexperiment:**

Eine Urne enthält neun Kugeln, zwei weiße, drei rote und vier schwarze. Ein Ball wird nach dem Zufallsprinzip ausgewählt; Seine Farbe wird notiert und er wird zurück in die Urne gelegt. Dies wird 5 Mal wiederholt.
Wir können dieses **Zufallsexperiment** wie folgt beschreiben:

Wir haben drei mögliche **Ereignisse:**

E1: Einen weißen Ball auswählen

E2: Einen roten Ball auswählen

E3: Einen schwarzen Ball wählen

und drei **Zufallsvariablen:**

X1: Anzahl der gewählten weißen Kugeln

X2: Anzahl der gewählten roten Kugeln

X3: Anzahl der gewählten schwarzen Kugeln

Die mit diesen drei verbundenen **Wahrscheinlichkeiten**
Veranstaltungen sind:

$$2\ p1 = \bar{9},$$

$$3\ p2 = \bar{9},$$

$$4\ p3 = \bar{9}.$$

Wir berechnen die Wahrscheinlichkeit , dass bei fünf ausgewählten Bällen *(n* = 5) einer weiß, zwei rot und zwei schwarz sein werden.

Die Anzahl der Möglichkeiten ist gleich:

$$5!/(1!2!2!).$$

Die **Wahrscheinlichkeit ,** die Folge (w, r, r, b, b) zu erhalten, wobei w, r und b jeweils weiße, rote und schwarze Kugeln sind, beträgt:

$$\prod_{i=1}^{3} pxi = \frac{2}{9} \cdot \frac{3}{9}^{2} \cdot \frac{4}{9}^{2} = 0{,}0049 .$$

Somit,

$$P\,(X1 = x1, X2 = x2, X3 = x3) = \frac{5!}{\prod_{i=1}^{3} xi!} \cdot \frac{2}{9}^{x1} \cdot \frac{3}{9}^{x2} \cdot \frac{4}{9}^{x3},$$

Wo

$$P(X1 = 1, X2 = 2, X3 = 2) = 30 \cdot 0{,}0049 = 0{,}147 .$$

WEITERLESEN

Binomialverteilung

Diskrete Wahrscheinlichkeitsverteilung

Multipler Korrelationskoeffizient

Siehe **Korrelationskoeffizient.**

Multiple lineare Regression

Eine **Regressionsanalyse ,** bei der **die unabhängige Variable** *Y* linear von vielen unabhängigen Variablen X1, *X2,...,Xk* abhängt, wird als multiple lineare Regression bezeichnet.
Eine multiple lineare Regressionsgleichung hat die Form:

$$Y = f\,(X1, X2,...,Xk) ,$$

wobei *f* (X1, *X2,...,Xk)* eine lineare Funktion von X1, *X2,...,Xk ist.*

GESCHICHTE

Siehe **Regressionsanalyse.**

MATHEMATISCHE ASPEKTE

Ein allgemeines **Modell** der multiplen linearen Regression mit *k = (p* ÿ 1) unabhängigen Variablen (und *n* Beobachtungen) wird wie folgt geschrieben:

$$Yi = ÿ0 + \sum_{j=1}^{pÿ1} Xjiÿj + ÿi , \qquad i = 1,...,n,$$

Machine Translated by Google

Dabei ist *Yi* die **abhängige Variable,** *Xji, j = 1,...,p* ÿ 1 die unabhängigen Variablen, *ÿj, j = 0,...,p* ÿ 1 die zu schätzenden Parameter und *ÿi* der Begriff des zufälligen, nicht beobachtbaren **Fehlers.**

In der Matrixform wird dieses **Modell** wie folgt geschrieben:

$$\mathbf{Y} = \mathbf{X}\ddot{y} + \ddot{y} \quad ,$$

wobei **Y** der Vektor (n×1) der Beobachtungen in Bezug auf die abhängige Variable *(n* Beobachtungen) ist, **ÿ** der Vektor (p×1) der zu schätzenden Parameter ist, **ÿ** der Vektor (n×1) der Fehler ist,

$$\text{undX} = \begin{matrix} \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} 1 & X11 & \ldots & X1(p\ddot{y}1) \\ \vdots & \vdots & & \vdots \\ \ddot{y} & 1 & Xn1 & \ldots & Xn(p\ddot{y}1) \end{matrix} \begin{matrix} \\ \ddot{y} \end{matrix}$$

ist der (n×p) **-Matrix** der unabhängigen Variablen.

Schätzung des Vektors ÿ

Ausgehend vom Modell

$$\mathbf{Y} = \mathbf{X}\ddot{y} + \ddot{y} \quad ,$$

Wir erhalten die Schätzung ÿˆ des Vektors **ÿ** mit der Methode der **kleinsten Quadrate :**

$$\ddot{y}\hat{} = (\mathbf{X}\ X)\ \ddot{y}1\mathbf{X}\ \mathbf{Y} \quad ,$$

und wir berechnen einen Schätzwert Y für Y:

$$\mathbf{Y} = \mathbf{X} \cdot \ddot{y}\hat{} \ .$$

In diesem Schritt können wir die durch den Vektor e bezeichneten Residuen berechnen , die wir auf folgende Weise finden:

$$\mathbf{e} = \mathbf{Y}\ \ddot{y}\ \mathbf{Y} \quad .$$

Maß für die Zuverlässigkeit der Schätzung von Y

Um zu wissen, welcher Maßnahme man für den Cho vertrauen kann

Bei einem linearen Modell ist es nützlich, eine **Varianzanalyse** durchzuführen und die Hypothesen auf dem Vektor **ÿ** des Regressionsmodells zu testen. Um diese Tests durchzuführen, müssen wir die folgenden Annahmen treffen:

- Für jeden **Wert** von *Xji* und für alle *i = 1,...,nandj* = 1 ...,pÿ1,gibt es **eine zufällige Variable Y ,** die gemäß der **Normalverteilung verteilt ist.**
- Die **Varianz** von **Y** ist für alle *Xji* gleich ; es ist gleich ÿ2 (unbekannt).
- Die unterschiedlichen Beobachtungen zu **Y** sind unabhängig voneinander abhängig, aber durch die Werte von *Xji bedingt.*

Varianzanalyse in Matrixform

In der Matrixform sieht die Tabelle der **Varianzanalyse** für die Regression wie folgt aus:

Varianzanalyse

Quelle von Variation	Grade von Freiheit	Die Summe von Quadrate	Bedeutung von Quadrate
Regression	*p* ÿ 1	ÿˆ **XY** ÿ ny¯2	$\frac{\ddot{y}\hat{}\ \mathbf{XY}\ \ddot{y}\ \mathbf{nY}\hat{}\ 2}{p\ \ddot{y}\ 1}$
Rest *n* ÿ *p*		**YY** ÿ ÿˆ **XY**	$\frac{\mathbf{YY}\ \ddot{y}\ \ddot{y}\hat{}\ \mathbf{XY}}{n\ \ddot{y}\ p}$
Gesamt	*n* ÿ 1	**Y Yÿny¯2**	

Wenn das **Modell** korrekt ist, dann

$$S2 = \frac{\mathbf{YY}\ \ddot{y}\ \ddot{y}\hat{}\ \mathbf{XY}}{n\ \ddot{y}\ p}$$

ist ein erwartungstreuer **Schätzer** von ÿ2.

Die Varianzanalyse ermöglicht es uns, die **Nullhypothese zu testen:**

$$H0 : \ddot{y}j = 0 \text{ für } j = 1,...,p\ \ddot{y}\ 1$$

gegen die **Alternativhypothese:**

H1: Mindestens einer der Parameterÿj, *j* = 0, ist von Null verschieden

Berechnung der **Statistik:**

$$F = \frac{EMSE}{RMSE} = \frac{EMSE}{S2} \ ,$$

M

Machine Translated by Google

Dabei ist *EMSE* der Mittelwert der Regressionsquadrate, *RMSE* der Mittelwert der Residuenquadrate und *TMSE* der Gesamtmittelwert von Quadrate.

Diese *F* -Statistik muss mit dem **Wert** Fÿ,pÿ1,nÿp der **Fisher-Tabelle verglichen werden,** wobei ÿ das **Signifikanzniveau** des Tests ist:

ÿ wenn *F* ÿ Fÿ,pÿ1,nÿp, akzeptieren wir H0;
wenn *F* > Fÿ,pÿ1,nÿp, lehnen wir H0 für H1 ab.

Das Bestimmtheitsmaß R2 wird berechnet auf folgende Weise durchgeführt:

$$R2 = \frac{ESS}{TSS} = \frac{\ddot{y}\hat{} \mathbf{XY} \ddot{y}\, ny^{-}2}{\mathbf{YY} \ddot{y}\, ny^{-}2},$$

Dabei ist ESS die Summe der Quadrate der Regression und *TSS* die Gesamtsumme der Quadrate.

DOMÄNEN UND EINSCHRÄNKUNGEN

Siehe **Analyse der Regression.**

BEISPIELE

Die folgenden Daten betreffen zehn Unternehmen der chemischen Industrie. Wir versuchen, einen Zusammenhang zwischen der Produktion, den Arbeitsstunden und dem Kapital herzustellen.

Produktionsarbeit		Hauptstadt
(100 Tonnen)	(H)	(Maschinenstunden)
60	1100	300
120	1200	400
190	1430	420
250	1500	400
300	1520	510
360	1620	590
380	1800	600
430	1820	630
440	1800	610
490	1750	630

Das Matrixmodell **wird** wie folgt geschrieben:

$$\mathbf{Y} = \mathbf{X}\ddot{y} + \ddot{y} \quad ,$$

Wo

$$\mathbf{Y} = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} 60 \\ 120 \\ 190 \\ 250 \\ 300 \\ 360 \\ 380 \\ 430 \\ 440 \\ 490 \end{matrix} \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix}, \quad \mathbf{X} = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} 1\ 1100\ 300 \\ 1\ 1200\ 400 \\ 1\ 1430\ 420 \\ 1\ 1500\ 400 \\ 1\ 1520\ 510 \\ 1\ 1620\ 590 \\ 1\ 1800\ 600 \\ 1\ 1820\ 630 \\ 1\ 1800\ 610 \\ 1\ 1750\ 630 \end{matrix} \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix},$$

$$\ddot{y} = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} \ddot{y}1 \\ \ddot{y}2 \\ \ddot{y}3 \\ \ddot{y}4 \\ \ddot{y}5 \\ \ddot{y}6 \\ \ddot{y}7 \\ \ddot{y}8 \\ \ddot{y}9 \\ \ddot{y}10 \end{matrix} \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix}, \quad \ddot{\mathbf{y}} = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} \ddot{y}0 \\ \ddot{y}1 \\ \ddot{y}2 \end{matrix} \begin{matrix} \\ \\ \ddot{y}\ \ddot{y}. \end{matrix}$$

Die Gleichungen, die das Modell bilden, sind die folgenden:

$$60 = \ddot{y}0 + 1100\ddot{y}1 + 300\ddot{y}2 + \ddot{y}1,$$

$$120 = \ddot{y}0 + 1200\ddot{y}1 + 400\ddot{y}2 + \ddot{y}2,$$

...

$$490 = \ddot{y}0 + 1750\ddot{y}1 + 630\ddot{y}2 + \ddot{y}10.$$

Wir berechnen die Schätzer ÿˆ mit Ergebnis:

$$\ddot{y}\hat{} = \mathbf{XX}^{\ddot{y}1}\ \mathbf{XY},$$

mit

$$(\mathbf{X}\ X) = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} 10 \\ 15540 \\ 5090 \end{matrix} \begin{matrix} 15540 \\ 24734600 \\ 8168700 \end{matrix} \begin{matrix} 5090 \\ 8168700 \\ 2720500 \end{matrix} \quad \ddot{y}\ \ddot{y},$$

$$(\mathbf{X}\ X)^{\ddot{y}1} = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} 6{,}304288 \\ \ddot{y}0{,}007817 \\ 0{,}011677 \end{matrix} \begin{matrix} \ddot{y}0{,}007817 \\ 0{,}000015 \\ \ddot{y}0{,}000029 \end{matrix} \begin{matrix} 0{,}011677 \\ \ddot{y}0{,}000029 \\ 0{,}000066 \end{matrix} \quad \ddot{y}\ \ddot{y},$$

Machine Translated by Google

Und

$$\mathbf{XY} = \begin{pmatrix} 3020 \\ 5012000 \\ 1687200 \end{pmatrix}.$$

Das gibt:

$$\hat{\gamma} = \begin{pmatrix} \gamma \\ 0{,}283 \\ 0{,}591 \end{pmatrix} = \begin{pmatrix} \gamma\hat{\gamma}0\hat{\gamma} \\ 1 \end{pmatrix}.$$

Somit erhalten wir die Gleichung für die geschätzte Linie:

$$\hat{Y}i = \gamma 439{,}269 + 0{,}283Xi1 + 0{,}591Xi2 .$$

Varianzanalyse in Matrixform

Die Berechnung verschiedener Elemente der Tabelle gibt uns

1. Berechnung der Freiheitsgrade:

$$\text{dlreg} = p - 1 = 3 - 1 = 2$$
$$\text{dlres} = n - p = 10 - 3 = 7$$
$$\text{dltot} = n - 1 = 10 - 1 = 9$$

Dabei sind dlreg die Freiheitsgrade der Regression, dlres die Freiheitsgrade der Residuen und dltot die gesamten Freiheitsgrade.

2. Berechnung der Quadratsumme:

$$ESS = \hat{\gamma}'\mathbf{XY} - n\bar{Y}^2 = [\gamma 439{,}269 \; 0{,}283 \; 0{,}591] \cdot \begin{pmatrix} 3020 \\ 5012000 \\ 1687200 \end{pmatrix} - 912040 = 179063{,}39 .$$

$$TSS = \mathbf{YY} - n\bar{Y}^2 = [60 \; \dots \; 490] \cdot \begin{pmatrix} 60 \\ \vdots \\ 490 \end{pmatrix} - 912040 = 187160 .$$

$$RSS = \mathbf{YY} - \hat{\gamma}'\mathbf{XY} = TSS - ESS = 187160{,}00 - 179063{,}39 = 8096{,}61 .$$

Wir erhalten die folgende Tabelle:

Varianzanalyse

Quelle von Variation	Grad frei dom	Die Summe von Quadrate	Bedeutung von Quadrate
Regression	2	179063,39	89531,70
Rest	7	8096,61	1156,66 = S2
Gesamt	9	187160,00	

Dies ermöglicht uns, die **Nullhypothese** zu testen :

$$H0 : \gamma 1 = \gamma 2 = 0 .$$

Der berechnete **Wert** von *F* ist:

$$F = \frac{EMSE}{RMSE} = \frac{EMSE}{S2} = \frac{89531.70}{1156{,}66} = 77{,}41 .$$

Der Wert von F in **der Fisher-Tabelle** mit einem **Signifikanzniveau** $\gamma = 0{,}05$ beträgt

$$F0{,}05{,}2{,}7 = 4{,}74 .$$

Folglich wird die **Nullhypothese** H0 abgelehnt. Die **Alternativhypothese** kann eine der drei folgenden Möglichkeiten sein:

$$H1 : \gamma 1 = 0 \text{ und } \gamma 2 = 0$$
$$H2 : \gamma 1 = 0 \text{ und } \gamma 2 = 0$$
$$H3 : \gamma 1 = 0 \text{ und } \gamma 2 = 0$$

M

Machine Translated by Google

Wir können das **Bestimmtheitsmaß** R2 berechnen :

$$R2 = \frac{ESS}{TSS} = \frac{179063.39}{187160,00} \ddot{y}= 0,96 .$$

WEITERLESEN

Analyse von Residuen
Varianzanalyse
Bestimmtheitsmaß
Kollinearität
Korrelationskoeffizient
Hutmatrix
Kleinsten Quadrate
Hebelpunkt
Normale Gleichungen
Regressionsanalyse
Restwert
Einfache lineare Regression

VERWEISE

Weiseberg, S.: Angewandte lineare Regression. Wiley, New York (2006)

Chatterjee, P., Hadi, AS: Regressionsanalyse anhand von Beispielen. Wiley, New York (2006)

Machine Translated by Google

Negative Binomialverteilung

Eine **Zufallsvariable** *X* folgt einem Negativ Binomialverteilung mit Parameternsk und *r* , wenn seine **Wahrscheinlichkeitsfunktion** die Form hat:

$$P(X = k) = C_{k+r-1}^{k} \cdot p^r \cdot q^k ,$$

Dabei **ist** *p* die Erfolgswahrscheinlichkeit und *q* die Ausfallwahrscheinlichkeit und C_{N}^{x} ist die Zahl von Kombinationen von *x* Objekten unter n.
In der Situation, in der ein „Erfolg-Misserfolg“

Wird ein **Zufallsexperiment** unabhängig voneinander wiederholt, ist die Erfolgswahrscheinlichkeit **gleich** mit *p* und die Ausfallwahrscheinlichkeit mit *q* = 1 ÿ p bezeichnet. Das **Experiment** wird so oft wie nötig wiederholt *r* Erfolge erzielen . Die Anzahl der erhaltenen

Die Anzahl der Fehler, bevor dieses Ziel erreicht wird, ist eine **Zufallsvariable, die** einer oben beschriebenen negativen Binomialverteilung folgt. Das negative
Die Binomialverteilung ist eine **diskrete Binomialverteilung Verteilung.**

GESCHICHTE

Der erste, der die negative Binomialverteilung behandelte, war Pascal, Blaise (1679). Dann de Montmort, PR (1714) wendete das Negativ an Binomialverteilung, die angibt, wie oft eine Münze geworfen werden sollte
eine bestimmte Anzahl an Köpfen erhalten. Student (1907) verwendeten die negative Binomialverteilung Bution als Alternative **zur Poisson-Distri Bution.**

Greenwood, M. und Yule, GU (1920) und Eggenberger, F. und Polya, G. (1923) gefunden Anwendungen der negativen Binomialverteilung. Seitdem gibt es immer mehr Anwendungen dieser Verteilung und der darauf basierenden statistischen Techniken
zu dieser Distribution wurden in entwickelt ein paralleler Weg.

MATHEMATISCHE ASPEKTE

Wenn X1, *X2,...,Xn n* unabhängige **Zufallszahlen** sind **Variablen ,** die einer **geometrischen Verteilung folgen,** dann die Zufallsvariable

$$X = X1 + X2 + ... + Xr$$

folgt einer negativen Binomialverteilung.
Um den **erwarteten** Wert von X zu berechnen, wird die folgende Eigenschaft verwendet, wobei *Y* und *Z* gleich sind **zufällige Variablen:**

$$E[Y + Z] = E[Y] + E[Z] .$$

Wir haben daher:

$$\begin{aligned} E[X] &= E[X1 + X2 + ... + Xr] \\ &= E[X1] + E[X2] + ... \\ &\quad + E[Xr] \\ &= \frac{Q}{P} + \frac{Q}{P} + ... + \frac{Q}{P} \\ &= r \cdot \frac{Q}{P} . \end{aligned}$$

Machine Translated by Google

Um die **Varianz** von X zu berechnen, wird die folgende Eigenschaft verwendet, wobei *Y* und *Z* unabhängige **Variablen sind:**

$$\mathrm{Var}(Y + Z) = \mathrm{Var}(Y) + \mathrm{Var}(Z) .$$

Wir haben daher:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(X1 + X2 + \ldots + Xr) \\ &= \mathrm{Var}(X1) + \mathrm{Var}(X2) + \ldots \\ &\quad + \mathrm{Var}(Xr) \\ &= \frac{Q}{p2} + + \frac{Q}{\ldots} + p2\ p2 \quad \frac{Q}{} \\ &= r \cdot \frac{Q}{p2} .\end{aligned}$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Zu den spezifischen Feldern, für die die negative Binomialverteilung angewendet wurde sind **Unfallstatistik,** Biowissenschaften, Ökologie, Marktstudien, medizinische Forschung, und Psychologie.

BEISPIELE

Eine Münze wird mehrmals geworfen. Uns interessiert die **Wahrscheinlichkeit ,** Kopf zu bekommen ein drittes Mal beim sechsten Wurf.
Wir haben dann:

Anzahl Erfolge: $r = 3$

Anzahl der Ausfälle: $k = 6 \ddot{y} r = 3$

Wahrscheinlichkeit eines Erfolgs: $p = \frac{1}{2}$

(Köpfe erhalten)

Wahrscheinlichkeit eines Ausfalls: $q = \frac{1}{2}$

(Schwänze erhalten)

Die **Wahrscheinlichkeit ,** zuvor *k* Schwänze zu erhalten Der dritte Kopf ist gegeben durch:

$$P(X = \mathrm{k}) = Ck_{3+k\ddot{y}1} \cdot \left(\frac{1}{2}\right)^{3} \cdot \left(\frac{1}{2}\right)^{k} .$$

Die Wahrscheinlichkeit , einen dritten zu bekommen, steigt der sechste Wurf, also die Wahrscheinlichkeit von Erhalten Sie dreimal Schwänze, bevor Sie sie erhalten die dritten Köpfe, ist daher gleich:

$$\begin{aligned}P(X = 3) &= C3_{3+3\ddot{y}1} \cdot \frac{1}{23} \cdot \frac{1}{23} \\ &= C3_5 \cdot \frac{1}{23} \cdot \frac{1}{23} \\ &= \frac{5!}{3!(5 \ddot{y} 3)!} \cdot \frac{1}{23} \cdot \frac{1}{23} \\ &= 0{,}1563 .\end{aligned}$$

WEITERLESEN

Bernoulli-Verteilung
Binomialverteilung
Diskrete Wahrscheinlichkeitsverteilung
Poisson-Verteilung

VERWEISE

Eggenberger, F. und Polya, G.: Über die Statistik verketteter Vorgänge. Zeitschrift für angewandte Mathematik und Mechanik 3, 279–289 (1923)

Greenwood, M., Yule, GU: Eine Untersuchung zu die Art der Häufigkeitsverteilungen, die für mehrere Ereignisse mit repräsentativ sind besonderer Hinweis auf das Auftreten von mehrfache Krankheitsanfälle oder wiederholte Unfälle. J. Roy. Stat. Soc. Ser. A **83,** 255–279 (1920)

Montmort, PR de: Essai d'analyse sur les Jeux de Hasard, 2. Aufl. Quillau, Paris (1713)

Pascal, B.: Varia Opera Mathematica. D. Petri de Fermat. Tolosae (1679)

Gosset, SW („Student"): Zum Fehler der Zählung mit einem Hämozytometer. Biometrie 5, 351–360 (1907)

Machine Translated by Google

Nelder, John A.

John Nelder wurde 1924 in Dulver geboren
Tonne (Somerset, England). Nach Erhalt seines
Nach seinem Diplom in mathematischer Statistik in
Cam Bridge wurde er zum Präsidenten der Section of
Statistics des National Plant ernannt
Forschungsstation in Wellesbourn, eine Stelle
Er hatte das Amt von 1951 bis 1968 inne. Er erhielt die Auszeichnung
Titel eines Doktors der Naturwissenschaften in Birmingham
und wurde dann zum Präsidenten der Statistikabteilung
von Rothamsted Experimental gewählt
Station in Harpenden von 1968 bis 1984. Er
ist jetzt Gastprofessor am Imperial College of Science,
Technology and Medicinein
London. Er wurde 1984 zum Mitglied der Royal Society
gewählt und war deren Präsident
renommierte Gesellschaften: die International Bio
metric Society und die Royal Statistical Society.

Nelder ist der Autor der heute verwendeten statistischen
Rechensysteme Genstat und GLIM
in mehr als 70 Ländern. Autor von mehr
mehr als 140 Artikel in statistischen und
Über den Autor und weitere *Mitwirkende*

McCullagh von einem Buch, das das Verallgemeinerte behandelt
lineares Modell. Auch Nelder entwickelte die Idee
von allgemein ausgewogenen Designs.

Ausgewählte Werke und Artikel von John A. Nelder:

1965 (mit Mead, R.) Eine Simplex-Methode
zur Funktionsminimierung. Computational
Journal, 7, 303–333.
1971 Computer in der Biologie. Wykeham,
London und Winchester, S. 149.
1972 (mit Wedderburn, RWM) Generalisierte lineare
Modelle. J. Roy. Stat. Soc.
Ser. A 135, 370–384.
1974 Genstat: ein statistisches System. In:
COMPSTAT: Verfahren in Com
putationale Statistik. Physica, Wien
n / A.
1983 (mit McCullagh, P.) Verallgemeinert
Lineare Modelle. Chapman & Hall,
London, S. 261.

WEITERLESEN

Statistiksoftware

Newcomb, Simon

Newcomb, Simon wurde 1835 in Nova geboren
Scotia (Kanada) und starb 1909. Dieser Astronom
kanadischer Herkunft trug dazu bei
Behandlung von Ausreißern in der Statistik, zur
Anwendung der Wahrscheinlichkeitstheorie für die
Interpretation von Daten und zur Entwicklung dessen
Wir nennen heute robuste Methoden in der Statistik.
Bis zu seinem 16. Lebensjahr lernte er hauptsächlich
die Bücher konsultieren, die sein Vater gefunden hat
ihn. Danach begann er ein Medizinstudium
mit Plänen, Arzt zu werden. Im Jahr 1857 war er
war mit der amerikanischen Ephemeride beschäftigt und
Nautischer Almanach in Cambridge im Bundesstaat
ofMassachussetts. Gleichzeitig studierte er in Harvard
und schloss sein Studium 1858 ab
1861 wurde er Professor für Mathematik
am Marineobservatorium.
Die Sammlung von Notizen zur Theorie von
Wahrscheinlichkeiten (Mathematics Monthly, 1859–
61) stellt ein Werk dar, das auch heute noch als
modern gilt. Newcombs wichtigste Bemerkung
Sein Ansatz ist es, einen wertvollen Beitrag zur Statistik zu leisten
zur Behandlung von Ausreißern in der Astronomie
Daten; Bestätigung, dass die Normalverteilung
passt nicht, er hat die normale kontaminierte Verteilung
erfunden.

Ausgewählte Werke und Artikel von Newcomb, Simon:

1859–61 Anmerkungen zur Wahrscheinlichkeitstheorie.
Mathematics Monthly, 1, 136–

Machine Translated by Google

139, 233–235, 331–355, 349–350; 2, 134–140, 272–275; 3, 68, 119–125, 341–349.

Neyman, Jerzy

Neyman, Jerzy (1894–1981) war einer der Begründer der modernen **Statistik.** Einer von seinen herausragende Beiträge wurde festgestellt, **mit Pearson, Egon Sharpe, die** Grundlage der Theorie des Hypothesentests. Geboren in Russland, Neyman, Jerzy studierte Physik und Mathematik an der Universität Charkow. Im Jahr 1921 war er ging nach Polen, seinem Herkunftsland, wo er eine Zeit lang als Statistiker am Nationalen Institut für Landwirtschaft in Bydgoszcz arbeitete. Im Jahr 1924 gab er einige aus
Zeit in London, wo er am University College unter der Leitung von **Pears Sohn Karl studieren konnte.** Zu dieser Zeit lernte er Pearson und Egon kennen Sharpe **Gosset, William Sealy** und **Fisher, Ronald Aylmer.** Im Jahr 1937, Ende
Nach einer Reise in die Vereinigten Staaten, wo er auf zahlreichen Konferenzen Vorträge hielt, nahm er die Stelle eines Professors an der University of California in Berkeley an. Er hat das geschaffen Fakultät für Statistik und beendete seine glänzende Karriere an dieser Universität.

Ausgewählte Artikel von Neyman, Jerzy:

1928 (mit Pearson, ES) Über die Verwendung und Interpretation bestimmter Testkriterien zum Zwecke statistischer Schlussfolgerungen, I, II. Biometrie **20A,** 175–240, S. 263–295.

1933 (mit Pearson, ES) Prüfung statistischer Hypothesen in Bezug auf Wahrscheinlichkeiten a priori. Proz. Camb. Philos. Soc. **29,** 492–510.

1934 Über die zwei verschiedenen Aspekte der repräsentative Methode: die Methode von geschichtete Probenahme und die Methode von gezielte Auswahl. J. Roy. Stat. Soc. **97,** 558–606, Diskussion S. 607–625.

1938 Beitrag zur Theorie der Stichprobenziehung menschlicher Populationen. Marmelade. Stat. Assoc. **33,** 101–116.

WEITERLESEN

Hypothesentest

Nichtlineare Regression

Eine **Regressionsanalyse, bei der** die **abhängige Variable** *Y* von einer oder mehreren abhängt Die unabhängigen Variablen *X1,...,Xk* werden aufgerufen eine nichtlineare Regression, wenn die Gleichung von Regression $Y = f\,X1,...,Xk; ÿ0,..., ÿp$ ist nicht linear in den Parametern *ÿ0,..., ÿp.*

GESCHICHTE

Die nichtlineare Regression stammt aus den 1920er Jahren an **Fisher, Ronald Aylmer** und Mackenzie,WADie Verwendung ist jedoch detaillierter

Die Untersuchung dieser Modelle musste warten für Fortschritte bei automatischen Berechnungen im 1970er Jahre.

MATHEMATISCHE ASPEKTE

Seien Y1, *Y2,...,Yr* und *X11,...,X1r, X21,...,X2r,..., Xk1,...,Xkr* seien die *Beobachtungen* (bzw.) des **Abhängigen Variable** *Y* und die unabhängigen Variablen *X1,...,Xk.*
Ziel ist die Schätzung der Parameter *ÿ0,..., ÿp* des Modells:

$$Yi = f\,X1i,...,Xki; ÿ0,..., ÿp + ÿi\,, \qquad i = 1,...,r.$$

Machine Translated by Google

Mit der Methode der **kleinsten Quadrate** und unter der entsprechenden Fehlerhypothese (siehe **Residuenanalyse)** schätzen wir die Parameter *ÿ0,..., ÿp* anhand der Werte ÿˆ *0,..., ÿˆ*, die die Quadratsumme minimieren Fehler, bezeichnet mit *S*

$$S_{\ddot{y}0,\ldots,\ddot{y}p} = \sum_{i=1}^{R} \left[Yi \ddot{y} f\,X1i,\ldots,Xki;\ddot{y}0,\ldots,\ddot{y}p \right]^2 :$$

$$\min_{\ddot{y}0,\ldots,\ddot{y}p} \sum_{i=1}^{R} \ddot{y}2 = \min S_{\ddot{y}0,\ldots,\ddot{y}p} \,.\, \ddot{y}0,\ldots,\ddot{y}p$$

Iff ist nichtlinear, die Auflösung der **Normalgleichungen** (von denen keine linear ist):

$$\sum_{i=1}^{R} \left[Yi \ddot{y} f\,X1i,\ldots,Xki;\ddot{y}0,\ldots,\ddot{y}p \right] . \frac{\ddot{y}f\,X1i,\ldots,Xki;\ddot{y}0,\ldots,\ddot{y}p}{\ddot{y}\ddot{y}j} = 0$$

(j = 0,...,p) kann schwierig sein.

Das Problem besteht darin, dass die gegebenen Gleichungen nicht unbedingt eine Lösung haben oder mehr als eine haben.

Im Folgenden diskutieren wir die verschiedenen Ansätze zur Lösung des Problems.

1. *Gauß–Newton*

   Das iterative Verfahren besteht in der Linearisierung von *f* mit Hilfe der Taylor-Entwicklung bei der sukzessiven Schätzung von Parametern durch lineare **Regression.**

   Wir werden diese Methode im folgenden Beispiel näher erläutern .

2. *Steilster Abstieg*

   Die Idee dieser Methode besteht darin, die Parameter *ÿj* für Schätzungen näherungsweise festzulegen und dann eine Näherung hinzuzufügen

   $$\ddot{y}\ddot{y}S_{\ddot{y}0,\ldots,\ddot{y}p}\,\ddot{y}\ddot{y}j$$

   das entspricht dem Maximum des Abstiegs von der Funktion S.

   Iterativ ermitteln wir die Parameter, die S minimieren.

3. *Marquard-Levenburg* Diese hier nicht entwickelte Methode integriert die Vorteile zweier genannter Verfahren und erzwingt und beschleunigt die Konvergenz der Approximationen der zu schätzenden Parameter.

DOMÄNEN UND EINSCHRÄNKUNGEN

Bei der **multiplen linearen Regression** gibt es Funktionen, deren Parameter nicht linear sind, die aber nach einer Transformation linear werden.

Ist dies nicht der Fall, nennen wir diese Funktionen nichtlinear und behandeln sie mit nichtlinear Methoden.

Ein Beispiel für eine nichtlineare Funktion ist:

$$Y = \frac{\ddot{y}1}{\ddot{y}1 \ddot{y} \ddot{y}2} \exp(\ddot{y}\ddot{y}2X) \ddot{y} \exp(\ddot{y}\ddot{y}1X) + \ddot{y} .$$

BEISPIELE

Es seien P1 = (x1, y1), P2 = (x2, y2) und P3 = (x3, y3) drei Punkte in einem Diagramm und d1, d2 und d3 die Näherungsabstände (die „Annäherungen“ werden entsprechend verteilt). (zur Gauß-Verteilung der gleichen Varianz) zwischen den Datenpunkten und einem Punkt *P = (x,* y) unbekannter und gesuchter Koordinaten.

Ein mögliches Regressionsmodell **zur** Schätzung der Koordinaten von *P* sieht wie folgt aus:

$$di = f\,(xi, yi; x, \mathrm{y}) + \ddot{y}i\,, \qquad ich = 1, 2, 3\,,$$

wobei die Funktion *f* den Abstand zwischen Punkt *Pi* und Punkt P darstellt, das heißt:

$$f\,(xi, yi; x, \mathrm{y}) = \sqrt{(x \ddot{y} xi)^2 + (y \ddot{y} yi)^2}\,.$$

N

Machine Translated by Google

Die Distanzfunktion ist eindeutig nichtlinear und kann nicht in eine lineare Form umgewandelt werden, daher ist die entsprechende Regression nichtlinear. Wir wenden die Methode von *Gauß-Newton an.* Betrachtet man *f* als Funktion der zu schätzenden Parameter *x* und *y* , ist die Taylor-Entwicklung von *f* , bis es zum linearen Term in (x0, y0) wird, gegeben durch:

$$f(xi, yi; x, y) \approx f(xi, yi; x0, y0) + \left.\frac{\partial f(xi, yi; x, y)}{\partial x}\right|_{\substack{x=x0\\y=y0}} \cdot (x - x0) + \left.\frac{\partial f(xi, yi; x, y)}{\partial x}\right|_{\substack{x=x0\\y=y0}} \cdot (y - y0),$$

Wo

$$\left.\frac{\partial f(xi, yi; x, y)}{\partial x}\right|_{\substack{x=x0\\y=y0}} = \frac{xi - x0}{f(xi, yi; x0, y0)}$$

Und

$$\left.\frac{\partial f(xi, yi; x, y)}{\partial x}\right|_{\substack{x=x0\\y=y0}} = \frac{yi - y0}{f(xi, yi; x0, y0)}.$$

Solet(x0, y0) sei eine erste Schätzung *von P* , beispielsweise geometrisch gefunden.
Das linearisierte Regressionsmodell wird somit ausgedrückt durch:

$$di = f(xi, yi; x0, y0) + \frac{xi - x0}{f(xi, yi; x0, y0)} \cdot (x - x0) - \frac{yi - y0}{f(xi, yi; x0, y0)} \cdot (y - y0) + \varepsilon i, + f(xi, yi; \quad i = 1, 2, 3.$$

Mit Hilfe einer **multiplen linearen Regression** (ohne Konstanten) können wir die Parameter *(x,* y) = *(x* x0, *y* y0) von schätzen das folgende Modell (die „Beobachtungen“ des Modells erscheinen in Klammern):

$$di - f(xi, yi; x0, y0) = x \cdot \frac{xi - x0}{f(xi, yi; x0, y0)} + y \cdot \frac{yi - y0}{f(xi, yi; x0, y0)} + \varepsilon i, \quad i = 1, 2, 3.$$

Lassen $\hat{X}, \hat{y}$ sei die gefundene Schätzung von (x, y). Dadurch erhalten wir eine bessere Schätzung der Koordinaten von *P* durch

x0 + $\hat{x}$, y0 + $\hat{y}$.

Wenn wir diesen neuen Punkt als Ausgangspunkt verwenden, können wir das gleiche Verfahren erneut anwenden. Die übrigen Koordinaten werden konvergierend zum gewünschten Punkt gefunden, also zu dem Punkt, der sich am besten annähert.
Nehmen wir nun ein konkretes Beispiel: P1 = (0, 5); P2 = (5, 0); P3 = (10, 10); d1 = 6; d2 = 4; d3 = 6.
Wenn wir als Anfangspunkt *des Gauß-Newton-*Verfahrens den folgenden Punkt wählen: *P* = $\frac{0+5+10}{3}, \frac{5+0+10}{3}$ = (5, 5), erhalten wir die 3 Satz der folgenden Koordinaten:

(5, 5)

(6.2536, 4.7537)

(6.1704, 4.7711)

(6.1829, 4.7621)

(6.1806, 4.7639)

das konvergiert zum Punkt (6.1806,4.7639).
Die Abstände zwischen dem gefundenen Punkt und P1, P2 und P3 betragen 6,1856, 4,9078 bzw. 6,4811.

WEITERLESEN

Kleinsten Quadrate

Machine Translated by Google

Modell

Multiple lineare Regression

Normale Gleichungen

Regressionsanalyse

Einfache lineare Regression

VERWEISE

Bates, DM, Watts DG: Nichtlineare Regressionsanalyse und ihre Anwendungen. Wiley, New York (1988)

Draper, NR, Smith, H.: Applied Regression Analysis, 3. Aufl. Wiley, New York (1998)

Fisher, RA: Über die mathematischen Grundlagen der theoretischen Statistik. Philos. Trans. Roy. Soc. London. Ser. A **222,** 309–368 (1922)

Fisher,RA,Mackenzie,WA:Themanuri al-Reaktion verschiedener Kartoffelsorten. J. Landwirtschaft. Wissenschaft. **13,** 311–320 (1923)

Fletcher, R.: Praktische Optimierungsmethoden. Wiley, New York (1997)

Seber, AF, Wilt, SC: Nichtlineare Regression. Wiley, New York (2003)

Nichtparametrische Statistik

Statistische Verfahren, die es uns ermöglichen, Daten aus kleinen Stichproben zu Variablen zu verarbeiten worüber nichts bekannt ist deren Verbreitung. Konkret wurden nichtparametrische statistische Methoden entwickelt Wird in Fällen verwendet, in denen der Forscher es weiß nichts über die Parameter der interessierenden Variablen in der Bevölkerung (daher der Name *nichtparametrisch).* Nichtparametrisch Methoden basieren nicht auf der Schätzung von Parameter (wie der Mittelwert oder die Standardabweichung), die die Verteilung beschreiben der interessierenden Variablen in der Bevölkerung.

Nichtparametrische Modelle unterscheiden sich von parametrischen Modellen dadurch, dass es keine Modellstruktur gibt a priori spezifiziert, sondern determiniert wird aus Daten. Daher sind diese Methoden auch manchmal auch als *„parameterfreie Methoden" bezeichnet Verteilungsfreie Methoden.*

GESCHICHTE

Der Begriff *nichtparametrisch* wurde erstmals von verwendet Wolfowitz, 1942.

Siehe auch **nichtparametrischer Test.**

DOMAIN UND EINSCHRÄNKUNGEN

Die nichtparametrische Methode unterscheidet sich von der Analyse eines Einwegklassifizierungsmodells zum Vergleich von Behandlungen mit Regression und Probleme bei der Kurvenanpassung.

Die am häufigsten verwendete nichtparametrische Methode Tests sind **Anderson-Darling-Test, Chi-Quadrat-Test, Kendalls Tau, Kolmogorov-Smirnov-Test, Kruskall-Wallis-Test, Wil-Coxon-Rangsummentest und Spearmans Rang Korrelationskoeffizient und Wilcoxon-Zeichen Rangtest.**

Nichtparametrische Tests haben weniger Aussagekraft als die geeignete parametrische Tests, sind aber mehr robust, wenn die Annahmen zugrunde liegen Parametertest sind nicht erfüllt.

BEISPIELE

Ein Histogramm ist eine einfache nichtparametrische Schätzung einer Wahrscheinlichkeitsverteilung. Eine Verallgemeinerung des Histogramms ist die Kernelglättung Technik, mit der aus einem bestimmten Datensatz eine sehr glatte Wahrscheinlichkeitsdichtefunktion konstruiert werden kann.

Siehe auch **Nichtparametrische Tests.**

Machine Translated by Google

WEITERLESEN

Nichtparametrischer Test

VERWEISE

Parzen, E.: Zur Schätzung einer Wahrscheinlichkeitsdichtefunktion und -modus. Ann. Mathematik. Stat. **33,** 1065–1076 (1962)

Savage, R.: Bibliographie der nichtparametrischen Statistik und verwandte Themen: Einführung. Marmelade. Stat. Assoc. **48,** 844–849 (1953)

Gibbons, JD, Chakraborti, S.: Nonpara metric Statistical Inference, 4. Auflage. CRC (2003)

Nichtparametrischer Test

Ein nichtparametrischer Test ist eine Art **Hypothesentest,** bei dem es nicht erforderlich ist, die Form der Verteilung der untersuchten **Grundgesamtheit** anzugeben . Auf jeden Fall sollten wir unabhängige Beobachtungen haben, das heißt, dass die Auswahl der Personen, die die **Stichprobe** bilden , keinen Einfluss auf die Auswahl anderer Personen haben darf.

GESCHICHTE Der erste nichtparametrische Test erschien in den Werken von Arbuthnott, J. (1710), der den **Vorzeichentest** einführte . Die meisten nichtparametrischen Tests wurden jedoch zwischen 1940 und 1955 entwickelt.

Wir erwähnen besonders die Artikel von Kolmogorov, Andrey Nikolaevich (1933), Smirnov, Vladimir Ivanovich (1939), Wilcoxon, F. (1945, 1947), Mann, HB und Whitney, DR (1947), Mood, AM (1950).) und Kruskal, WH und Wallis, WA (1952).

Später wurden dieser Liste viele weitere Artikel hinzugefügt. Savage, IR (1962) veröffentlichte eine Bibliog Zusammenfassung von etwa 3000 Artikeln, die vor 1962 verfasst wurden und sich mit nichtparametrischen Tests befassen.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die schnelle Entwicklung nichtparametrischer Tests kann durch die folgenden Punkte erklärt werden: • Nichtparametrische Methoden erfordern wenige Annahmen über die untersuchte **Population ,** wie beispielsweise die Annahme einer Normalität.

• Nichtparametrische Methoden sind oft einfacher zu verstehen und anzuwenden als die entsprechenden parametrischen Tests.

• Nichtparametrische Methoden sind in Situationen anwendbar, in denen parametrische Tests nicht verwendet werden können, beispielsweise wenn die Variablen nur auf Ordinalskalen gemessen werden. • Trotz der Tatsache, dass nichtparametrische Methoden auf den ersten Blick einen wesentlichen Teil der in den Stichproben enthaltenen Informationen zu opfern scheinen, haben theoretische Untersuchungen gezeigt, dass nichtparametrische Tests ihren parametrischen Gegenstücken nur geringfügig unterlegen sind, wenn die Verteilung der untersuchten Population gleich ist spezifiziert beispielsweise die **Normalverteilung.** Nichtparametrische Tests hingegen sind parametrischen Tests überlegen, wenn die Verteilung der Grundgesamtheit weit von der vorgegebenen (Normal-)Verteilung entfernt ist.

BEISPIELE

Siehe **Kolmogorov-Smirnov-Test, Kruskal-Wallis-Test, Mann-Withney-Test, Wilcox-Test mit Vorzeichen, Vorzeichentest.**

WEITERLESEN

Anpassungstest

Hypothesentest

Kolmogorov-Smirnov-Test

Kruskal-Wallis-Test

Machine Translated by Google

Mann-Whitney-Test
Zeichentest
Rangkorrelationskoeffizient nach Spearman
Test der Unabhängigkeit
Von Wilcoxon unterzeichneter Test
Wilcoxon-Test

VERWEISE

Arbuthnott, J.: Ein Argument für die göttliche Vorsehung, basierend auf der konstanten Regelmäßigkeit, die bei den Geburten beider Geschlechter beobachtet wird. Phi los. Trans. **27,** 186–190 (1710)(3.4).

Kolmogorov, AN: Die empirische Bestimmung einer Vertriebslinie.
Giornale dell'Instituto Italiano degli Attuari 4, 83–91 (6.1) (1933)

Kruskal, WH, Wallis, WA: Verwendung von Rängen in der Einkriterium-Varianzanalyse. Marmelade. Stat. Assoc. **47,** 583–621 und Errata, ebenda. **48,** 907–911 (1952)

Mann, HB, Whitney, DR: Bei einem Test, ob eine von zwei Zufallsvariablen vorliegt stochastisch größer als die anderen. Ann. Mathematik. Stat. **18,** 50–60 (1947)

Mood, AM: Einführung in die Theorie der Statistik. McGraw-Hill, New York (1950) (Kap. 16).

Savage, IR: Bibliographie der nichtparametrischen Statistik. Harvard University Press, Cambridge, MA (1962)

Smirnov, NV: Schätzung der Abweichung zwischen empirischen Verteilungsfunktionen in zwei unabhängigen Stichproben. (Russisch). Stier. Moskauer Universität. 2(2), 3–16 (6.1, 6.2) (1939)

Wilcoxon, F.: Einzelvergleiche nach Ranking-Methoden. Biometrie 1, 80–83 (1945)

Wilcoxon, F.: Einige schnelle ungefähre statistische Verfahren. Amerikanisches Cyanamid, Stamford Research Laboratories, Stamford, CT (1957)

Norm eines Vektors

Die Norm eines **Vektors** gibt die Länge dieses Vektors an, die sich aus dem **Skalarprodukt** des Vektors selbst ergibt. Die Norm eines Vektors erhält man, indem man aus diesem Skalarprodukt die Quadratwurzel zieht.

Ein Vektor der Norm 1 wird Einheitsvektor genannt.

MATHEMATISCHE ASPEKTE

Für einen gegebenen **Vektor** x,

$$x = \begin{pmatrix} x1 \\ x2 \\ \vdots \\ xn \end{pmatrix},$$

Wir definieren das **Skalarprodukt** von *x* selbst durch:

$$x \cdot x = \sum_{i=1}^{N} x2_{ch.}$$

Die Norm des Vektors *x* ist somit gleich:

$$x = \sqrt{x \cdot x} = \sqrt{\sum_{i=1}^{N} x2_{ch.}}$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Als Norm wird ein **Vektor** mit berechnet

Mit Hilfe des **Skalarprodukts ergeben sich die** Eigenschaften der Norm aus denen des Skalars Produkt. Somit gilt:

- Die Norm eines Vektors *x* ist streng positiv, wenn der Vektor nicht Null ist, und Null, wenn er:

$$x > 0, \text{ wenn } x = 0 \text{ und}$$

$$x = 0 \text{ genau dann, wenn } x = 0 .$$

N

Machine Translated by Google

- Die Norm des Ergebnisses einer Multiplikation eines Vektors x mit einem Skalar ist gleich dem Produkt der Norm von x und dem Absolutwert von der Skalar:

$$k \cdot x = |k| \cdot x .$$

- Für zwei Vektoren x und y gilt:

$$x + y \ddot{y} x + y ,$$

das heißt, die Norm einer Summe zweier Vektoren ist kleiner oder gleich der Summe der Normen dieser Vektoren.

- Für zwei Vektoren x und y ist der Absolutwert des Skalarprodukts von x und y kleiner oder gleich dem Produkt der Normen:

$$x \cdot y \ddot{y} x \cdot y .$$

Dieses Ergebnis wird als Cauchy-Schwarz-Ungleichung bezeichnet.

Zur Ermittlung wird auch die Norm eines Vektors verwendet ein Einheitsvektor mit derselben Richtung wie der gegebene Vektor. In diesem Fall genügt es, den Anfangsvektor durch seine Norm zu dividieren:

$$y = \frac{x}{x} .$$

BEISPIELE

Betrachten Sie den folgenden **Vektor ,** der in definiert ist Euklidischer Raum mit drei Dimensionen:

$$x = \begin{matrix} 12 \\ \ddot{y} \ 15 \\ \ddot{y} \ 16 \ \ddot{y} \ddot{y} . \end{matrix}$$

Das **Skalarprodukt** dieses Vektors selbst ist gleich:

$$x \cdot x = 122 + 152 + 162$$
$$= 144 + 225 + 256 = 625$$

woraus wir die Norm von x erhalten:

$$x = \ddot{y} \ x \cdot x = \ddot{y} \ 625 = \overline{25} .$$

Der Einheitsvektor mit der Richtung von x erhält man, indem man jede Komponente von x durch 25 dividiert:

$$y = \begin{matrix} 0,48 \\ \ddot{y} \ 0,6 \\ \ddot{y} \ 0,64 \ \ddot{y} \ddot{y} , \end{matrix}$$

und wir wollen überprüfen, dass $y = 1$ ist. Daher:

$$j^2 = 0,482 + 0,62 + 0,642$$
$$= 0,2304 + 0,36 + 0,4096 = 1 .$$

WEITERLESEN

Kleinsten Quadrate
L1 -Schätzung
Vektor

VERWEISE

Dodge, Y.: Basismathematik für Wirtschaftswissenschaftler. Springer, Berlin Heidelberg New York (2002)

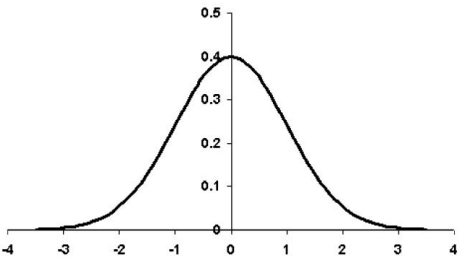
Dodge, Y., Rousson, V.: Multivariater L1-Mittelwert. Metrika **49,** 127–134 (1999)

Normalverteilung

Die Zufallsvariable X ist entsprechend verteilt zu einer Normalverteilung, wenn sie eine **Dichtefunktion** der Form hat:

$$f(x) = \exp \ddot{y} \ddot{y} \frac{1}{2\ddot{y}} \quad \frac{\ddot{y} \ (x \ddot{y} \ddot{y})^2}{2\ddot{y}2} ,$$

$$(\ddot{y} > 0) .$$



Normalverteilung, $\ddot{y} = 0$, $\ddot{y} = 1$

Machine Translated by Google

Wir werden sagen, dass *X* einer Normalverteilung des Mittelwerts ÿ und der Varianz ÿ2 folgt. Die Normalverteilung ist eine **kontinuierliche Wahrscheinlichkeitsverteilung.**

GESCHICHTE

Die Normalverteilung wird oft zugeschrieben **Laplace, PS** und **Gauss, CF,** deren Namen es trägt. Sein Ursprung geht jedoch auf die Werke von **Bernoulli, J.** zurück , der in seinem Werk *Ars Conjectandi* (1713) die ersten Grundelemente des **Gesetzes der großen Zahlen lieferte.** Im Jahr 1733 erlangte A. de **Moivre** als erster die Normalverteilung als Näherung **der Binomialverteilung.** Diese Arbeit war in lateinischer Sprache verfasst und 1738 in einer englischen Version veröffentlicht. De Moivre, A. nannte das, was er fand, eine „Kurve"; Er entdeckte diese Kurve, als er die **Gewinnwahrscheinlichkeiten** für verschiedene Glücksspiele berechnete . **Laplace, PS,** untersuchte nach **de Moivre, A.** diese Verteilung und erhielt mehr formales und allgemeines Ergebnis der de Moivre-Näherung. 1774 erhielt er die Normalverteilung als Näherung der **hypergeometrischen Verteilung.**

Gauss, CF untersuchte diese Verteilung anhand von Messproblemen in der Astronomie. Seine Werke in den Jahren 1809 und 1816 begründeten die Technik niques basierend auf der Normalverteilung, die im 19. Jahrhundert zu Standardmethoden wurde. Beachten Sie, dass selbst wenn die erste Näherung dieser Verteilung auf de zurückzuführen ist Moivre , Galilei hatte bereits festgestellt, dass die **Beobachtungsfehler** verteilt waren auf symmetrische Weise und neigten dazu, sich um ihren wahren **Wert herum neu zu gruppieren.**

In der Literatur finden sich viele Bezeichnungen der Normalverteilung.

Quetelet, A. (1796–1874) sprach von der „Kurve der Möglichkeiten" oder der „Verteilung der Möglichkeiten".

Beachten Sie auch, dass **Galton, F.** von der **„Häufigkeit** der **Fehlerverteilung** " oder von der „Verteilung der **Abweichungen** von einem **Mittelwert"** sprach .

Stigler, S. (1980) präsentiert eine ausführlichere Diskussion über die verschiedenen Namen davon Kurve.

MATHEMATISCHE ASPEKTE

Der **Erwartungswert** der Normalverteilung ergibt sich aus:

$$E[X] = \ddot{y} .$$

Die **Varianz** ist gleich:

$$\mathrm{Var}(X) = \ddot{y}2 .$$

Wenn der **Mittelwert ÿ** gleich 0 und die **Varianz** ÿ2 gleich 1 ist, erhalten wir eine zentrierte und reduzierte Normalverteilung (oder Standardnormalverteilung), deren **Dichtefunktion** gegeben ist durch:

$$f(x) = \frac{1}{\ddot{y}2\ddot{y}} \exp\ddot{y}x2\frac{}{2} \quad .$$

Wenn eine **Zufallsvariable** *X* einer Normalen folgt Verteilung von Mittelwertÿ und Varianz ÿ2, dann die Zufallsvariable

$$Z = \frac{X\ddot{y}\ \ddot{y}}{\ddot{y}}$$

folgt einer zentrierten und reduzierten Normalverteilung Bution (von Mittelwert 0 und Varianz 1).

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Normalverteilung ist die bekannteste **kontinuierliche Wahrscheinlichkeitsverteilung.** Sie spielt eine zentrale Rolle in der Wahrscheinlichkeitstheorie **und** ihren statistischen Anwendungen. Viele Messungen wie die Größe oder das Gewicht von Personen, der Durchmesser einer Maschine, die Ergebnisse eines IQ-Tests usw.

N

Machine Translated by Google

folgen ungefähr einer Normalverteilung.
Die Normalverteilung wird häufig als Näherungswert verwendet, entweder wenn die Normalität einer Verteilung im Konstrukt zugeschrieben wird
tion eines **Modells** oder wenn eine bekannte Verteilung vorliegt
wird durch eine Normalverteilung mit demselben **Erwartungswert** oder **derselben Varianz ersetzt.** Es wird für die Approximation **der Chi-Quadrat-Verteilung,** der **Student-Verteilung** und **diskreter Wahrscheinlichkeitsverteilungen** wie der **Binomialverteilung** und der **Poisson-Verteilung verwendet.**
Die Normalverteilung ist ebenfalls ein Fundamentalelement der Stichprobentheorie , wo seine Rolle bei der Untersuchung von Korrelation, **Regressionsanalyse , Varianzanalyse** und **Kovarianzanalyse wichtig ist.**

WEITERLESEN

Kontinuierliche Wahrscheinlichkeitsverteilung
Lognormalverteilung
Normaler Tisch

VERWEISE

Bernoulli, J.: ArsConjectandi, Opus Posthu mum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Basel (1713)

Gauss, CF: Theoria Motus Corporum Coelestium. Werke, 7 (1809)

Gauss, CF: Bestimmung der Genauigkeit der Beobachtungen, vol. 4, S. 109–117 (1816). In: Gauss, CF Werke (erschienen 1880). Dieterichsche Universitätsdruckerei, Göttingen

Laplace, PS de: Erinnerung an die wahrscheinliche Ursache der Ereignisse. Mem. Acad. Roy. Wissenschaft. (vorgetragen von verschiedenen Wissenschaftlern) 6, 621–656 (1774) (oder Laplace PS de (1891) Œuvres complètes, Bd. 8. Gauthier-Villars, Paris, S. 27–65)

Moivre, A. de: Approximatio ad summam terminorum binomii *(a + b)n,* in seriem expansi. Supplementum II zu Miscellanae Analytica, S. 1–7 (1733). Fotografisch nachgedruckt in einer seltenen Broschüre über Moivre und einige seiner Entdeckungen. Kneipe gelistet von Archibald, RC Isis 8, 671–683 (1926)

Moivre, A. de: Die Lehre vom Zufall: oder eine Methode zur Berechnung der Wahrscheinlichkeit von Ereignissen im Spiel. Pearson, London (1718)

Stigler, S.: Stiglers Gesetz der Eponymie, Transactions of the New York Academy of Sciences, 2. Serie **39,** 147–157 (1980)

Normale Gleichungen

Normale Gleichungen sind Gleichungen, die man erhält, indem man die partiellen Ableitungen der Summe der Fehlerquadrate **(kleinste Quadrate) gleich Null setzt;** Normalgleichungen ermöglichen die Schätzung der Parameter einer multiplen linearen Regression.

GESCHICHTE Siehe **Analyse der Regression.**

MATHEMATISCHE ASPEKTE

Betrachten Sie ein **allgemeines Modell der multiplen linearen Regression:**

$$Yi = \ddot{y}0 + \sum_{j=1}^{p\ddot{y}1} \ddot{y}j \cdot Xji + \ddot{y}i, \qquad i = 1,...,n,$$

wobei *Yi* die **abhängige Variable ist,** *Xji, j = 1,...,pÿ1,* die unabhängigen Variablen sind, *ÿi* der Term des zufälligen nicht beobachtbaren **Fehlers ist,**

Machine Translated by Google

ÿj, j = 0,...,p ÿ 1, sind die zu schätzenden Parameter und *n* ist die Anzahl der Beobachtungen.

Um die Methode der **kleinsten Quadrate anzuwenden,** sollten wir die Summe der Fehlerquadrate minimieren:

$$f(\ddot{y}0, \ddot{y}1, \ldots, \ddot{y}p\ddot{y}1) = \sum_{i=1}^{N} \ddot{y}2_{\ddot{y}Yi\ \ddot{y}\ \ddot{y}0\ \ddot{y}} = \sum_{i=1}^{N} \left(\ddot{y}_{\ddot{y}} - \sum_{j=1}^{p\ddot{y}1} \ddot{y}j \cdot Xji \right)^2 \ddot{y}.$$

Wenn wir die *p* partiellen Ableitungen relativ zu den zu schätzenden Parametern gleich Null setzen, erhalten wir die *p* Normalgleichungen:

$$\begin{aligned} \ddot{y} \frac{\ddot{y}f(\ddot{y}0,\ddot{y}1,\ldots,\ddot{y}p\ddot{y}1)}{\ddot{y}\ddot{y}0} &= 0 \ , \\ \frac{\ddot{y}f(\ddot{y}0,\ddot{y}1,\ldots,\ddot{y}p\ddot{y}1)}{\ddot{y}\ddot{y}1} &= 0 \ , \\ &\vdots \\ \frac{\ddot{y}f(\ddot{y}0,\ddot{y}1,\ldots,\ddot{y}p\ddot{y}1)}{\ddot{y}\ddot{y}p\ddot{y}1} &= 0 \ . \end{aligned}$$

Wir können dasselbe Modell auch in Matrixform ausdrücken:

$$\mathbf{Y} = \mathbf{X}\ddot{\mathbf{y}} + \ddot{\mathbf{y}} \quad ,$$

Dabei ist **Y** der **Vektor** (n×1) der Beobachtungen in Bezug auf die abhängige Variable *(n* Beobachtungen), **X** die **Matrix** *(n* × p) des Plans in Bezug auf die unabhängigen Variablen und **ÿ** der *(n* × 1)-Vektor der Fehler und **ÿ** ist der *(p* × 1)-Vektor der zu schätzenden Parameter.

Mit der Methode der **kleinsten Quadrate** sollten wir Folgendes minimieren:

$$\begin{aligned} \ddot{\mathbf{y}}\ \ddot{\mathbf{y}} &= (\mathbf{Y}\ \ddot{y}\ \mathbf{X}\ddot{\mathbf{y}}\ )(\mathrm{Y}\ \ddot{y}\ \mathbf{X}\ddot{\mathbf{y}}) \\ &= \mathbf{YY}\ \ddot{y}\ \ddot{\mathbf{y}}\ \mathbf{XY}\ \ddot{y}\ \mathbf{Y}\ \mathbf{X}\ddot{\mathbf{y}} + \ddot{\mathbf{y}}\ \mathbf{X} \\ &\quad \mathbf{X}\ddot{\mathbf{y}} = \mathbf{YY}\ \ddot{y} \\ &\quad \mathbf{2}\ddot{\mathbf{y}}\ \mathbf{XY} + \ddot{\mathbf{y}}\ \mathbf{X}\ \mathbf{X}\ddot{\mathbf{y}}\ . \end{aligned}$$

Die Ableitungen relativ zu **ÿ** (in Matrixform) auf Null setzen, entsprechend dem Partial Ableitungen, aber in Vektorform geschrieben, finden wir die Normalgleichungen:

$$\mathbf{X}\ \mathbf{X}\ddot{\mathbf{y}}\hat{} = \mathbf{XY} \quad .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Wir können das Konzept der Normalgleichungen im Fall einer **nichtlinearen Regression** (mit *r* Beobachtungen und *p* + 1 zu schätzenden Parametern) verallgemeinern.

Sie werden ausgedrückt durch:

$$\sum_{i=1}^{R} \frac{Yi\ \ddot{y}\ f\ X1i,\ldots,Xki;\ \ddot{y}0,\ldots,\ \ddot{y}p\ \ddot{y}\ddot{y}j}{} = \sum_{i=1}^{R} Yi\ \ddot{y}\ f\ X1i,\ldots,Xki;\ \ddot{y}0,\ldots,\ \ddot{y}p \cdot \frac{\ddot{y}f\ X1i,\ldots,Xki;\ \ddot{y}0,\ldots,\ \ddot{y}p\ \ddot{y}\ddot{y}j}{} = 0 \ ,$$

wobei *j = 0,...,p.* Da *f* in den zu schätzenden Parametern nichtlinear ist, sind auch die Normalgleichungen nichtlinear und können daher sehr schwer zu lösen sein. Darüber hinaus kann das Gleichungssystem mehr als eine Lösung haben, die der Möglichkeit vieler Minima von *f* entspricht . Normalerweise versuchen wir, ein iteratives Verfahren zur Lösung der Gleichungen zu entwickeln oder beziehen uns auf die in der nichtlinearen Regression erläuterten Techniken.

BEISPIELE

Wenn das **Modell** beispielsweise 3 Parameter ÿ0, ÿ1 und ÿ2 enthält, lauten die Normalgleichungen wie folgt:

$$\begin{aligned} \ddot{y} \frac{\ddot{y}f\ (\ddot{y}0,\ \ddot{y}1,\ \ddot{y}2)}{\ddot{y}\ddot{y}0} &= 0 \ , \\ \frac{\ddot{y}f\ (\ddot{y}0,\ \ddot{y}1,\ \ddot{y}2)\ \ddot{y}\ddot{y}1}{\ddot{y}f} &= 0 \ , \\ \frac{(\ddot{y}0,\ \ddot{y}1,\ \ddot{y}2)\ \ddot{y}\ddot{y}1}{} &= 0 \ . \end{aligned}$$

N

Machine Translated by Google

Das ist:

$$\ddot{y}\,\ddot{y}0 \cdot n + \ddot{y}1 \cdot \sum_{i=1}^{N} X1i + \ddot{y}2 \cdot \sum_{i=1}^{N} X2i = \sum_{i=1}^{N} Yi\,,$$

$$\ddot{y}0 \cdot \sum_{i=1}^{N} X1i + \ddot{y}1 \cdot \sum_{i=1}^{N} X2_{1i} + \ddot{y}2 \cdot \sum_{i=1}^{N} X1i \cdot X2i = \sum_{i=1}^{N} X1i \cdot Yi\,,$$

$$\ddot{y}0 \cdot \sum_{i=1}^{N} X2i + \ddot{y}1 \cdot \sum_{i=1}^{N} X2i \cdot X1i + \ddot{y}2 \cdot \sum_{i=1}^{N} X2_{2i} = \sum_{i=1}^{N} X2i \cdot Yi\,.$$

Siehe **einfache lineare Regression.**

WEITERLESEN

- **Fehler**
- **Kleinsten Quadrate**
- **Multiple lineare Regression**
- **Nichtlineare Regression**
- **Parameter**
- **Einfache lineare Regression**

REFERENZEN

Legendre, AM: Neue Methoden zur Bestimmung der Umlaufbahnen von Kometen.

Courcier, Paris (1805)

Normales Wahrscheinlichkeitsdiagramm

Mithilfe eines Normalwahrscheinlichkeitsdiagramms kann überprüft werden, ob ein **Datensatz** gemäß einer **Normalverteilung verteilt ist.** Wenn dies der Fall ist

In diesem Fall ist es möglich, den **Mittelwert** und die **Standardabweichung** aus dem Diagramm abzuschätzen . Die kumulierten **Häufigkeiten** werden auf einer Gauß-**Skala** auf der **Ordinate** dargestellt , abgestuft durch Setzen des **Werts** der **Verteilungsfunktion** *F(t)* der **Normalverteilung** (zentriert und reduziert) gegenüber dem Punkt, der sich im **Abstand** *t* vom Ursprung befindet. Auf **der Abszisse** sind die **Beobachtungen** dargestellt auf einer arithmetischen Skala. In der Praxis werden die am Markt üblichen Wahrscheinlichkeitspapiere verwendet.

MATHEMATISCHE ASPEKTE Auf einem Achsensystem erfolgt die Wahl des Ursprungs und der Größe der Einheitslänge entsprechend den **Beobachtungen ,** die dargestellt werden sollen.
Die **Abszisse** ist dann wie Millimeterpapier arithmetisch abgestuft.
Auf der **Ordinate** sind die Werte der **Verteilungsfunktion** *F(t)* der **Normalverteilung** auf der Höhe *t* eines fiktiven Ursprungs in der Mitte des Blattes niedergeschrieben .

Deshalb:

F()	T
0,9772	2
0,8412	1
0,5000	0
0,1588 ÿ1	
0,0	

wird platziert und nur die **Skala** auf der linken Seite, die von 0 bis 1 (oder von 0 bis 100, wenn in %) abgestuft ist, wird beibehalten.
Betrachten Sie einen Satz von *n* Punktdaten , die in aufsteigender Reihenfolge klassifiziert werden sollen. Nennen wir diese Reihe x1, *x2,...,xn,* und für jede **Abszisse** *xi (i = 1,...,n)* wird die **Ordinate** berechnet als

$$100 \cdot \frac{3\,i\,\ddot{y}\,8}{1\,n + 4}$$

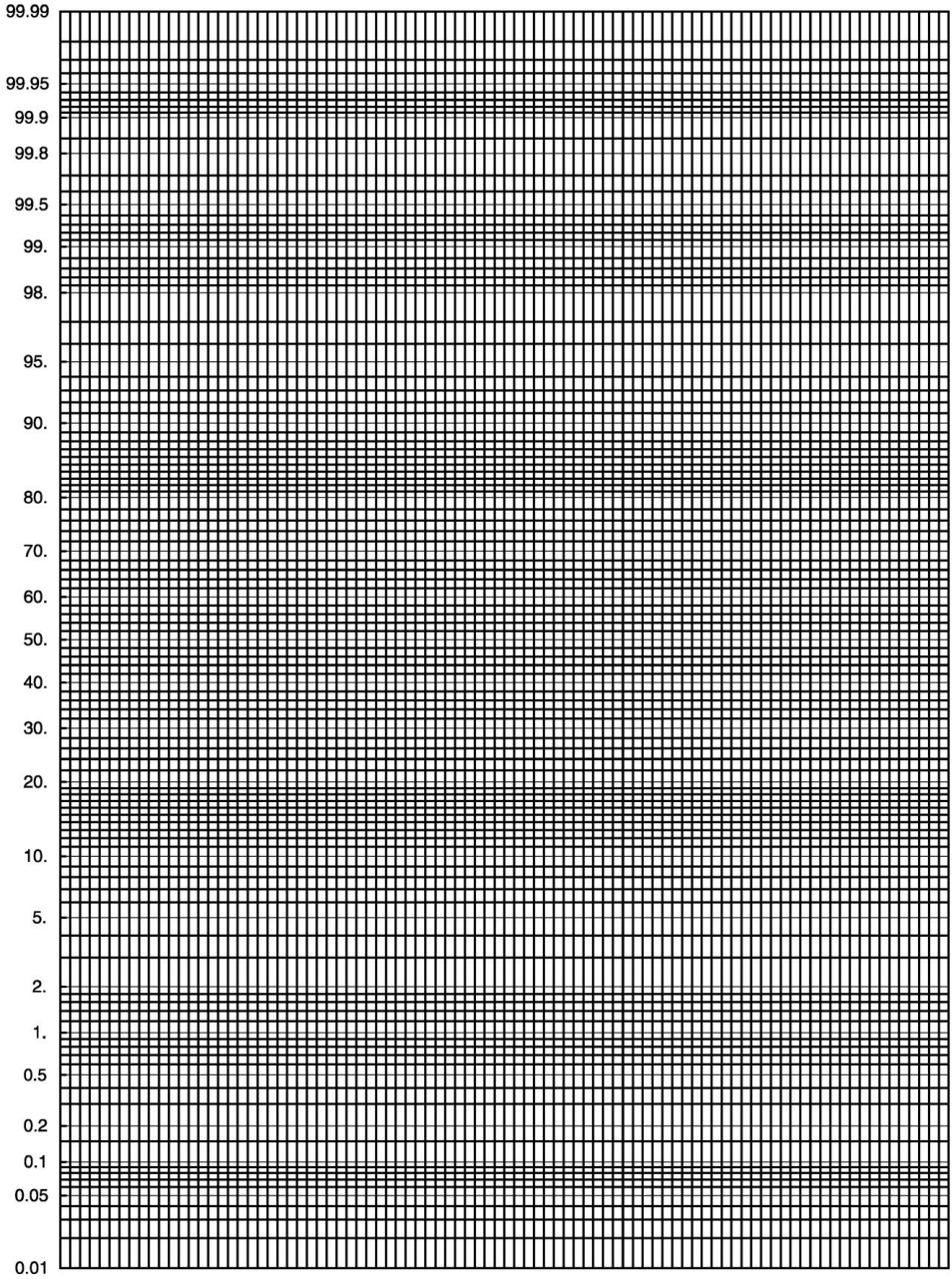
und in einem Normalwahrscheinlichkeitsdiagramm dargestellt. Es ist auch eine Transkribierung möglich

$$2\,100 \cdot \frac{1\,i\,\ddot{y}}{N}$$

als Funktion von *xi* im Diagramm. Diese zweite Näherung der kumulierten **Häufigkeit** wird wie folgt erklärt. Wenn die Fläche unter der Normalenkurve liegt (mit einer Fläche gleich

Machine Translated by Google

Papier gausso-arithmetique. Echelle verticale : 0.01 - 99.99 % (made with Mathematica™)



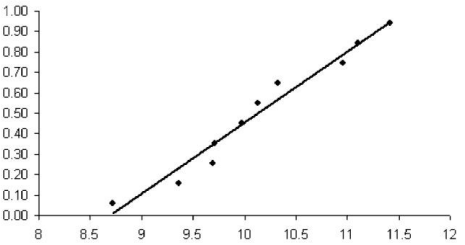
Normalwahrscheinlichkeitspapier. Vertikale Skala: 0,01 - 99,99 %

Machine Translated by Google

1) in *n* gleiche Teile geteilt ist, kann man annehmen, dass jede unserer *n* **Beobachtungen** sein kann in einem dieser Teile gefunden. Daher das ith Beobachtung xi (in aufsteigender Reihenfolge) liegt in der Mitte des i-ten Teils, der in Begriffen der kumulierten **Häufigkeit** entspricht

$$\frac{i - \frac{1}{2}}{N}.$$

Die Normalität der beobachteten Verteilung kann man dann beobachten, indem man die Ausrichtung der Punkte untersucht: Wenn die Punkte ausgerichtet sind, dann ist die Verteilung normal.



DOMÄNEN UND EINSCHRÄNKUNGEN

Durch Darstellung der kumulierten **Häufigkeiten** als Funktion der **Werte** der Statistik **Variable** in einem normalen Wahrscheinlichkeitsnetz möglich, zu überprüfen, ob die beobachtete Verteilung folgt einer **Normalverteilung.** In der Tat, wenn Dies ist der Fall, die Punkte sind ungefähr ausgerichtet.

Daher die grafische Anpassung einer Linie zu einer Menge von Punkten ermöglicht:

- Visualisieren Sie den normalen Charakter des Verteilung besser als mit **einem Togramm.**
- Erstellen Sie eine grafische **Schätzung Mittelwert** und die **Standardabweichung** von die beobachtete Verteilung.

Die Linie der **kleinsten Quadrate** schneidet die **Ordinate** 0,50 (oder 50 %) bei der **Schätzung** des Mittelwerts (auf der Abszisse ablesbar). Der Die geschätzte Standardabweichung ergibt sich aus der Kehrwert der Steigung der Linie.

Die Steigung lässt sich leicht ermitteln, indem man einfach zwei Punkte berücksichtigt. Lasst uns wählen (ÿˆ,0,5) und der Punkt auf der **Ordinate**

$$0{,}9772 = F(2) = F\left(\frac{X0{,}9772 - \hat{y}}{\hat{y}}\right)$$

deren **Abszisse** X0,9772 auf der Zeile abgelesen wird. Wir haben:

$$\frac{1}{\hat{y}} = \frac{F^{-1}(0{,}9772) - F^{-1}(0{,}5)}{X - \bar{y}}$$

= Steigung der Linie der **kleinsten Quadrate ,**

woraus wir schließlich erhalten:

$$\hat{y} = \frac{X0{,}9772 - \hat{y}}{2}.$$

BEISPIELE

Betrachten Sie zehn Teile, die von einer Maschine hergestellt wurden *X* zufällig ausgewählt: Sie haben Folgendes Durchmesser:

9.36 , 9.69 , *11.10* , 8,72 , *10.13* ,
9,98 , *11.42* , *10.33* , 9.71 , *10,96* .

Diese vermeintlich normalverteilten Stücke **sind** dann in aufsteigender Reihenfolge sortiert und für jeden Durchmesser

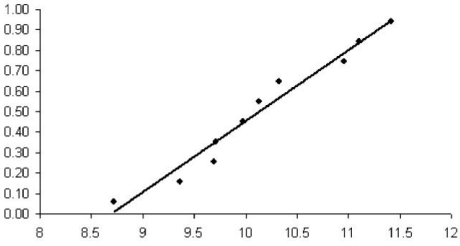
$$fi = \frac{i - \frac{3}{8}}{10 + \frac{1}{4}},$$

werden berechnet, *wobei i* die Bestellnummer von ist jede **Beobachtung:**

xi	fi
8,72	5/82
9,36	13/82
9,69	21/82
9,71	29/82
9,98	37/82

Machine Translated by Google

xi	fi
10,13	45/82
10,33	53/82
10,96	61/82
11.10	69/82
11,42	77/82



Die **grafische Darstellung** von *fi* in Bezug auf *xi* auf einem Blatt Normalwahrscheinlichkeitspapier ergibt für unsere Zufallsstichprobe : geschätzter **Mittelwert** der Durchmesser: ÿˆ = *10,2,* geschätzte **Standardabweichung:** ÿˆ = 0,9.

WEITERLESEN

- **Verteilungsfunktion**
- **Frequenz**
- **Häufigkeitsverteilung**
- **Normalverteilung**

Normaler Tisch

Die Normaltabelle, auch Gausstabelle oder Normalverteilungstabelle genannt , gibt die Werte der **Verteilungsfunktion** eines **Laufs** an **Domvariable,** die einer Standardnormalverteilung folgt, d. h. deren **Mittelwert** gleich 0 und deren **Varianz** gleich 1 ist.

GESCHICHTE

de Laplace, Pierre Simon (1774) erhielt die **Normalverteilung** aus der **hypergeometrischen Verteilung** und in einem zweiten Werk (datiert 1778, aber veröffentlicht 1781) er gab einen normalen Tisch. **Pearson, Egon Sharpe** und Hartley, HO (1948) erstellten eine Normaltabelle basierend auf den von Shepard, WF (1903, 1907) berechneten Werten. Ausführliche Listen der auf dem Markt erhältlichen Standardtische wurden vom National Bureau of Standards (bis 1952) sowie von Greenwood, JA und Hartley, HO (bis 1958) herausgegeben.

DOMÄNEN UND EINSCHRÄNKUNGEN

Siehe **Zentraler Grenzwertsatz.**

MATHEMATISCHE ASPEKTE

Die **Zufallsvariable** *Z* folge einer **Normalen Verteilung** des **Mittelwerts** 0 und der **Varianz** 1.

Die **Dichtefunktion** *f (t)* ist somit gleich:

$$f(t) = \cdot \exp \frac{1}{\ddot{y}t} \quad \frac{2}{2} \quad .$$

Die **Verteilungsfunktion** der Zufallsvariable fähig *Z* ist definiert durch:

$$F(z) = P(Z \ddot{y} z) = \int_{\ddot{y}\ddot{y}}^{z} f(t)\, dt .$$

Die hier dargestellte normale Tabelle gibt die Werte von *F (z)* für die verschiedenen Werte von *z* zwischen 0 und 3,5 an.

Mit einer zum Mittelwert symmetrischen **Normalverteilung** gilt:

$$F(z) = 1 \ddot{y} F(\ddot{y}z) ,$$

womit sich *F (z)* für einen negativen Wert von z bestimmen lässt .

Manchmal müssen wir die Normaltabelle in umgekehrter Richtung verwenden, um den Wert von *z* zu finden , der einer gegebenen Wahrscheinlichkeit entspricht. Im Allgemeinen bezeichnen wir mit *z* = zÿ den Wert der Zufallsvariablen *Z* , für die

$$P(Z \ddot{y} z\ddot{y}) = 1 \ddot{y} \ddot{y} .$$

N

Machine Translated by Google

BEISPIELE

Siehe Anhang G.

Beispiele für den normalen Tischgebrauch:

$$P(Z \le 2{,}5) = 0{,}9938 \quad ,$$

$$P(1 \le Z \le 2) = P(Z \le 2) - P(Z \le 1) = 0{,}9772 - 0{,}8413 = 0{,}1359 \quad ,$$

$$P(Z \le -1) = 1 - P(Z \le 1) = 1 - 0{,}8413 = 0{,}1587 \quad ,$$

$$P(-0{,}5 \le Z \le 1{,}5) = P(Z \le 1{,}5) - P(Z \le -0{,}5) = P(Z \le 1{,}5) - [1 - P(Z \le 0{,}5)] = 0{,}9332 - [1 - 0{,}6915] = 0{,}9332 - 0{,}3085 = 0{,}6247 .$$

Umgekehrt können wir die Normaltabelle verwenden, um den Grenzwert z_α zu bestimmen, **für** den die Wahrscheinlichkeit , dass die **Zufallsvariable Z** einen kleineren Wert als diesen Grenzwert (z_α) annimmt, einem festen Wert ($1 - \alpha$) entspricht, d. h.:

$$P(Z \le z_\alpha) = 0{,}95 \Rightarrow z_\alpha = 1{,}645 \quad P(Z \le z_\alpha) = 0{,}975 \quad , \Rightarrow z_\alpha = 1{,}960 .$$

Dies ermöglicht es uns, beim **Hypothesentest** den **kritischen Wert** z_α relativ zu einem gegebenen **Signifikanzniveau** α zu bestimmen .

Anwendungsbeispiel:

Wenn bei einem **einseitigen Test** das Signifikanzniveau α gleich 5 % ist, dann können wir im Folgenden den kritischen Wert *z0,05* ermitteln Benehmen:

$$P(Z \le z_{0{,}05}) = 1 - \alpha = 1 - 0{,}05 = 0{,}95$$

$$\Rightarrow z_{0,05} = 1{,}645 .$$

Im Falle eines **zweiseitigen Tests** sollten wir den Wert $z_{\alpha/2}$ ermitteln . Bei einem Signifikanzniveau von 5 % ergibt sich der kritische Wert z0,025 auf folgende Weise:

$$P(Z \le z_{0{,}025}) = 1 - \frac{\alpha}{2} = 1 - 0{,}025 = 0{,}975$$

$$\Rightarrow z_{0{,}025} = 1{,}960 .$$

WEITERLESEN

- **Hypothesentest**
- **Normalverteilung**
- **Einseitiger Test**
- **Statistische Tabelle**
- **Zweiseitiger Test**

VERWEISE

Greenwood, JA, Hartley, HO: Leitfaden zu Tabellen in der mathematischen Statistik. Prinz Tonne University Press, Princeton, NJ (1962)

Laplace, PS de: Erinnerung an die wahrscheinliche Ursache der Ereignisse. Mem. Acad. Roy. Wissenschaft. (vorgetragen von verschiedenen Wissenschaftlern) 6, 621–656 (1774) (oder Laplace, PS de (1891) Œuvres complètes, Bd. 8. Gauthier-Villars, Paris, S. 27–65)

Laplace, PS de: Mémoire sur les probabil ités.Mem. Acad. Roy. Wissenschaft. Paris, 227–332 (1781) (oder Laplace, PS de (1891) Œuvres complètes, Bd. 9. Gauthier-Villars, Paris, S. 385–485.)

National Bureau of Standards.: Ein Leitfaden zu Tabellen des Normalwahrscheinlichkeitsintegrals. US-Handelsministerium. Reihe Angewandte Mathematik, 21 (1952)

Pearson, ES, Hartley, HO: Biometrika Tables for Statisticians, 1 (2. Aufl.). Cambridge University Press, Cambridge (1948)

Machine Translated by Google

Sheppard, WF: Neue Tabellen des Wahrscheinlichkeitsintegrals. Biometrie 2, 174–190 (1903)

Sheppard,WF: Tabelle der Abweichungen von der Normalkurve. Biometrie 5, 404–406 (1907)

Normalisierung

Normalisierung ist die Transformation von eine normalverteilte **Zufallsvariable** zu eine Zufallsvariable, die einer Standardnormalverteilung folgt. Es ermöglicht die Berechnung und Vergleichen Sie die Werte, die zu den Normalkurven des **Mittelwerts** und der Differenz gehören **Abweichungen** auf Basis einer Referenz noch Malverteilung, also die Standardnormalverteilung Verteilung.

MATHEMATISCHE ASPEKTE

Die **Zufallsvariable** *X* folge der **Normalverteilung** *N* *ÿ,* ÿ2 . Es normalisiert sich tion gibt uns eine Zufallsvariable

$$Z = \frac{X \ddot{y} \ddot{y}}{\ddot{y}}$$

das folgt einer Standardnormalverteilung *N* (0, 1).
Jeder **Wert** der VerteilungN *ÿ,* ÿ2 kann in die **Standardvariable** Z transformiert werden , Jedes *Z* stellt eine Abweichung vom dar **Mittelwert,** ausgedrückt in Einheiten der **Standardabweichung.**

BEISPIELE

Die Schüler einer Berufsschule haben hatte zwei Prüfungen. Jede Prüfung wurde benotet eine Skala von 1 bis 60, und die Noten werden als Realisierungen zweier Zufallsvariablen betrachtet einer **Normalverteilung.** Vergleichen wir die Noten, die die Studierenden hierfür erhalten haben zwei Prüfungen.

Hier sind die Mittelwerte und die Standardabweichung tionen jeder Prüfung, berechnet für alle Stu Dellen:

Prüfung 1: ÿ1 = 35 , ÿ1 = 4 Prüfung ,

2: ÿ2 = 45 , ÿ2 = *1,5* ,

Ein Student namens Marc hat Folgendes erhalten Ergebnisse:

Prüfung 1: X1 = 41 ,

Prüfung 2: X2 = 48 .

Die Frage ist zu wissen, welche Prüfung Marc hat im Vergleich zu den anderen Stu besser abgeschnitten Dellen.Wir können die Ergebnisse nicht direkt vergleichen von zwei Prüfungen, weil die Ergebnisse dazu gehören Verteilungen mit unterschiedlichen Mitteln und Standards merkliche Abweichungen.

Untersuchen wir einfach den Unterschied jedes einzelnen Notieren Sie dies und vergleichen Sie es mit dem Mittelwert Verteilung erhalten wir:

Prüfung 1: X1 ÿ ÿ1 = 41 ÿ 35 = 6 ,

Prüfung 2: X2 ÿ ÿ2 = 48 ÿ 45 = 3 .

Beachten Sie, dass Marcs Punktzahl um 6 Punkte höher war als der Durchschnitt der Prüfung 1 und nur 3 Punkte höher als der Mittelwert in Prüfung 2. Eine voreilige Schlussfolgerung würde uns nahelegen, dass Marc es getan hat besser bei Prüfung 1 als bei Prüfung 2, relativ zu die anderen Studenten.

Diese Schlussfolgerung berücksichtigt nur die Differenz jedes Ergebnisses vom Mittelwert. Es berücksichtigt nicht die Streuung der Schüler Noten um den Mittelwert. Um die Ergebnisse vergleichbar zu machen, dividieren wir die Differenz vom Mittelwert durch die Standardabweichung:

$$\text{Prüfung 1: } Z1 = \frac{X1 \ddot{y} \ddot{y}1}{\ddot{y}1} = \frac{6}{} = 1{,}5\,4 ,$$

$$\text{Prüfung 1: } Z2 = \frac{X2 \ddot{y} \ddot{y}2}{\ddot{y}2} = \frac{3}{1.5} = 2 .$$

Durch diese Berechnung haben wir die normalisiert Punkte X1 und X2. Wir können daraus schließen, dass die

N

Machine Translated by Google

Der normalisierte **Wert** ist für Prüfung 2 *(Z2* = 2) größer als für Prüfung 1 *(Z1* = 1,5) und dass Marc bei Prüfung 1 im Vergleich zu anderen Studenten besser abgeschnitten hat.

WEITERLESEN

Normalverteilung

Nullhypothese

Bei der Erfüllung des **Hypothesentests** ist die Nullhypothese die Hypothese, die getestet werden soll. Es ist mit H0 bezeichnet. Die entgegengesetzte Hypothese **wird als Alternativhypothese bezeichnet.** Es ist die Alternativhypothese, die akzeptiert wird, wenn der Test zur Ablehnung der Nullhypothese führt.

GESCHICHTE

Siehe **Hypothese** und **Hypothesentests.**

MATHEMATISCHE ASPEKTE

Bei der Durchführung von **Hypothesentests** für einen **Parameter** einer **Grundgesamtheit** ist die Nullhypothese normalerweise eine Annahme über den angenommenen **Wert** dieses Parameters. Die Nullhypothese wird dann wie folgt dargestellt:

$$H0 : ÿ = ÿ0 ,$$

ÿ ist der unbekannte Parameter der Bevölkerung und ÿ0 der angenommene Wert dieses Parameters. Dieser **Parameter** kann beispielsweise der **Mittelwert** der Grundgesamtheit sein.

Wenn Hypothesentests darauf abzielen, zwei **Populationen zu vergleichen,** geht die Nullhypothese im Allgemeinen davon aus, dass die **Parameter** gleich sind:

$$H0 : ÿ1 = ÿ2 ,$$

Dabei ist ÿ1 der Parameter der **Population ,** aus der die erste Stichprobe stammt, und ÿ2 die Population, aus der die zweite Stichprobe stammt.

BEISPIELE

Wir werden die Nullhypothese anhand von drei Beispielen für **Hypothesentests untersuchen:** 1. *Hypothesentest anhand des Prozentsatzes einer Stichprobe*

Ein Kandidat, der sich um ein Amt bewirbt, möchte wissen, ob er mehr als 50 % davon erhalten wird Die Wahl.

Die **Nullhypothese für** dieses Problem kann folgendermaßen aufgestellt werden:

$$H0 : ÿ = 0,5 ,$$

Dabei ist ÿ der zu schätzende **Prozentsatz** der **Bevölkerung .**

2. *Hypothesentest am Mittelwert der Grundgesamtheit*

Ein Hersteller möchte die Präzision einer neuen Maschine testen, die Teile mit einem Durchmesser von 8 mm produzieren soll.

Wir können die folgende **Nullhypothese** aufstellen :

$$H0 : ÿ = 8 ,$$

wobei ÿ der **Mittelwert** der zu schätzenden **Bevölkerung** ist.

3. *Hypothesentest zum Vergleich der Mittelwerte zweier Grundgesamtheiten*

Eine Versicherungsgesellschaft hat beschlossen, ihre Büros mit Computern auszustatten. Sie will Computer von zwei verschiedenen Anbietern kaufen, wenn es keinen signifikanten Unterschied zwischen der Zuverlässigkeit der beiden Marken gibt.

Es werden zwei Stichproben von PCs jeder Marke gezogen. Dann wird aufgezeichnet, wie Es dauerte viel Zeit, bis für jede Probe ein Problem auf dem PC auftrat.

Gemäß der Nullhypothese ist der Mittelwert der seit dem ersten Problem verstrichenen Zeit für jede Marke gleich:

$$H0 : ÿ1 ÿ ÿ2 = 0 ,$$

Machine Translated by Google

wobei ÿ1 und ÿ2 die jeweiligen Mittelwerte zweier Populationen sind.

Diese **Hypothese** kann auch wie folgt geschrieben werden:

$$H0 : ÿ1 = ÿ2 .$$

WEITERLESEN

- **Alternative Hypothese**
- **Varianzanalyse**
- **Hypothese**
- **Hypothesentest**

Machine Translated by Google

Ö

Überwachung

Eine Beobachtung ist das Ergebnis einer wissenschaftlichen Arbeit Studie zur Zusammenstellung von Informationen über eine statistische Einheit, die zu einer bestimmten **Bevölkerung gehört.**

WEITERLESEN

- **Daten**
- **Ausreißer**
- **Bevölkerung**
- **Probe**

Quoten und Quotenverhältnis

Quoten sind definiert als das Verhältnis der Wahrscheinlichkeit eines Ereignisses zur Wahrscheinlichkeit des Ereignisses Komplementärereignis, also das Verhältnis der Wahrscheinlichkeit des Eintretens eines Ereignisses (insbesondere einer Epidemiologie, einer Krankheit) und die Wahrscheinlichkeit, dass es nicht eintritt.
Das Odds Ratio ist, wie das **relative Risiko,** ein Maß für die relative Wirkung einer Ursache. Mehr

Genauer gesagt ist es das Verhältnis der Wahrscheinlichkeiten, die innerhalb der Bevölkerung berechnet werden, die einem bestimmten Risiko ausgesetzt ist Faktor in unterschiedlichem Ausmaß.

MATHEMATISCHE ASPEKTE

Chancen, dass ein Ereignis „eine Krankheit hat“ sind formal das folgende Verhältnis, wobei *p* bezeichnet die Wahrscheinlichkeit, an der Krankheit zu erkranken:

$$\text{Quoten} = \frac{P}{1 \ddot{y} S}.$$

Lassen Sie es einen Risikofaktor für eine Gruppe geben E ist exponiert und eine Gruppe NE ist nicht exponiert. Das entsprechende Quotenverhältnis, OR, ist Verhältnis der Quoten der Gruppe E und der Quoten von Gruppe NE:

$$\text{ODER} \quad \frac{\text{Quoten der Gruppe E}}{= \text{Quoten der Gruppe NE}}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Nehmen wir als Beispiel die Berechnung von das Quotenverhältnis aus den Daten im Folgenden Tabelle zum Vergleich zweier Personengruppen angegriffen von der gleichen potenziell gefährlichen Erkrankung; eine Gruppe („behandelte“ Gruppe) erhält medizinische Behandlung, während der andere dies nicht tut ("Referenzgruppe):

Gruppe	Todesfälle	Hinterbliebene	Quoten	ODER
Behandelt	5 %	95 %	0,053	0,47
Referenz	10 %	90 %	0,111	1,0*

*) Per Definition ist das ODER für die Gruppe von gleich 1 Referenz.

Wenn wir hier **„Risiko“** nennen, ist das damit verbundene Risiko Fall „Tod“, dann ist das OR eine gute Annäherung an das relative Risiko, unter der Bedingung, dass die Sterberisiken ähnlich sind oder,

wenn sie stark unterschiedlich sind, sind sie klein. Wir gehen grundsätzlich von einem Risiko kleiner als 10 % aus, wie in diesem Fall, als klein. Das Quotenverhältnis bleibt keine Annäherung an das Relative

Machine Translated by Google

Risiko, weil es eine **Verzerrung** des OR gibt erscheint für kleine Risiken. Das ist:

$$\text{ODER} = \frac{\text{Risiko der behandelten Gruppe, Überlebensrate in der behandelten Gruppe, Risiko in der Referenzgruppe, Überlebensrate}}{\text{in der Referenzgruppe, Risiko in der behandelten Gruppe, Risiko in}} \times \frac{\text{der Referenzgruppe, Überlebensrate in der Referenzgruppe, Überlebensrate in der behandelten Gruppe}}{}$$

$$= \text{relatives Risiko} \times \text{Bias des Odds Ratio}$$

mit:

$$\text{Bias von OR} = \frac{\text{Überlebensrate in der Referenzgruppe,}}{\text{Überlebensrate in der behandelten Gruppe}}.$$

Die Schätzungen der relativen und Chancen Verhältnisse werden verwendet, um die Kraft der Kausalität zu messen Zusammenhänge zwischen der Exposition gegenüber einem Risikofaktor und einer Krankheit. Die beiden Verhältnisse können dauern Werte zwischen 0 und unendlich, die durch dieselbe allgemeine Interpretation gerechtfertigt sind. Wir können zwei klar unterschiedliche Fälle unterscheiden: 1. Ein OR-Wert größer als 1 weist auf ein Risiko oder eine viel größere Erkrankungswahrscheinlichkeit hin, wenn die untersuchte Gruppe einem bestimmten Risikofaktor ausgesetzt ist. In einem solchen Fall liegt eine positive Assoziation vor.

2. Ein kleinerer Wert impliziert eine geringere Exposition gegenüber einem Risikofaktor. In diesem Fall gibt es eine negative Assoziation.

Konfidenzintervall bei 95 % des Odds Ratio

Die für ein Konfidenzintervall von 95 % des Quotenverhältnisses verwendete Technik und Notation sind die gleichen wie für das **relative Risiko:**

$$\exp(\ln \text{OR}) \pm 1{,}961 \sqrt{\frac{1}{A} + \frac{1}{N} + \frac{1}{C} + \frac{1}{M}}.$$

a, b, c und *d* müssen groß sein. Wenn das Konfidenzintervall den Wert 1,0 umfasst, ist das Odds Ratio statistisch nicht signifikant.

BEISPIELE

Betrachten Sie eine Studie über den Zusammenhang zwischen dem Risikofaktor „Rauchen“ und Brustkrebs. Die Wahrscheinlichkeit, an Brustkrebs zu erkranken, ist das Verhältnis von die Wahrscheinlichkeit, an dieser Krebserkrankung zu erkranken, und die Wahrscheinlichkeit, diese Krankheit während eines bestimmten Zeitraums nicht zu haben. Die Wahrscheinlichkeit, für jede Gruppe an Brustkrebs zu erkranken, wird durch den Grad der Rauchexposition der Gruppe bestimmt. Die folgende Tabelle veranschaulicht dieses Beispiel:

Gruppenkrebs	(100.000/2 Jahre)	NEIN Krebs (100000/ 2 Jahre)	Chancen auf Brust Krebs (/2 Jahre)	ODER
Nicht ausgesetzt	114	99886	0,00114	1,0
Passiv Raucher	252	99748	0,00252	2,2
Aktiv Raucher	276	99724	0,00276	2,4

In diesem Beispiel beträgt die Wahrscheinlichkeit, an Brustkrebs zu erkranken, bei der nicht exponierten Gruppe 0,00114 (114 : 99886), bei den Passivrauchern 0,00252 (252 : 99748) und bei den Aktivrauchern 0,00276 (276 : 99274). Nimmt man als Referenzgruppe die nicht exponierten Probanden, so beträgt das Wahrscheinlichkeitsverhältnis für Brustkrebs für die aktiven Raucherinnen:

$$\frac{\frac{276}{99724}}{\frac{114}{99886}} = 2{,}4 = \frac{276}{114} \cdot \frac{99886}{99724}.$$

Da das Brustkrebsrisiko über einen Zeitraum von zwei Jahren sehr gering ist, liegt die Überlebensrate sowohl aktiver Raucherinnen als auch nicht exponierter Personen bei nahezu 100 %, so dass die Überlebensquote (der Bias des Odds Ratio) praktisch bei 1,0 liegt wird als nahezu das Fehlen jeglicher OR-Verzerrung interpretiert. In diesem Beispiel ist das Wahrscheinlichkeitsverhältnis für Brustkrebs identisch mit dem relati

Machine Translated by Google

Brustkrebsrisiko bei Passivraucherinnen
und die aktiven Raucher. Seit Brustkrebs

Wenn es sich um eine seltene Erkrankung handelt, ist das Odds Ratio ein guter Näherungswert für das relative Risiko.

WEITERLESEN

Zurechenbares Risiko

Vermeidbares Risiko

Ursache und Wirkung in der Epidemiologie

Vorfall

Inzidenzrate

Quoten und Quotenverhältnis

Häufigkeit

Prävalenzrate

Relatives Risiko

Risiko

VERWEISE

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Kleiner Brauner, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Morabia, A.: L'Épidémiologie Clinique. Ausgaben „Que sais-je?“. Presses Univer Sitaires de France, Paris (1996)

Offizielle Statistik

Das Adjektiv „offiziell“ bedeutet kommend aus dem Regierung, die Verwaltung; autorisiert durch eine zuständige Behörde. So wird das Wort *„offiziell“* in Verbindung mit dem Wort „ Statistik“ weist auf eine Beziehung zum Staat hin. Darüber hinaus, die Etymologie des Wortes *Statistik,* von der lateinische *Status* (Staat), was bedeutet, Gouverneur zu sein ernmentelle Statistiken über Daten, die von staatlichen Stellen oder privaten Stellen gesammelt werden für die Regierung arbeiten. Der Begriff *offiziell Statistiken beziehen sich auf* ein sehr allgemeines Konzept kann auf unterschiedliche Weise erlangt werden.

GESCHICHTE

Die **Volkszählung,** die bekannteste amtliche Statistik, ist eine fast ebenso alte Praxis als soziale Organisation der Menschen. Seit Antike, die Könige von Babylon und Ägypten, sowie Chinesisch, Griechisch, Hebräisch, Persisch, und römische Häuptlinge versuchten, das zu bestimmen Anzahl der Soldaten unter ihrem Kommando und die Güter, die sie besaßen. Die ersten statistischen Versuche reichen mehr als 6000 Jahre zurück. Aber es waren nur sporadische Versuche, Menschen und Güter zu zählen. Statistiker, in der modernen Sinne des Begriffs, erscheinen nur in Ende des 17. Jahrhunderts mit der Gründung des modernen Staates und die Bedürfnisse von Colo Regierungsverwaltung in Frankreich und England. Im 18. Jahrhundert wird in der Enzyklopädie von Diderot und Alembert „arithmetische Politik“ erwähnt. wird definiert als „die Anwendung der Arithmetik“. Berechnungen zu politischen Themen und Gebrauch“. Die Französische Revolution und dann die Geburt von Demokratie in den Vereinigten Staaten, begründete das Neue Entwicklungen in der amtlichen Statistik. Der neue Staaten, um deren politisches Gewicht zu kennen ihre verschiedenen Entitäten, über die man Bescheid wissen musste ihre Populationen. Aus diesem Grund wurden in den USA im Jahr 2010 erstmals alle zehn Jahre Volkszählungen durchgeführt 1790.

Seit dem Zweiten Weltkrieg entstanden Imperien, Kolonien und Stammesgesellschaften wurden nach und nach durch Nationalstaaten ersetzt. Organisieren Wahlen und Versammlungen und erleichtern die Entwicklung moderner Unternehmen, diese Neue Staaten müssen statistische Systeme einrichten Das waren wesentliche Bedingungen ihrer Entwicklung Option. Die Vereinten Nationen und insbesondere

Ö

Machine Translated by Google

Insbesondere seine Statistische Kommission unternahm erhebliche Anstrengungen zur Förderung und Entwicklung amtlicher Statistiken in verschiedenen Bundesstaaten. Unter ihrer Schirmherrschaft wurde 1950 die erste Weltbevölkerungszählung durchgeführt. Eine weitere folgte 1960.

Mittlerweile sind offizielle Statistiken allgegenwärtig. Die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) möchte Statistiken erstellen, die bei der Formulierung einer präzisen, umfassenden und ausgewogenen Bewertung des Zustands der wichtigsten Bereiche der Gesellschaft helfen können. Diese als Indikatoren bezeichneten Statistiken sind seit den 1960er Jahren Gegenstand umfangreicher Forschung in den Bereichen Wirtschaft, Gesundheit und Bildung. Ziel ist es, für Politiker ein Informationssystem zu schaffen, das es ihnen ermöglicht, die Auswirkungen ihrer Wirtschafts-, Sozial- und Bildungspolitik zu bestimmen.

BEISPIELE

Amtliche Statistiken behandeln viele Bereiche, von denen die wichtigsten sind: • *Bevölkerung* (Familien, Haushalte, Geburten, Adoptionen, Anerkennungen, Todesfälle, Ehen, Scheidungen, Migrationen usw.) • *Umwelt (Land,* Klima, Wasser, Fauna, Flora). usw.) • *Beschäftigung und aktives Leben* (Unternehmen und Organisationen, Beschäftigung, Arbeitsbedingungen usw.) • *Volkswirtschaftliche Gesamtrechnungen* (Volkswirtschaftliche Gesamtrechnung, Zahlungsbilanz usw.) • *Preis* (Preise und Preisindizes) • *Produktion, Handel usw Konsum* (Leistung der Unternehmen, Produktion, Umsatz, Außenhandel etc.) • *Land- und Forstwirtschaft* (Ressourceneinsatz, Arbeitskräfte, Maschinen, Wälder, Fischerei etc.) • *Energie* (Bilanz, Produktion, Konsum etc.) • *Bauen und Wohnen* (Gebäude- und Wohnstrukturen, Lebensbedingungen usw.) • *Tourismus* (Hotels, Reisen usw.) • *Transport und Kommunikation* (Fahrzeuge, Anlagen, Verkehr, Verkehrsunfälle usw.) • *Finanzmärkte und Banken* (Bankensystem, Finanzbuchhaltung usw.) • *Versicherungen* (Krankenversicherung, Privatversicherung usw.) • *Gesundheit* (Gesundheitssystem, Mortalitäts- und Morbiditätsursachen, Gesundheitspersonal, Ausrüstung, Gesundheitskosten usw.) • *Bildung und Wissenschaft* (Grundschule und Sekundarschulbildung, Hochschulbildung, Forschung und Entwicklung usw.) • *Sport, Kultur und Lebensstandard* (Sprache, Religion, kulturelle Veranstaltungen, Sport usw.) • *Politik* (Wahlen, Teilnahme am politischen Leben usw.) • *Öffentliche* Finanzen (Bundes- und Landeseinnahmen und -ausgaben, Steuern usw.) • *Rechtssystem* (Strafgesetzbuch, Gerichte, Rückfallismus usw.)

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Geschichte zeigt die gegenseitige Abhängigkeit zwischen Statistiken und der Entwicklung der Zustand. Diese gegenseitige Abhängigkeit erklärt zumindest teilweise die Schwierigkeit, die Statistiker haben, in ihrem Fachgebiet klare Grenzen zu definieren.

Im Laufe der Geschichte und auch heute noch bestimmt die politische Macht die Grenzen des Studiums der amtlichen Statistik und erweitert oder begrenzt die Inhalte je nach Bedarf.

Politische Entscheidungen spielen in vielen Bereichen der amtlichen Statistik eine Rolle: bei Definitionen von solche Konzepte wie Arbeitslosigkeit, Armut, Wohlbefinden usw.; in den Definitionen von Variablen wie Berufskategorien, Niveau

Machine Translated by Google

des Gehalts usw. Dieses Problem der Wahl führt zu zwei weiteren Problemen:

1. Die Vergleichbarkeit von Statistiken
   Die Bedürfnisse und Ziele des Staates bestimmen die Wahl in diesem Bereich. Weil Bedürfnisse sind nicht immer gleich, die Auswahl wird an unterschiedlichen Orten und auf internationaler Ebene getroffen Statistiken werden nicht immer vergleichbar sein. Der Unterschied in den Mitteln verschiedener Staaten zur Datenerhebung und -nutzung erläutert der Unterschied in den Ergebnissen zwischen den Ländern.
2. Objektivität
   Amtliche Statistiken erfordern Objektivität. Auch wenn der politische Wille nicht vollständig sein kann beseitigt werden, ermöglichen die Weiterentwicklung statistischer Theorien und Verbesserungen der Tools viel wichtiger und zuverlässiger Datensammlung. Das behandelte Volumen Die Datenmenge ist stark gestiegen und die Objektivität der Informationen hat sich verbessert.

WEITERLESEN

- Volkszählung
- Daten
- Indexnummer
- Indikator
- Bevölkerung

VERWEISE

Duncan, JW, Gross, AC: Statistiken für das 21. Jahrhundert: Vorschläge zur Verbesserung der Statistik für eine bessere Entscheidungsfindung. Irwin ProfessionalPublishing, Chica go (1995)

Dupâquier, J. und Dupâquier, M.: Histoire de la démographie. Académie Perrin, Paris (1985)

INSEE: Für eine Geschichte der Statistik. Bd. 1: Beiträge. INSEE, Econo Glimmer, Paris (1977)

INSEE: Für eine Geschichte der Statistik. Bd. 2: Matériaux. J. Afichard (Hrsg.) INSEE, Economica, Paris (1987)

ISI: Die Zukunft der Statistik: Eine Internationale al-Perspektive. Voorburg (1994)

Stigler, S.: Die Geschichte der Statistik, die Messung der Unsicherheit vorher 1900. Belknap, London (1986)

Spaß der Statistikkommission der Vereinten Nationen Grundprinzipien der amtlichen Statistik

Ogive

Eine Ogive ist eine **grafische Darstellung** von eine kumulierte **Häufigkeitsverteilung.** Das ist eine Art Frequenzgrafik und wird auch genannt ein kumuliertes **Häufigkeitspolygon.** Es dient um die Anzahl (oder den Anteil) der Beobachtungen anzugeben, die kleiner oder gleich einer bestimmten Zahl sind **Wert.**

MATHEMATISCHE ASPEKTE

Eine Ogive ist auf einem System senkrechter Pendelachsen aufgebaut. Auf der horizontalen Achse tragen wir die Grenzen der zuvor ermittelten Klassenintervalle ein. Basierend auf jedem dieser Punkte Grenzwerte bestimmen wir deren Ordinate die Höhe entspricht der kumulierten **Häufigkeit** diesem Wert entspricht. Zeilenweise verbinden Segmentiert man die auf diese Weise ermittelten aufeinanderfolgenden Punkte, so erhält man eine Linie, die man Ogive nennt. Wir Geben Sie normalerweise die kumulierte Häufigkeit ein die linke vertikale Achse und der **Prozentsatz** von kumulierte Häufigkeit auf der rechten Vertikale Achse.

DOMÄNEN UND EINSCHRÄNKUNGEN

Ogiven eignen sich besonders gut für Schätzungen Perzentile in einer Verteilung. Zum Beispiel wir kann den zentralen Punkt kennen, so dass 50 % der

Ö

Machine Translated by Google

Beobachtungen würden unterhalb dieses Punktes liegen und 50 % darüber. Dazu ziehen wir eine Linie von der Punkt von 50 % auf der Prozentachse bis es die Kurve schneidet. Dann ver Projizieren Sie den Schnittpunkt manuell auf die horizontale Achse. Die letzte Kreuzung gibt uns das gewünschten **Wert.** Das **Frequenzpolygon** und Ogive werden zum Vergleich zweier statistischer Sätze verwendet deren Anzahl unterschiedlich sein könnte.

BEISPIELE

Aus der Tabelle, die die Verteilung darstellt Aus dem Tagesumsatz von 75 Filialen werden wir eine Ogive konstruieren.

Verteilung des Tagesumsatzes von 75 Filialen (in Dollar):

Umsatz (Klasse Intervalle)	Frei quenz	Cumu verspätet häufig cy	Prozent Tag von kumuliert Frequenz
215–235	4	4	5.3
235–255	6	10	13.3
255–275	13	23	30.7
275–295	22	45	60,0
295–315	15	60	80,0
315–335	6	66	88,0
335–355	5	71	94,7
355–375	4	75	100,0
Gesamt	75		

In diesem Beispiel erhalten wir die **Schätzung** dass 50 % der Geschäfte einen geringeren Umsatz haben als oder gleich etwa US&289.

WEITERLESEN

Häufigkeitsverteilung
Frequenzpolygon
Grafische Darstellung

Olkin, Ingram

Olkin, Ingram wurde 1924 in Waterbury, CT, USA geboren. Er erhielt sein Diplom in 1947 am City College of New York und setzte sein Studium an der Columbia University fort Stadt, wo er einen Master-Abschluss erhielt Statistik (1949). Er erhielt seinen Doktortitel in statistischer Mathematik an der Universität von North Carolina im Jahr 1951. Anschließend unterrichtete er an der Michigan State University und an der University of Minnesota, bevor er benannt wurde Professor für Statistik an der Stanford University, eine Position, die er derzeit innehat. Ingram Olkin war Herausgeber des *Annals of Statistics,* Mitherausgeber von *Psychometrika, Journal of Educational Statistics, Journal of the American Statistical Association,* und *Lineare Algebra and Its Applications,* neben anderen mathematisch-statistischen Fachzeitschriften. Seine Forschung Der Schwerpunkt liegt auf multivariater Analyse, Ungleichheiten sowie Theorie und Anwendung der Metaanalyse in den Sozialwissenschaften und der Biologie. Er ist Mitautor zahlreicher Werke, darunter *Ungleichheitstheorie der Majorisierung und ihre Applications* (1979), *Selecting and Order ing Populations (1980), Probability Models and Applications* (1980) und in jüngerer Zeit *Statistical Methods in Meta-Analysis* (1985).

Ausgewählte Werke und Artikel von Ingram Olkin:

1979 (mit Marshall, A.) Ungleichheiten: Die Theorie der Majorisierung und ihre Anwendungen. Akademisch, New York.

1984 (mit Marsaglia, G.) Generating Cor-relation Matrices. SIAM Journal of Wissenschaftliches und statistisches Rechnen 5, S. 470–475.

Machine Translated by Google

1985 (mit Hedges, LV) Statistical Methods for Meta-Analysis. Akademisch, New York.

1994 (mit Gleser, LJ und Derman, C.) Wahrscheinlichkeitsmodelle und Anwendungen. Macmillan, New York. **1995** (mit Moher, D.) Metaanalyse randomisierter kontrollierter Studien. Ein Betrug Interesse an Standards. Marmelade. Med. Assoc. 274, 1962–1963. **1996** (mit Begg C., Cho M., Eastwood S., Horton R., Moher D., Pitkin R., Ren nie D., Schulz KF., Simel D., Stroup DF) Verbesserung der Qualität der Berichterstattung über randomisierte kontrollierte Studien. Die CONSORT-Anweisung. Marmelade. Med. Assoc. 276(8), 637–639.

Einseitiger Test

Ein einseitiger oder einseitiger Test für einen **Populationsparameter** ist eine Art **Hypothesentest ,** bei dem die mit H0 bezeichneten Werte, für die wir die Nullhypothese ablehnen können, vollständig in einem Ende der Wahrscheinlichkeitsverteilung liegen.

MATHEMATISCHE ASPEKTE

Ein einseitiger Test an einer **Stichprobe** ist eine Art **Hypothesentest ,** bei dem es sich um **Hypothesen** handelt die Form:

(1) H0 : ÿ ÿ ÿ0

H1 : ÿ<ÿ0

oder

(2) H0 : ÿ ÿ ÿ0

H1 : ÿ>ÿ0 ,

Dabei ist H0 die **Nullhypothese,** H1 die **Alternativhypothese,** ÿ ein unbekannter **Parameter** der **Grundgesamtheit** und ÿ0 der angenommene **Wert** dieses Parameters.

Bei einem einseitigen Test an zwei Stichproben lauten die Hypothesen wie folgt:

(1) H0 : ÿ1 ÿ ÿ2

H1 : ÿ1 < ÿ2

oder

(2) H0 : ÿ1 ÿ ÿ2

H1 : ÿ1 > ÿ2 ,

wobei ÿ1 und ÿ2 die unbekannten Parameter zweier verglichener Populationen sind.

Der kritische Bereich, der als Wertesatz der Teststatistik definiert ist

Die Nullhypothese H0 wird abgelehnt, denn bei einem einseitigen Test ist die Wertemenge kleiner als die **kritischer Wert** des Tests oder die Gruppe von Werten, die größer als der **kritische Wert** des Tests sind.

BEISPIELE

Einseitiger Test am Bevölkerungsmittelwert

Ein Parfümhersteller möchte sicherstellen, dass seine Flaschen tatsächlich mindestens 40 ml Parfüm enthalten. Eine **Probe** von 50 Flaschen **ergibt durchschnittlich** 39 ml, mit **einer Standardabweichung** von 4 ml.

Die Hypothesen lauten wie folgt:

Nullhypothese H0 : ÿ ÿ 40 ,

Alternativhypothese H1 : ÿ < 40 .

Wenn die Stichprobengröße groß genug ist, kann die **Stichprobenverteilung** des Mittelwerts durch eine **Normalverteilung angenähert werden.** Für ein **Signifikanzniveau** ÿ = 1 % erhalten wir den **Wert** *von zÿ* in der **Normaltabelle:** zÿ = 2,33. Der **kritische Wert** der **Nullhypothese** wird durch den Ausdruck berechnet:

$$ÿx^- ÿ zÿ \cdot ÿx^-,$$

wobei $ÿx^-$ der Mittelwert der Stichprobenverteilung der Mittelwerte ist

$$ÿx^- = ÿ$$

Ö

Machine Translated by Google

und $\ddot{y}x^{-}$ ist die Standardabweichung der Stichprobenverteilung der Mittelwerte **oder der Standardfehler** des Mittelwerts

$$\ddot{y}x^{-} = \frac{\ddot{y}}{\ddot{y}n} .$$

Wir erhalten somit den folgenden kritischen Wert:

$$40 \ddot{y} 2{,}33 \cdot \frac{4}{\sqrt{50}} = 38{,}68$$

Da der Stichprobenmittelwert $x^{-} = 39$ größer als der kritische Wert für die einseitige Linke ist

Test kann der Hersteller die Nullhypothese nicht ablehnen und kann daher davon ausgehen, dass seine Parfümflaschen tatsächlich mindestens 40 ml Parfüm enthalten, wobei der Mittelwert der Stichprobe zufällig verteilt ist.

Einseitiger Test des Anteils zweier Bevölkerungsgruppen

Der Antragsteller eines Gesetzentwurfs zum Straßenverkehr geht davon aus, dass sein Gesetzentwurf von der Stadtbevölkerung eher akzeptiert wird als von der Landbevölkerung. Es wurden zwei Stichproben von 100 Personen aus städtischen und ländlichen Umgebungen befragt. Im städtischen Umfeld (Bevölkerung 1) reagierten 82 Personen positiv auf seinen Gesetzentwurf, während im ländlichen Umfeld (Bevölkerung 2) nur 69 Personen positiv reagierten. Um den Verdacht des Antragstellers zu bestätigen oder zu entkräften, formulieren wir einen **Hypothesentest ,** der die folgenden Hypothesen aufstellt:

Nullhypothese H0 : $\ddot{y}1 \ddot{y} \ddot{y}2$ oder $\ddot{y}1 \ddot{y} \ddot{y}2 \ddot{y} 0$ **Alt.**

Hypothese H1 : $\ddot{y}1 > \ddot{y}2$ oder $\ddot{y}1 \ddot{y} \ddot{y}2 > 0$

wobei $\ddot{y}1$ und $\ddot{y}2$ den günstigen Anteil der städtischen bzw. ländlichen Bevölkerung darstellen.

Abhängig vom gemessenen Prozentsatz der beiden Proben *(p1* = 0,82 und p2 = 0,69),

Wir können den **Wert** des Standards abschätzen Abweichung der Stichprobenverteilung der prozentualen Differenz bzw. Standardfehler der prozentualen Differenz:

$$\ddot{y}\hat{}p1\ddot{y}p2 = \sqrt{\frac{p1 \cdot (1 \ddot{y} p1)}{n1} \quad \frac{p2 \cdot (1 \ddot{y} p2) +}{n2}}$$

$$= \sqrt{\frac{0{,}82 \cdot (1\ddot{y}0{,}82)}{100} \quad +\frac{0{,}69 \cdot (1\ddot{y}0{,}69)}{100}}$$

$$= 0{,}06012 .$$

Wenn wir einen Test mit einem **Signifikanzniveau** ÿ von 5 % entwerfen, beträgt der **Wert** von zÿ in der **Normaltabelle** 1,645. Der kritische Wert der Nullhypothese wird berechnet durch:

$$\ddot{y}1 \ddot{y} \ddot{y}2 + z\ddot{y} \cdot \ddot{y}p1\ddot{y}p2 = 0$$
$$+ 1{,}645 \cdot 0{,}06012 = 0{,}0989 .$$

Denn der Unterschied zwischen dem Beobachteten Wenn die Proportionen p1 ÿ p2 = 0,82 ÿ 0,69 = 0,13 größer als der kritische Wert sind, müssen wir die Nullhypothese zugunsten der Alternativhypothese ablehnen und damit die Annahme des Vorschlagenden bestätigen.

WEITERLESEN

- **Akzeptanzregion**
- **Alternative Hypothese**
- **Hypothesentest**
- **Nullhypothese**
- **Ablehnungsbereich**
- **Stichprobenverteilung**
- **Signifikanzniveau**
- **Zweiseitiger Test**

VERWEISE

Miller, RG, Jr.: Simultaneous Statistical Inference, 2. Aufl. Springer, Berlin Heidelberg New York (1981)

Machine Translated by Google

Einfaktorielle Varianzanalyse

Die einfaktorielle Varianzanalyse für einen **Faktor** mit *t*- **Werten** (oder *t*- **Behandlungen)** ist eine Technik, mit deren Hilfe festgestellt werden kann, ob ein Faktor vorhanden ist ein signifikanter Unterschied zwischen den *Behandlungen* (Populationen). Das Prinzip der Einweg-Varianzanalyse besteht darin, die zu vergleichen Variabilität innerhalb von **Proben ,** die aus dem entnommen wurden **Populationen** mit der Variabilität zwischen diese Proben. Die Variationsquelle heißt **„Fehler“** entspricht der Variabilität in den **Stichproben** und der Quelle der Variation Der sogenannte „Effekt“ entspricht der Variabilität zwischen den Proben.

GESCHICHTE

Siehe **Varianzanalyse.**

MATHEMATISCHE ASPEKTE

Das lineare **Modell** für einen **Faktor** mit *t*- **Stufen (Behandlungen)** ist wie folgt:

$$Yij = ÿ + ÿi + ÿij ,$$
$$i = 1, 2,...,t, j = 1, 2,...,ni ,$$

wobei *Yij* **die Beobachtung** jreceiving darstellt **Treatmenti,** ÿ ist der allgemeine **Mittelwert** Bei allen Behandlungen ist *ÿi* der tatsächliche Effekt der Behandlung, die Beobachtungen zu erwähnen, *und ÿij ist* der experimentelle **Fehler** der Beobachtung *Yij.* Dieses **Modell** unterliegt den Grundlagen **Hypothesen** im Zusammenhang mit **der Varianzanalyse,** wenn wir annehmen, dass die Fehler *ij* sind unabhängige **Zufallsvariablen** folgen eine **Normalverteilung** N(0, ÿ2). Um zu sehen, ob es einen Unterschied zwischen den gibt *Bei den* **Behandlungen** wird ein **Hypothesentest** durchgeführt. Die folgende **Nullhypothese:**

$$H0 : ÿ1 = ÿ2 = \cdots = ÿt$$

bedeutet, dass die *t*- Behandlungen identisch sind.

Die **Alternativhypothese** wird wie folgt formuliert: folgt:

H1 : nicht alle **Werte** von *ÿi (i = 1, 2,...,t)* sind identisch.

Das Prinzip der **Varianzanalyse** besteht darin Vergleichen Sie die Variabilität innerhalb jeder **Stichprobe** mit der Variabilität zwischen den Proben. Für Hierzu wird der **Fisher-Test** verwendet, bei dem ein Verhältnis ermittelt wird gebildet, dessen Zähler eine **Schätzung** ist der **Varianz** zwischen Behandlungen (Stichproben) und dessen Nenner eine Schätzung von ist die Varianz innerhalb der Behandlungen.

Dieses mit F bezeichnete Verhältnis folgt einem **Fisher Verteilung** mit tÿ1 undN *ÿt* **Graden von Freiheit (N** ist die Gesamtzahl der **Beobachtungen).** Die **Nullhypothese** H0 wird sein auf dem **Signifikanzniveau** ÿ abgelehnt, wenn *F* Das Verhältnis ist größer oder gleich dem **Wert** im **Fisher-Tisch,** also wenn

$$F ÿ Ftÿ1,Nÿt,ÿ .$$

Wenn **die Nullhypothese H0** nicht abgelehnt werden kann, Dann können wir davon ausgehen, dass **die Proben** kommen aus der gleichen **Bevölkerung.**

Berechnung der Varianz zwischen Behandlungen

Für die **Varianz** zwischen den Behandlungen die Summe der Quadrate zwischen den Behandlungen wird berechnet als folgt:

$$SSStr = \sum_{i=1}^{T} ni(Y¯i. ÿ Y¯..)^2 ,$$

wo *Y¯i.* ist der Mittelwert der i-ten Behandlung. Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** ist gleich *t* ÿ 1, wobei *t* ist Anzahl der **Proben** (oder Gruppen). Die Unterschiede zwischen den Behandlungen sind gleich Zu:

$$s^2_{Tr} = \frac{SSStr}{t ÿ 1} .$$

Ö

Machine Translated by Google

Berechnung der Varianz innerhalb der Behandlungen

Für die **Varianz** innerhalb der Behandlungen die Summe von Quadraten innerhalb von Behandlungen, auch genannt Fehler, muss wie folgt berechnet werden:

$$SSE = \sum_{i=1}^{T} \sum_{j=1}^{ni} (Yij - \bar{Y}i.)^2 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** ist gleich *N*ÿ*t*, wobei *N* die ist Gesamtzahl der **Beobachtungen** und *t* die Anzahl Anzahl der **Proben.** Die **Varianz** innerhalb der Behandlungen ist dann gleich:

$$s^2_E = \frac{SSE}{N - t} .$$

Die Gesamtquadratsumme *(SST)* ist die Summe von Quadrate der **Abweichungen** jeder **Beobachtung** *Yij* vom allgemeinen **Mittelwert** *Y*¯*..*:

$$SST = \sum_{i=1}^{T} \sum_{j=1}^{ni} (Yij - \bar{Y}..)^2 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** beträgt *N*ÿ*1, wobei N* gilt Gesamtzahl der **Beobachtungen.** Wir können auch Finden Sie *SST* auf folgende Weise:

$$SST = SSTr + SSE .$$

Fisher-Test

Wir können jetzt das *F*- Verhältnis berechnen und abschrecken **Meiner Meinung nach** kann davon ausgegangen werden, dass die Proben derselben Population entnommen wurden . Das *F*- Verhältnis entspricht der **Schätzung** des **Varianz** zwischen den Behandlungen geteilt durch Schätzung der Varianz innerhalb von Behandlungen, Bedeutung

$$F = \frac{s^2_{Tr}}{s^2_E} .$$

Wenn *F* größer oder gleich dem **Wert** in ist **Fisher-Tabelle** für *t* ÿ 1 und *N* ÿ *t* **Grad der Freiheit,** dann können wir daraus schließen, dass die Unterschied zwischen den **Proben** ist darauf zurückzuführen die experimentellen **Behandlungen** und nicht nur zu Zufälligkeit.

Tabelle der Varianzanalyse

Es ist üblich, die Informationen zusammenzufassen **einer Varianzanalyse** in einer Tabelle namens an Tabelle zur Varianzanalyse, dargestellt wie folgt:

Quelle von va ration	Abschlüsse von Freiheit	Die Summe von Quadrate	Bedeutung von Quadrate	F
Unter behandeln gen	t-1	*SSStr*	s^2_{Tr}	$\frac{s^2_{Tr}}{s^2_E}$
Innerhalb behandeln gen	Nt	*SSE*	s^2_E	
Gesamt	N-1	*SST*		

Paarweiser Mittelwertvergleich

Wenn der Fishertest die **Nullhypothese ablehnt,** was bedeutet, dass es einen signifikanten Unterschied zwischen den **Mittelwerten** der **Stichproben gibt,** sagen wir Kann mich fragen, wo dieser Unterschied liegt. Sev Zur Beantwortung dieser Frage wurden verschiedene Tests entwickelt diese Frage. Es gibt den LSD- **Test** (Least **Signifikanter Unterschied)** (auf Französisch: *test de la différence minimale) ,* das macht Vergleiche paarweise ermittelter **Mittelwerte .** Für Damit dieser Test verwendet werden kann, muss das *F*- Verhältnis einen signifikanten Unterschied zwischen den anzeigen bedeutet.

BEISPIELE

Während des Kochens nehmen Croissants Fett auf variable Mengen. Wir wollen sehen, ob die Die aufgenommene Menge hängt von der Art ab

Machine Translated by Google

Fett. Es werden vier verschiedene Fettsorten zubereitet und pro Sorte sechs Croissants gebacken Fett.

Die **Daten** finden Sie in der folgenden Tabelle (die Zahlen beziehen sich auf die pro Jahr aufgenommene Fettmenge). Croissant):

Fett			
1	2	3	4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85	63	70
95	77	76	68

Die **Nullhypothese** besagt, dass es keinen Unterschied zwischen den Behandlungen gibt, was bedeutet, dass die Fettmenge, die beim Kochen aufgenommen wird ist unabhängig von der Fettsorte:

$$H0 : ÿ1 = ÿ2 = ÿ3 = ÿ4 .$$

Um diese **Hypothese zu testen,** dient der **Fisher-Test** verwendet, für die das *F* -Verhältnis ermittelt werden muss. Daher die **Varianz** zwischen den Behandlungen und die Varianz innerhalb der Behandlungen muss berechnet werden.

Unterschiede zwischen den Behandlungen

Der allgemeine **Mittelwert** $Y¯..$ und die Mittelwerte der vier Sätze $Y¯i. = \frac{1}{ni}\sum_{j=1}^{ni} Y1i$ *(i = 1, 2, 3,* 4) haben berechnet werden. Da alle *ni* gleich 6 sind und *N* gleich 24 ist, erhalten wir:

$$Y¯.. = \frac{64 + 72 + ... + 70 + 68}{24}$$

$$= \frac{1770}{24}$$

$$= 73{,}75 .$$

Wir berechnen auch:

$$Y¯1. = \frac{1}{6}\sum_{j=1}^{6} Y1j = = \frac{432}{6} = 72 ,$$

$$Y¯2. = \frac{1}{6}\sum_{j=1}^{6} Y2j = = \frac{510}{6} = 85 ,$$

$$Y¯3. = \frac{1}{6}\sum_{j=1}^{6} Y3j = = \frac{456}{6} = 76 ,$$

$$Y¯4. = \frac{1}{6}\sum_{j=1}^{6} Y4j = = \frac{372}{6} = 62 .$$

Dann können wir die Summe der Quadrate berechnen unter den Behandlungen:

$$SSStr = \sum_{i=1}^{4} ni(Y¯i. ÿ Y¯..)^2$$

$$= 6(72 ÿ 73{,}75)^2 + 6(85 ÿ 73{,}75)^2$$

$$+ 6(76 ÿ 73{,}75)^2 + 6(62 ÿ 73{,}75)^2$$

$$= 1636{,}5 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** beträgt 4 ÿ 1 = 3.

Die Unterschiede zwischen den Behandlungen sind gleich Zu:

$$s^2_{Tr} = \frac{SSStr}{t ÿ 1} = \frac{1636{,}5}{3} = 545{,}5 .$$

Varianz innerhalb der Behandlungen

Die Summe der Quadrate innerhalb von Behandlungen oder Fehlern ist gleich:

$$SSE = \sum_{i=1}^{4}\sum_{j=1}^{6} (Yij ÿ Y¯i.)^2$$

$$= (64 ÿ 72)^2 + (72 ÿ 72)^2 + ... +$$

$$(70 ÿ 62)^2 + (68 ÿ 62)^2$$

$$= 2018 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** beträgt 24 ÿ 4 = 20.

Ö

Machine Translated by Google

Die **Varianz innerhalb** der Behandlungen ist dann gleich Zu:

$$s_E^2 = \frac{SSE}{N \ddot{y} t} = \frac{2018}{20} = 100{,}9 .$$

Fisher-Test

Alle zur Berechnung von *F* notwendigen Elemente sind nun bekannt:

$$F = \frac{s_{Tr}^2}{s_E^2} = \frac{545{,}5}{100{,}9} = 5{,}4063 .$$

Wir müssen den **Wert** der **Fisher-Tabelle** finden

mit 3 und **20 Freiheitsgraden mit** einem **Signifikanzniveau** ÿ = 0,05:

$$F3.20,0.05 = 3.10 .$$

Das merken wir

$$F \ddot{y} F3,20,0,05 ,$$

$$5{,}4063 \ddot{y} 3{,}10 ,$$

was bedeutet, dass H0 verworfen werden muss.

Dies bedeutet, dass es einen erheblichen Unterschied zwischen den **Behandlungen gibt.** Deshalb

Die aufgenommene Fettmenge hängt davon ab

die Art des verwendeten Fettes.

Tabelle der Varianzanalyse

Diese Informationen sind im Folgenden zusammengefasst

Tabelle der **Varianzanalyse:**

Quelle von Variation	Abschlüsse von Freiheit	Die Summe von Quadrate	Bedeutung von Quadrate	F
Unterbehandelngen	3	1636,5	545,5	5.4063
Innerhalb behandelngen	20	2018	100,9	
Insgesamt	23	3654,5		

WEITERLESEN

- Varianzanalyse
- Kontrast
- Fisher-Tisch
- Fisher-Test
- Test auf den geringsten signifikanten Unterschied
- Zwei-Wege-Varianzanalyse

VERWEISE

Cox, DR, Reid, N.: Theorie des Designs von Experimenten. Chapman & Hall, London (2000)

Montgomery, DC: Design und Analyse von Experimente, 4. Aufl. Wiley, Chichester (1997)

Sheffé, H.: Die Varianzanalyse. C'n'H (1959)

Unternehmensforschung

Operations Research ist ein Bereich der angewandten Forschung Mathematik, die wissenschaftliche Methoden verwendet, um eine notwendige Entscheidungsgrundlage liefern. Es wird im Allgemeinen auf den Komplex angewendet Probleme der Personen- und Ausrüstungsorganisation, um die beste Lösung zu finden irgendein Ziel. Eingeschlossen in Operations Research Methoden sind **Simulation, lineare Programmierung,** mathematische Programmierung (nichtlinear), Spieltheorie usw.

GESCHICHTE

Das ursprüngliche Ziel des Operations Research sollte in den 1930er Jahren im Vereinigten Königreich militärische Anwendungen finden. Diese Militärs Anwendungen wurden dann weitgehend erweitert, und Seit dem Zweiten Weltkrieg wurde der Einsatz der Operationsforschung auf Nicht-Operationen ausgeweitet militärische öffentliche Anwendungen und sogar an die privater Sektor. Die Auswirkungen auf den Betrieb

Machine Translated by Google

Die Erforschung der rasanten Entwicklung von Computern war sehr wichtig.

BEISPIELE

Optimierung, lineare Programmierung und **Simulation** sind Beispiele für Operationen Forschung.

WEITERLESEN

- **Lineares Programmieren**
- **Optimierung**
- **Simulation**

VERWEISE

Bealle, EML: Einführung in die Optimierung. Wiley, New York (1988)

Collatz, L., Wetterling, W.: Optimierung Probleme. Springer, Berlin Heidelberg New York (1975)

Optimales Design

Ein optimales Design ist ein **Experiment Design ,** das bestimmte Kriterien der Optimalität erfüllt (zum Beispiel die Minimierung von Fehler im Zusammenhang mit den von uns vorgenommenen Schätzungen muss machen). Wie die Behandlungen angewendet werden Verschiedene Versuchseinheiten werden durch festgelegt Der Experimentator, wir können versuchen, das zu verbessern Qualität der Ergebnisse. Wir können auch zunehmen Die Präzision der Schätzer verringert sich Konfidenzintervalle ändern oder die Leistung erhöhen des Hypothesentests, Auswahl des optimalen Designs innerhalb der **Kategorie ,** zu der es gehört gehört.

GESCHICHTE

Die Theorie optimaler Designs wurde von **Kiefer, Jack** (1924–1981) entwickelt . Einer von Die Hauptpunkte seiner Arbeit basieren auf die statistische Effizienz des Designs. In seinem Artikel von 1959 stellt er die wichtigsten Kriterien zur Erzielung des optimalen Designs vor. Kiefers Die gesammelten Artikel im Bereich experimentelles Design wurden von Springer veröffentlicht die Zusammenarbeit des Instituts für Mathe matisch Statistik.

DOMÄNEN UND EINSCHRÄNKUNGEN

Da wir nicht alle Fehler **minimieren** können Gleichzeitig wird das optimale Design durch die Wahl der Optimalitätskriterien bestimmt. Diese unterschiedlichen Kriterien werden in offenbart Kiefers Artikel von 1959.

WEITERLESEN

- **Versuchsplanung**

VERWEISE

Brown, LD, Olkin, I., Sacks, J., Wynn, HP: Jack Carl Kiefer, Gesammelte Papiere III: Versuchsplanung. Springer, Berlin Heidelberg New York (1985)

Fedorov, VV: Theorie optimaler Experimente. Akademisch, New York (1972)

Kiefer, J.: Optimale experimentelle Designs. J. Roy. Stat. Soc. Ser. B **21,** 272–319 (1959)

Optimierung

Eine große Anzahl statistischer Probleme kann mithilfe von Optimierungstechniken gelöst werden. Wir verwenden Optimierungstechniken in statistischen Verfahren wie **kleinste Quadrate, maximale Wahrscheinlichkeit,** L1- **Schätzung** usw. In In vielen anderen Bereichen **der Statistik** wie der Regressionsanalyse, dem Testen von Hypothesen, dem experimentellen Design usw. spielt die Optimierung eine versteckte, aber wichtige Rolle.

Machine Translated by Google

Optimierungsmethoden können in klassifiziert werden
Folgende Weise:

1. Klassische Optimierungsverfahren (Differentialrechnung, Lagrange-Multiplikatoren)
2. Mathematische Programmiermethoden **(lineare Programmierung,** nichtlineare Programmierung, dynamische Programmierung)

GESCHICHTE

Die Probleme der Optimierung wurden von Euklid formuliert, allerdings erst mit der Entwicklung der Differentialrechnung und der Variationsrechnung im 17. und 18 Jahrhunderte waren mathematische Werkzeuge fähig zur Lösung solcher Probleme zur Verfügung. Diese Methoden wurden zum ersten Mal eingesetzt die Lösung bestimmter Optimierungsprobleme in Geometrie und Physik.
Probleme dieser Art wurden formuliert und untersucht von bekannten Mathematikern wie z wie Euler, Leonhard **Bernoulli, Jakob,** Jaco bi, Carl Gustav Jacob und Lagrange, Joseph Louis, dem wir den Lagrange-Multiplikator verdanken.

Die Entdeckung wichtiger Anwendungen von die Optimierungstechnik bei militärischen Problemen während des Zweiten Weltkriegs und Fortschritte in der Technologie (Computer) erlaubt für die Lösung immer komplexer werdender Aufgaben Probleme der Optimierung in den verschiedensten Formen Felder.

WEITERLESEN

- **Lagrange-Multiplikator**
- **Lineares Programmieren**
- **Unternehmensforschung**

VERWEISE

Arthanari, TS, Dodge, Y.: Mathematik Programmierung in der Statistik. Wiley, Neu York (1981)

Ausreißer

Ein Ausreißer ist eine Beobachtung, die gut vom Rest der Daten getrennt ist.

Ausreißer werden in Bezug auf eine angenommene zugrunde liegende Verteilung definiert oder theoretisch Modell. Wenn sich diese ändern, kann die Beobachtung Seien Sie nicht mehr abseits.

GESCHICHTE

Das Problem der Ausreißer wurde behoben von vielen Wissenschaftlern seit 1750 **von Boscovich, Roger Joseph, de Laplace, Pierre Simon,** und **Legendre, Adrien Marie.**
Laut Stigler, Stephen M. (1973), Legendre, AM schlug 1805 die Ablehnung von Ausreißern vor; und im Jahr 1852 etablierte Peirce, Ben Jamin, ein erstes Kriterium zur Abschreckung von Ausreißern im Bergbau. Es gab Kritik, insbesondere von Airy, George Biddell (1856), gegen die Verwendung eines solchen Kriteriums.

DOMÄNEN UND EINSCHRÄNKUNGEN

Im Rahmen statistischer Studien können wir Formulieren Sie die Frage nach Ausreißern auf zwei verschiedene Arten:

- Die erste besteht aus der Anpassung von Methoden sind resistent gegen das Vorhandensein von Ausreißern in die **Probe.**
- Der Zweite versucht, die Ausreißer zu beseitigen. Löschen Sie sie nach der Identifizierung aus dem Probe, bevor Sie eine statistische Analyse durchführen.

Das Löschen von Ausreißern aus einem Datensatz ist ein Nachteil umstrittenes Problem, insbesondere in einem kleinen Datensatz. Das Problem der Ausreißer führte zur Forschung von neue statistische Methoden namens robustes Meth Odds. Schätzer reagieren nicht empfindlich auf Ausreißer soll robust sein.

Machine Translated by Google

WEITERLESEN

Analyse von Residuen
Hutmatrix
L1 -Schätzung
Hebelpunkt
Überwachung
Robuste Schätzung

VERWEISE

Airy, GB: Brief von Professor Airy, Astronomer Royal, an den Herausgeber. Astronom. J. 4, 137–138 (1856)

Barnett, V., Lewis, T.: Ausreißer in statistischen Daten. Wiley, New York (1984)

Peirce, B.: Kriterium für die Ablehnung zweifelhafter Beobachtungen. Astron. J. 2, 161–163 (1852)

Stigler, S.: Simon Newcomb, Percy Daniell und die Geschichte der robusten Schätzung 1885–1920. Marmelade. Stat. Assoc. **68,** 872–879 (1973)

Machine Translated by Google

Paasche-Index

Der Paasche- **Index** ist ein **zusammengesetzter Index** **Anzahl** der Preise, die durch die Gewichtungs- und Summenmethode ermittelt wurden. Diese Indexnummer entspricht entspricht dem Verhältnis der Summe der Preise des tatsächlichen Zeitraums *n* und der Summe der Preise des Bezugszeitraums 0, diese Summen betragen gewichtet mit den jeweiligen Mengen der tatsächlicher Zeitraum.

Der Paasche-Index unterscheidet sich vom **Laspeyres-Index** nur durch die Wahl von Gewichtungsmethode. Tatsächlich in den Laspeyres Index, die Gewichte werden anhand von Mengen angegeben des Referenzzeitraums Q0 und nicht jene der aktuellen Periode *Qn.*

GESCHICHTE

Mitte des 19. Jahrhunderts deutscher Statistiker Paasche, Hermann entwickelte eine Formel für die **Indexnummer ,** die seinen Namen trägt.

Paasche, H. (1874) beschäftigte sich mit den in Hamburg erfassten Preisen.

MATHEMATISCHE ASPEKTE

Der Paasche- **Index** wird wie folgt berechnet:

$$In/0 = \frac{Pn \cdot Qn}{P0 \cdot Qn},$$

wobei *Pn* und *Qn* jeweils die Preise sind und verkaufte Mengen in der aktuellen Periode und P0 und Q0 sind die Preise und verkauften Mengen im Bezugszeitraum. Die Beträge beziehen sich auf die gelten als Güter und werden in Basis ausgedrückt 100.

$$In/0 = \frac{Pn \cdot Qn}{P0 \cdot Qn} \cdot 100 .$$

Das Paasche-Modell kann auch verwendet werden, um einen Mengenindex (auch Volumen genannt) zu berechnen Index). In diesem Fall sind es die Preise Konstante und die Größen, die variabel sind:

$$In/0 = \frac{Qn \cdot Pn}{Q0 \cdot Pn} \cdot 100 .$$

BEISPIELE

Betrachten Sie die folgende Tabelle, die Folgendes angibt jeweilige Preise für drei Lebensmittel im Referenzjahr 0 und im aktuellen Jahr n als die im laufenden Jahr verkauften Mengen:

Produkt	Verkaufte Menge im Jahr 1988 (Qn) (Tausende)	Preis (Euro) 1970 (P0)	1988 (Pn)
Milch	85,5	0,20	1,20
Brot	50,5	0,15	1.10
Butter	40,5	0,50	2,00

Aus der folgenden Tabelle haben wir:

$$PnQn = 239,15 \text{ und}$$

$$P0Qn = 44,925 .$$

Machine Translated by Google

Produkt	PnQn	P0 Qn
Milch	102,60	17.100
Brot	55,55	7.575
Butter	81,00	20.250
Gesamt	239.15	44.925

Wir können dann den Paasche-Index finden:

$$In/0 = \frac{Pn \cdot Qn}{P0 \cdot Qn} \cdot 100$$

$$= \frac{239.15}{44.925} \cdot 100 = 532,3 .$$

Mit anderen Worten, laut Paasche
Index ist die Preisindexzahl dieser Produkte im
Betrachtungszeitraum um 432,3 % (532,3ÿ100)
gestiegen.

WEITERLESEN

Zusammengesetzte Indexnummer
Fisher-Index
Indexnummer
Laspeyres-Index
Einfache Indexnummer

VERWEISE

Paasche, H.: Über die Preisentwicklung
der letzten Jahre nach dem Hamburger
Börsen-Notirungen. Jahrb. Natl. Stat. **23,**
168–178 (1874)

Paar Zufallsvariablen

Ein Variablenpaar , dessen **Werte** durch ein
Zufallsexperiment ermittelt werden , heißt
ein Paar Zufallsvariablen. Es gibt zwei
Arten von Paaren:

- Ein Paar **Zufallsvariablen** ist *diskret,* wenn
  Die Menge der Werte, die jede der Zufallsvariablen
  annimmt, ist eine endliche oder unendliche Anzahl
  fähiges Set.
- Ein Paar Zufallsvariablen ist *stetig*
  wenn die Menge der von jeder Zufallsvariablen
  angenommenen Werte eine unendliche, nicht zählbare Zahl ist
  Satz.

Im Fall eines Paars von Zufallsvariablen ist die
Die Menge aller möglichen Ergebnisse des Experiments
umfasst den **Probenraum**
befindet sich in einem zweidimensionalen Raum.

MATHEMATISCHE ASPEKTE

Im Allgemeinen handelt es sich um ein Paar **Zufallsvariablen**
gekennzeichnet durch Großbuchstaben wie X und Y.
Es handelt sich um eine Funktion mit zwei Variablen und wahren **Werten**
in 2.

Wenn *A* eine Teilmenge des **Probenraums ist,** dann gilt:

$$P(A) = \textbf{Wahrscheinlichkeit ,} \text{ dass } (X, Y) \ddot{y} A$$

$$= P[(X, Y) \ddot{y} A] .$$

Der Fall eines Paares von **Zufallsvariablen** kann
auf den Fall von *n* Zufallsvariablen verallgemeinert
werden : Wenn zur Beschreibung *n* Zahlen notwendig sind
die Durchführung eines **Zufallsexperiments,**
Eine solche Erkenntnis wird durch *n* zufällig dargestellt
Variablen X1, *X2,...,Xn.* Der **Probenraum**
liegt dann in einem n-dimensionalen Raum. Wenn
A ist eine Teilmenge des Probenraums, dann können wir
Sag das:

$$P(A) = P[(X1, X2,...,Xn) \ddot{y} A] .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Bestimmte zufällige Situationen zwingen einen zum Betrügen
sider nicht eine, sondern mehrere numerische Enti
Krawatten gleichzeitig. Bedenken Sie zum Beispiel:
ein System aus zwei Elementen in einer Reihe und
von zufälligen Lebensspannen *X* und Y; das Studium der
Die Lebensdauer des gesamten Satzes muss gleichzeitig
Ereignisse berücksichtigen, die *X* und *Y* betreffen
Zeit. Oder betrachten Sie einen Zylinder, hergestellt von
eine Maschine mit Durchmesser *X* und Höhe Y, mit

Machine Translated by Google

Diese Werte sind zufällig. Die auferlegten Toleranzen berücksichtigen in der Regel beide Einheiten gleichzeitig; die Struktur ihrer Wahrscheinlichkeit Abhängigkeit sollte bekannt sein.

WEITERLESEN

- Kovarianz
- Funktion der Gelenkdichte
- Randdichtefunktion
- Zufällige Variable

VERWEISE

Hogg, RV, Craig, AT: Einführung zu Mathematische Statistik, 2. Aufl. Macmillan, New York (1965)

T-Test für gepaarte Schüler

Der gepaarte Student-Test ist ein Hypothesentest das wird verwendet, um die Mittelwerte von zwei zu vergleichen Populationen, wenn jedes Element einer Population mit einem Element des anderen in Beziehung steht eins.

MATHEMATISCHE ASPEKTE

Sei *xij* die **Beobachtung** j für das Paari(j = *1,* 2 und *i = 1, 2,...,n).* Für jedes Beobachtungspaar berechnen wir die Differenz

$$di = xi2 \ddot{y} xi1 .$$

Wir suchen den geschätzten **Standard Fehler** des **Mittelwerts** von *di,* bezeichnet mit $\bar{d}$:

$$sd^{-} = \frac{1}{\ddot{y}n} \cdot Sd ,$$

wobei *Sd* die **Standardabweichung** von *di ist:*

$$Sd = \sqrt{\frac{\sum_{i=1}^{N} di \ddot{y} d^{-2}}{n \ddot{y} 1}} .$$

Der resultierende statistische Test ist definiert durch:

$$T = \frac{D}{sd^{-}} .$$

Hypothesen

Der Paired Student's Test **ist ein zweiseitiger Test.**

Die Hypothesen sind:

H0 : ÿ = 0 (es gibt keinen Unterschied unter den Behandlungen)

H1 : ÿ = 0 (es gibt einen Unterschied unter den Behandlungen) ,

wobei ÿ die Differenz zwischen den Mittelwerten ist von zwei Populationen (ÿ = ÿ1 ÿ ÿ2).

Entscheidungsregeln

Wir lehnen die **Nullhypothese zur Signifikanz** ab **Ebene** ÿ wenn

$$|T| > tn\ddot{y}1, \ddot{y}_{\overline{2}} ,$$

wobei $tn\ddot{y}1, \ddot{y}_{\overline{2}}$ ist der **Wert** des **Studenten Tabelle** mit *n* ÿ 1 Freiheitsgraden.

BEISPIELE

Angenommen, es werden zwei Behandlungen durchgeführt zehn Beobachtungspaare. Die erhaltenen Daten und die entsprechenden Unterschiede bezeichnet von *di* sind in der folgenden Tabelle dargestellt:

Paar	Behandlung		di = xi 2ÿ xi 1
	1	2	
1	110	118	8
2	99	104	5
3	91	85	ÿ6
4	107	108	1
5	82	81	ÿ1
6	96	93	ÿ3
7	100	102	2
8	87	101	14
9	75	84	9
10	108	111	3

P

Machine Translated by Google

Der **Mittelwert** beträgt:

$$d^{-} = \frac{1}{10}\sum_{i=1}^{10} di = \frac{32}{10} = 3{,}2.$$

Die **Standardabweichung** wird folgendermaßen berechnet:

$$Sd = \sqrt{\frac{\sum_{i=1}^{10} di \ddot{y} d^{-}}{10 \ddot{y} 1}^{2}} = \sqrt{\frac{323{,}6}{9}} = 6{,}00 \quad ,$$

und der **Standardfehler:**

$$sd^{-} = \frac{1}{\sqrt{\ddot{y}n\ 1}} \cdot Sd$$

$$= \frac{}{\sqrt{\ddot{y}10}} \cdot 6{,}00 = 1{,}90 .$$

Der statistische Test ergibt dann:

$$T = \frac{D}{sd^{-}} = \frac{3.2}{1{,}90} = 1{,}69 .$$

Wenn wir ein **Signifikanzniveau** ÿ = 0,05 wählen, Der **Wert** von *t9.0,025* beträgt 2,26. Also die **Null Hypothese** H0 : ÿ = 0 kann nicht verworfen werden weil *|T| < t9.0,025.*

WEITERLESEN

- **Hypothesentest**
- **Studententisch**
- **Schülertest**

Panel

Das Panel ist eine Art **Volkszählung ,** die sich wiederholt regelmäßig. Es basiert auf einer dauerhaften (oder semipermanente) **Stichprobe** von Einzelpersonen, Haushalte usw., die regelmäßig zu ihrem Verhalten oder ihrer Meinung befragt werden. Der Panel bietet den Vorteil, dem zu folgen individuelles Verhalten der befragten Personen bzw Einheiten und messen alle Änderungen in ihren Verhalten im Laufe der Zeit. Die Informationen sammeln Die im Panel erfassten Daten sind häufig umfangreicher als bei einer einfachen Volkszählung, wobei die Kosten begrenzt bleiben. Wir Passen Sie die Probenerneuerungsrate je nach Bedarf an das Ziel der Untersuchung. Wenn das Ziel von Das Gremium soll die Entwicklung von Einzelpersonen im Laufe der Zeit verfolgen (z. B. die Verfolgung von Karrierewegen). die Probe bleibt gleich; Wir nennen dies eine Längsuntersuchung. Wenn das Ziel darin besteht dienen der Charakteristikabschätzung der **Bevölkerung** in verschiedenen Zeiträumen, die Das Muster wird von Zeit zu Zeit teilweise erneuert Zeit; Wir nennen dies ein Wechselpanel. Und so kam es dass der Daten des Panels stellen eine Quelle von sehr dar wichtige Informationen, weil sie enthalten eine individuelle und zeitliche Dimension gleichzeitig mäßig.

GESCHICHTE

Die Geschichte der Panelforschung reicht bis zurück 1759, als der französische Graf von Montbeil, Philibert Guéneau, mit der Aufnahme seines Schaffens begann Die Statur der Söhne erfolgt in Abständen von sechs Monaten ab der Geburt bis zum Alter von 18 Jahren. Abgesehen davon haben seine Aufzeichnungen wenig mit zeitgenössischen Panelstudien zu tun aus der Systematik seiner Wiederholungen Beobachtungen. Das aktuelle Konzept des Panels Die Forschung wurde etwa in den 1920er Jahren etabliert und 1930er Jahre, als mehrere Studien über den Menschen durchgeführt wurden Wachstum und Entwicklung begannen. Seit der Panelstudien haben in der Mitte des 20. Jahrhunderts in den Sozialwissenschaften stark zugenommen.

Mehrere Faktoren haben zum Wachstum des Panels beigetragen Forschung in den Sozialwissenschaften. Viel Wachstum in den 1960er und 1970er Jahren kam eine Reaktion zu einer Erhöhung der Bundesmittel für Panelstudien in den Vereinigten Staaten. Beispielsweise führten politische Bedenken hinsichtlich der Arbeit und des Wirtschaftsverhaltens zur Finanzierung laufender Studien wie der National Longitudinal Surveys (NLS) im Jahr 1966 und die Panel Study of Income

Machine Translated by Google

Dynamics im Jahr 1968. Letzteres begann beispielsweise mit 4802 Familien mit einer Liste von Variablen über 5000; arme Haushalte wurden überbewertet. An der NLS-Umfrage nahmen 5.020 ältere Männer, 5.225 junge Männer, 5.083 reife Frauen, 5.159 junge Frauen und 12.686 Jugendliche teil.

Technologische Fortschritte haben die Datenerfassung auf nationaler Ebene erleichtert und zugenommen

Die Fähigkeit von Forschern, große, komplexe Datensätze zu verwalten, zu analysieren und zu teilen. Darüber hinaus haben Fortschritte bei den statistischen Methoden zur Analyse von Längsschnittdaten Forscher dazu ermutigt, nicht nur Paneldaten zu sammeln, sondern auch neue Fragen zu vorhandenen Datensätzen zu stellen. Die Zunahme von Panelstudien ist auch in anderen Industrieländern zu beobachten. zu.

DOMÄNEN UND EINSCHRÄNKUNGEN

In der Praxis kommt es selten vor, dass es Panels im engeren Sinne des Begriffs gibt. Wir sollten daher die folgenden Probleme des Panels berücksichtigen: Schwierigkeiten bei der Rekrutierung bestimmter Kategorien von Panelisten, Ermüdung der Panelisten, die einer der Hauptgründe für Nichtantworten ist, und die daraus resultierende Entwicklung der **Bevölkerung** die Abweichung der **Stichprobe** von der Grundgesamtheit, die sie darstellen sollte. In der Praxis versuchen wir oft, dieses Problem zu beseitigen, indem wir ein wechselndes Panel einem strikten vorziehen.
Es gibt auch spezifische Messfehler in Panels: (1)

Teleskopeffekt: Datierungsfehler, der häufig von neuen Panelteilnehmern gemacht wird, wenn sie das Datum des Ereignisses angeben, nach dem sie gefragt werden.

(2) *Paneleffekt* oder *Konditionierungsbias:* Verhaltensänderung von Panelteilnehmern, die im Laufe der Zeit als Ergebnis wiederholter Interviews auftreten kann. Die Diskussionsteilnehmer werden zu ihrem Verhalten befragt und schließlich ändern sie es.

BEISPIELE

Umfrage zu Gesundheit und medizinischer Versorgung: Umfrage des INSEE in etwa 10.000 Haushalten. Jeder Haushalt wird überwacht

3 Monate, in denen alle 3 Wochen Haushaltsmitglieder von einem Forscher zu ihrer medizinischen Versorgung befragt werden.

Sofres-Verbraucherpanel: Das Metascope-Panel von Sofres basiert auf einer **Stichprobe** von 20.000 Haushalten. Jeden Monat erhalten die Diskussionsteilnehmer per Post einen selbst ausgefüllten Fragebogen zu ihren verschiedenen Einkäufen. Diese Fragebögen werden im Wesentlichen von Unternehmen verwendet, die wissen möchten, was ihre Kunden wollen.

Zielgruppe: Das Mediamat-Panel basiert auf einer Stichprobe von 2300 Haushalten, die mit Audimetern mit Drucktasten ausgestattet sind.
Es ermöglicht Forschern, das Publikum zu messen von Fernsehsendern.

WEITERLESEN

Voreingenommenheit

Bevölkerung

Probe

Umfrage

P

REFERENZEN

Duncan, GJ: Haushaltspanelstudien: Perspektiven und Probleme. Sozialwissenschaftliche Methodologie, Trient, Italien (1992)

Kasprzyk, D., Duncan, G. et al. (Hrsg.): Panel Surveys. New York, Chichester, Bris Bane, Toronto, Singapur. Wiley, New York (1989)

Rose, D.: Haushaltspanelstudien: ein Überblick. Innovation 8(1), 7–24 (1995)

Diggle, PJ, P. Heagerty, K.-Y. Liang und SL Zeger.: Analyse von Längsschnittdaten, 2. Aufl. Oxford University Press, Oxford (2002)

Machine Translated by Google

Parameter

Ein Parameter charakterisiert einen quantitativen Aspekt einer **Population.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Parameter einer **Population** sind oft unbekannt. Wir können jedoch einen Parameter anhand einer **Statistik** schätzen, die mithilfe einer Schätzmethode aus einer **Stichprobe** berechnet wird .

BEISPIELE

Ein Parameter wird im Allgemeinen mit einem griechischen Buchstaben bezeichnet:

ÿ: Mittelwert der **Grundgesamtheit ÿ: Standardabweichung** der Grundgesamtheit **ÿ: Prozentsatz** relativ zur Grundgesamtheit

WEITERLESEN

- **Einschätzung**
- **Schätzer**
- **Kleinsten Quadrate**
- **Maximale Wahrscheinlichkeit**
- **Moment**
- **Statistiken**

VERWEISE

Lehmann, EL: Theory of Point Estimation, 2. Aufl. Wiley, New York (1983)

Parametrischer Test

Ein parametrischer Test ist eine Form des **Hypothesentests ,** bei dem Annahmen über die zugrunde liegende Verteilung der beobachteten Daten getroffen werden.

GESCHICHTE

Einer der ersten parametrischen Tests war der Chi-Quadrat-Test, der 1900 von Pearson, K. eingeführt wurde.

BEISPIELE

Der **Student-Test** ist ein Beispiel für einen Parametertest. Ziel ist es, die Mittelwerte zweier normalverteilter Populationen zu vergleichen.
Zu den bekanntesten parametrischen Tests zählen der **Student-t-Test** und der **Fisher**

prüfen.

WEITERLESEN

- **Binomialtest**
- **Fisher-Test**
- **Hypothesentest**
- **Schülertest**

Partielle Autokorrelation

Die partielle Autokorrelation bei Verzögerung *k* ist die Autokorrelation zwischen *Yt* und Ytÿk , die nicht durch die Verzögerungen 1 bis *k* ÿ 1 erklärt wird .

Siehe **Autokorrelation.**

Teilweise Korrelation

Die teilweise Korrelation zwischen zwei Variablen wird als Korrelation zweier Variablen definiert

Fähigkeiten, während ein Drittel oder mehr andere Variablen kontrolliert werden. Ein *Maß für die teilweise Korrelation* zwischen den Variablen *X* und *Y* hat ein Drittel Variable *Z* als Maß für die direkte Rela

tion zwischen *X* und *Y* , die nicht dauert

Berücksichtigen Sie die Konsequenzen der linearen Beziehungen dieser beiden Variablen mit Z. Die

Berechnung der Teilkorrelation zwischen *X* und *Y* bei gegebenem *Z* dienen dazu, zu identifizieren, in welchem Maße die lineare Beziehung zwischen *X* und *Y* besteht kann auf die Korrelation zweier Variablen zurückzuführen sein mit Z.

GESCHICHTE

Die Interpretation von Analysen von Korrelationskoeffizienten zwischen zwei Variablen wurde zu einem

Machine Translated by Google

wichtige Frage in der Statistik während der zunehmenden Verbreitung von Korrelationsmethoden im frühen 20. Jahrhundert. Pearson und Karl wussten, dass eine große Korrelation zwischen zwei Variablen auf deren Korrelation zurückzuführen sein könnte mit einer dritten Variablen. Dieses Phänomen wurde erst 1926 erkannt, als George Udny es anhand eines Beispiels bewies, indem er die Korrelationskoeffizienten zwischen Zeitreihen ermittelte.

MATHEMATISCHE ASPEKTE

Aus einer Stichprobe von *n* Beobachtungen,(x1, y1, z1), (x2, y2, z2),...,(xn, *yn,* zn)aus einer unbekannten Verteilung von drei Zufallsvariablen X, Y und Z, die *Der Teilkorrelationskoeffizient,* den wir mit *rxy.z* bezeichnen , ist als der berechnete **Korrelationskoeffizient** definiert zwischenxi undyi mit

$$xi = \ddot{y}\ 0x + \ddot{y}\ 1xzi\ ,$$

$$yi = \ddot{y}\ 0y + \ddot{y}\ 1yzi\ ,$$

Dabei sind ÿ *0x* und ÿ *1x* die **Kleinste-Quadrate** -Schätzer, die man durch eine Regression von *xi* auf *zi* erhält , und ÿ *0y* und ÿ *1y* sind die Kleinste-Quadrate-Schätzer, die man durch eine Regression von *yi* auf *zi* erhält . Somit haben wir per Definition:

$$rxy.z = \frac{(xi \ddot{y}\ \bar{x})\overline{(yi \ddot{y}\ y)}}{\overline{(xi \ddot{y}\ \bar{x})^{2}}\ \overline{(yi \ddot{y}\ y)^{2}}}\ .$$

Wir können das folgende Ergebnis beweisen:

$$rxy.z = \frac{rxy \ddot{y}\ rxz \cdot ryz}{\overline{1 \ddot{y}\ r2_{xz}}\ \cdot\ \overline{1 \ddot{y}\ r2_{yz}}}\ ,$$

Dabei sind *rxy, rxz* und *ryz* die Korrelationskoeffizienten zwischen *xi* und *yi, xi* und *zi* sowie *yi* und *zi.* Beachten Sie, dass wir Folgendes haben, wenn *xi* und *yi* nicht mit *zi* korrelieren , d. h. wenn *rxz* = *ryz* = 0 ist:

$$rxy.z = rxy\ ,$$

Das heißt, der Korrelationskoeffizient entspricht dem Koeffizienten der teilweisen Korrelation, oder mit anderen Worten: Der Wert der Korrelation zwischen *xi* und *yi* ist überhaupt nicht auf das Vorhandensein von *zi zurückzuführen.* Wenn andererseits die Korrelationen mit *zi* wichtig sind, dann gilt *rxy.z* ÿ= 0, was darauf hindeutet, dass die beobachtete Korrelation zwischen *xi* und *yi* nur auf die Korrelation zwischen diesen Variablen mit zurückzuführen i *zi.* Daher sagen wir, dass zwischen *xi* und *yi* keine direkte Beziehung besteht (sondern eine indirekte Beziehung über zi).

Allgemeiner ausgedrückt können wir die partielle Korrelation zwischen *xi* und *yi* mit *zi1* und *zi2* relativ zu zwei Variablen Z1 und Z2 auf folgende Weise berechnen:

$$rxy.z1z2 = \frac{rxy \cdot z1 \ddot{y}\ rxz2.z1 \cdot ryz2.z1}{\overline{1 \ddot{y}\ r2_{xz2.z1}}\ \cdot\ \overline{1 \ddot{y}\ r2_{yz2.z1}}}\ .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Wir müssen darauf bestehen, dass eine Korrelation besteht misst nur die lineare Beziehung zwischen zwei Variablen, ohne funktionale Modelle oder die Vorhersage- oder Prognosekapazität eines bestimmten Modells zu berücksichtigen.

P

BEISPIELE

Wir haben beispielsweise eine starke positive Korrelation zwischen der Anzahl der registrierten Infarkte während eines bestimmten Zeitraums und der im gleichen Zeitraum verkauften Eiscrememenge beobachtet. Wir sollten nicht zu dem Schluss kommen, dass ein hoher Eiskonsum Infarkte provoziert oder umgekehrt, dass ein Infarkt ein besonderes Verlangen nach Eis hervorruft. In dieser Prüfung Die große Zahl der Infarkte ist jedoch nicht auf die Menge der verkauften Eiscreme zurückzuführen aber von der Hitze. Wir haben also eine starke Korrelation zwischen der Anzahl der Infarkte und die Temperatur. Die große Menge an

Machine Translated by Google

Die Temperatur des verkauften Eises lässt sich auch durch die Hitze erklären, was einen starken Zusammenhang zwischen der Menge des verkauften Eises und der Temperatur impliziert Temperatur. Die starke Korrelation zwischen der Anzahl der Infarkte und der Die Menge an verkauftem Eis ist dann die Konsequenz Folge zweier starker Korrelationen. Hinter der scheinbaren Beziehung der beiden vorherigen Variablen verbirgt sich eine dritte Variable.

WEITERLESEN

Korrelationskoeffizient

REFERENZEN

Yule, GU (1926) Warum erhalten wir manchmal unsinnige Korrelationen zwischen Zeitreihen? Eine Studie über Stichproben und die Natur von Zeitreihen. J. Roy. Stat. Soc. **(2)** 89, 1– 64

Teilweise am wenigsten absolut Abweichungsregression

Die partielle LAD-Regression (partielle kleinste absolute Abweichung) ist eine Regressionsmethode, die einen Antwortvektor mithilfe abgeleiteter Komponenten linear mit einer Reihe von Prädiktoren in Beziehung setzt. Dabei handelt es sich um eine L1-Regressionsmodellierungstechnik, die sich gut für Situationen eignet, in denen die Anzahl der Parameter in einem linearen Regressionsmodell die Anzahl der Beobachtungen übersteigt. Es wird hauptsächlich für Vorhersagezwecke und nicht für Rückschlüsse auf Parameter verwendet. Die Methode der partiellen LAD-Regression erweitert die **partielle Regression der kleinsten Quadrate** auf die L1- Norm, die mit der LAD-Regression verknüpft ist, statt auf die L2- Norm, die mit der partiellen Methode der kleinsten Quadrate verknüpft ist. Die partielle LAD-Regression folgt der Struktur des univariaten partiellen Regressionsalgorithmus der kleinsten Quadrate und extrahiert Komponenten (bezeichnet mit t) aus Richtungen (bezeichnet mit w), die von abhängen Antwortvariable. Die Richtungen werden durch eine Gnanadesikan-Ketteterning (GK)-Kovarianzschätzung bestimmt, die die übliche Varianz basierend auf der L2- Norm durch MAD, die mittlere absolute Abweichung, basierend auf L1, ersetzt. Daher verwenden wir für partielle KQP-Richtungen die Notation wmad_k .

GESCHICHTE Die partielle LAD-Regressionsmethode wurde von Dodge, Y. et al. eingeführt. (2004). Es wurde von Dodge, Y. et al. weiterentwickelt. (2004) und getestet mit dem BootstrapinKondylis,A. und Whittaker, J. (2005).

MATHEMATISCHE ASPEKTE

Für $i = 1,...,n$ und $j = 1,...,p$, die Beobachtungseinheiten und Prädiktoren bezeichnen, ist der partielle LAD-Regressionsalgorithmus unten angegeben: 1. Zentrieren oder standardisieren Sie sowohl **X** als auch y.

2. Für $k = 1,...,kmax:$

   Berechnen Sie wmad_k gemäß

   $$\text{wmad}_{j,k} = \frac{1}{4}\,\textbf{mad2(xj,kÿ1} + \text{y)}$$

   $$\ddot{y}\text{mad2}\,(xj,k\ddot{y}1\ \ddot{y}\ y)\ ,$$

   und wmad auf 1 $_k$ skalieren.

   Komponente extrahieren

   $$tk = \sum_{j=1}^{P} \text{wmad}_{j,k}\, xj,k\ddot{y}1\ ,$$

   Wo

   $$\text{wmad}_k = \text{wmad}\ ,...,\text{wmad}_{1,k_{j,k}} \quad .$$

   Orthogonalisieren Sie jedes $xj,k\ddot{y}1$ bezüglich $tk: xj,k = xj,k\ddot{y}1\ \ddot{y}\ L(xj,k\ddot{y}1|tk)$.

3. Geben Sie die resultierende Folge der plattierte Vektoren $_k$ **y** Anpassung = **Tkqk an,** wobei $Tk = (t1,..., tk)$ die Bewertungsmatrix und $qk = (q1,...,qk)$ der LAD-Regressionskoeffizientenvektor ist.

Machine Translated by Google

4. Stellen Sie das implizite partielle LAD-Regres- Plad wieder her = sionskoeffizienten nach ÿ W $_{k}^{\text{zerrückt}}$ *qk*, wobei die MatrixW ihre $_{k}^{\text{bin verrückt}}$ Pools in Spalten die Vektoren wmad sind ausgedrückt Begriffe des Originals *xj*.

Die univariate partielle Regression der kleinsten Quadrate und die partielle LAD-Methode haben das gleiche folgende Eigenschaften:

1. **y** = **q1t1** + ... + **qktk** + 2. **X** = , **p1t1** + ... + **pktk** + *f* ,

3. cor(ti, *tj)* = 0 für *i* = j,

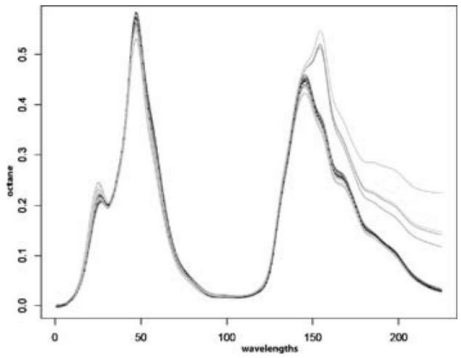
wobei cor(,) die Pearson-Korrelation bezeichnet Koeffizient und *f* entsprechen dem Residuum Begriffe und *pk* die X-Ladung.

Die partielle LAD-Methode erstellt ein Regressionsmodell, das die Prädiktoren mit dem in Beziehung setzt Antwort gemäß:

$$\mathbf{j}_{k}^{\text{plad}} = \sum_{j=1}^{P} \ddot{y}_{J}^{\text{plad}} xj .$$

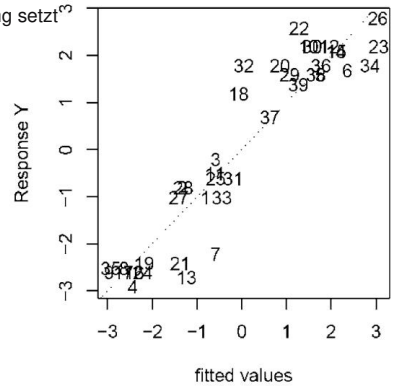
BEISPIEL

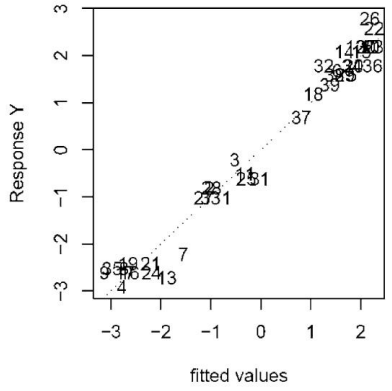
Wir geben hier ein Beispiel aus dem nahen Infrarot Experimente in der Spektroskopie. Wir benutzen das Oktanzahl-Datensatz, der aus 39 Benzinleitungsproben besteht, für die die Oktanzahl angegeben ist wurde bei 225 Wellenlängen (Zoll) gemessen Nanometer). Die Regressoren sind hochgradig multikollineare Spektren mit unterschiedlicher Anzahl Wellenlängen, und ihre Anzahl überschreitet die Stichprobengröße n.



In der Abbildung geben wir die Linien für die 39 an Benzinproben über ihre gesamten 225Wellenlängen.

Wir verwenden die partielle LAD-Regressionsmethode ein lineares Modell basierend auf zwei abgeleiteten zu erstellen Komponenten. Die resultierende Handlung des Antwort **y** im Vergleich zu den angepassten Werten **y** wird angegeben im rechten Bereich der folgenden Abbildung unter dem Haupttitel PLAD (teilweise LAD Regression). Das rechte Feld derselben Abbildung ist das entsprechende Diagramm der Antwort **y** vs. die angepassten Werte y , die für eine Univariate angegeben sind Basierend auf einem partiellen Regressionsmodell der kleinsten Quadrate auf zwei Komponenten.





Das partielle Regressionsmodell der kleinsten Quadrate wird eine zusätzliche Komponente in der erforderlich sein

P

Machine Translated by Google

endgültiges Modell, um gute Vorhersageergebnisse wie die partielle LAD-Regression zu liefern Modell. Dies ist auf eine Gruppe von sechs Ausreißern zurückzuführen im Oktan-Datensatz. Das sind Beobachtungen 25, 26, 36, 37, 38 und 39, die Alkohol enthielten. Sie sind im Liniendiagramm für Wellenlängen über 140 nm sichtbar über.

WEITERLESEN

- **Regression der kleinsten absoluten Abweichung**
- **Partielle Regression der kleinsten Quadrate**

VERWEISE

Dodge, Y., Kondylis, A., Whittaker, J.: Erweiterung der PLS1- auf die PLAD-Regression und die Verwendung der L1-Norm im Soft-Mod elling. COMPSTAT 2004, S. 935–942, Physica/Verlag-Springer (2004)

Dodge, Y., Whittaker, J., Kondylis, A.: Fortschritte bei PLAD und PQR. In: Tagungsband 55. Sitzung des ISI 2005, S. 495–503 (2005)

Kondylis, A., Whittaker, J.: Verwendung des Boot-Straps bei der PLAD-Regression. PLS und Verwandte Methoden. In: Proceedings of the PLS'05 Internationales Symposium 2005, S. 395–402. Edition Aluja et al. (2005)

Teilweise kleinste Quadrate Rückschritt

Die partielle Regression der kleinsten Quadrate (PLSR) ist eine statistische Methode, die zwei Daten in Beziehung setzt Matrizen **X** und Y, üblicherweise *Blöcke genannt,* über alatentlineare Struktur. Es wird entweder auf angewendet ein einzelner Antwortvektor **y** (univariate PLSR) oder eine Antwortmatrix **Y** (multivariate PLSR). PLSR wird üblicherweise bei der Aufzeichnung verwendet Variablen sind stark korreliert. Es ist ein sehr geeigneteMethodein Fällen, in denendie Anzahl der Die Variablen überschreiten die Anzahl der verfügbaren gute Beobachtungen. Dies liegt daran, dass es verwendet stattdessen orthogonal abgeleitete Komponenten der ursprünglichen Variablen. Die Verwendung einiger weniger orthogonale Komponenten statt zahlreich korrelierte Variablen garantieren die Reduzierung des Regressionsproblems auf einem kleinen Unterraum, der häufig die Variabilität von stabilisiert die geschätzten Koeffizienten und liefert bessere Vorhersagen.

GESCHICHTE

Das PLSR wurde ursprünglich innerhalb von entwickelt der NIPALS-Algorithmus, obwohl es so war implementiert in verschiedenen Algorithmen, einschließlich der orthogonalen Scores und der orthogonalen Belastung PLSR, dem SIMPLS-Algorithmus für PLSR, der Helland PLSR-Algorithmus und der Kernel-PLSR-Algorithmus.

MATHEMATISCHE ASPEKTE

PLSR-Methoden lösen das folgende Maximierungsproblem:

$$\max_{w,q} \{cov(Xkÿ1wk, \mathbf{Ykÿ1qk})\}$$

unterliegen

$$w^T_k wk = 1, q^T_k qk = 1$$

Und

$$tk ÿ tj, uk ÿ uj \text{ für jedes } k = j,$$

wobei cov(·, ·) die Kovarianz bezeichnet und *wk, qk* und *tk, uk* sind Ladevektoren und Komponenten (oder Score-Vektoren) für *X* bzw. Y. Die Verwendung des Index *k* in **X** und **Y** zeigt, dass PLSR die Daten jeweils entleert extrahierte Dimension *k* und verwendet Residuen als neue Daten für die folgende Dimension.

Machine Translated by Google

Univariate PLSR ist viel einfacher zu interferieren
pret. Tatsächlich handelt es sich um eine Verallgemeinerung
der **multiplen linearen Regression ,** die die
Regressionskoeffizienten in Richtungen niedriger Kovarianz verkleinert
Beziehung zwischen Variablen und Antworten.Gegeben
die folgende Beziehung:

$$\text{cov}(Xw, y) \ddot{y} \text{cor}(Xw, y)\text{var}(X)^{1/2},$$

wobei *cor* und *var* Korrelation bezeichnen und
Varianz bzw. Varianz ist leicht zu überprüfen
dass das Maximierungskriterium im PLSR
ist ein Kompromiss zwischen der **Methode der kleinsten Quadrate**
Regression und **grundsätzliche** Regression
Komponenten. Die letzteren Methoden maximieren
cor(Xw, y) bzw. var(X) .
Das univariate PLSR beinhaltet Folgendes
Schritte:

1. Für **(X,** *y)gemeinsam zentriert:*
2. Für *k = 1,...,p:*
   Speichern Sie *cov(xj,kÿ1,* ykÿ1) auf dem Vektor *wk.*
   Pool *wk* in Matrix W(p×k).
   Extrahieren Sie die Komponente *tk* als *tk* = Xj,kÿ1wk.
   Orthogonalisieren Sie Xkÿ1 und ykÿ1 mit
   Respekt vor *tk.*
3. Nehmen Sie die Residuen der kleinsten Quadrate als neue Daten
   für *k* ÿ *k* + 1.

Sobald die Komponenten extrahiert sind, wird die
Für die Regression wird **die Regression der kleinsten Quadrate** verwendet
sie auf Antwortvektor y. Die Nummer
von Komponenten, die letztendlich sein sollten
„retained“ wird mithilfe von **Bootstrap** und definiert
Kreuzvalidierungsmethoden.

BEISPIEL

Die PLSR-Methode ist eine Verallgemeinerung **der**
multiplen linearen Regression. Das ist berechtigt
Verwendung eines Regressionsproblems, bei dem keine
Korrelation zwischen den Prädiktoren auftritt, d. h.
ein **orthogonales Design.** Die folgende Tabelle enthält
die Ergebnisse eines orthogonalen Designs mit
zwei erklärende Variablen und eine Antwort.

Tabelle: Orthogonales Design

y	x1	x2
18 ÿ2 4		
12	1	3
10	0 ÿ6	
16 ÿ1 ÿ5		
11	2	3
9	0 ÿ3	
11 ÿ1		6
8	1 ÿ1	
7	1	0
12 ÿ1 ÿ1		

Multiple lineare Regressionsanalyse für die
Daten in Tabelle 1 **(y** als Antwort und zwei
erklärende Variablen x1 und x2) führen zum
Folgende Schätzungen der kleinsten Quadrate:

$$\mathbf{y} = 11{,}40 \ddot{y} 1{,}8571x1 + 0{,}1408x2 .$$

Unter Verwendung der PLSR-Methode (Daten wurden
zunächst zentriert und skaliert) und Regression der
Antwort auf die abgeleiteten Komponenten, die wir sind
führte zu genau den gleichen Schätzungen mit nur
eine Komponente verwendet, also byy = t1 ·q1.
Wir bekommen es endlich

$$\mathbf{y} = \mathbf{T1q1} = Xw0q1 = X\ddot{y} \text{ PLS} .$$

Für standardisierte Daten gilt also das implizite
Regressionskoeffizienten sind

$$q\ddot{y} \text{ PLS} = (\ddot{y}0{,}68, 0{,}164),$$

was, zurück in die ursprüngliche Skala transformiert,
gleich (11, 40, ÿ1,8571, 0,1408) ist.
Letztere sind die multiple lineare Regression
Schätzungen.

DOMÄNEN UND EINSCHRÄNKUNGEN

PLSR-Methoden wurden häufig verwendet
in Nahinfrarot-Experimenten in der Spektroskopie.
Sie werden seit langem von Chemometrie-Experten zur
Durchführung multivariater Kalibrierungen verwendet

P

Machine Translated by Google

chemische Konzentrationen zu messen. PLSR
Methoden wurden in vielen statistischen Anwendungen
mit mehr Variablen gleichermaßen verwendet
Beobachtungen, zum Beispiel in der Umweltmetrik,
in Microarray-Experimenten in der Biostatistik,
und in der funktionalen Datenanalyse.
PLSR-Methoden konzentrieren sich hauptsächlich auf die Vorhersage
Genauigkeit und werden hauptsächlich verwendet, um
gute Vorhersagemodelle zu konstruieren.

WEITERLESEN

Bootstrap

Methode der kleinsten Quadrate

VERWEISE

Tenenhaus, M.: La régression PLS. Théorie
et pratique. Technip, Paris (1998)

Frank, I., Friedman, J.: Eine statistische Sicht
einiger chemometrischer Regressionswerkzeuge.
Technometrics **35,** 109–135 (1993)

Martens, H. und Naes, T.: Multivariate Kalibrierung.
Wiley, New York (1989)

Helland: Zur Struktur der PLS-Regression, Comm. in
Stat-Simul. und Comp.,
17, 581–607 (1988)

Pearson, Egon Form

Pearson, Egon Shape (1895–1980), Sohn von
Der Statistiker **Pearson, Karl,** studierte Mathematik
am Trinity College in Cambridge. In
1921 trat er in die Statistikabteilung **des** University
College in London ein. Er
lernte **Neyman, Jerzy,** kennen, mit dem er während
dessen Besuchs in London 1924–1925
zusammenzuarbeiten begann, und arbeitete auch mit
Gosset, William Sealy, zusammen. Wenn
Sein Vater verließ das University College im Jahr 1933
wechselte in die Abteilung für Angewandte Statistik
und wurde 1936 dort Direktor
nal *Biometrie* nach dem Tod von ihm
Vater, der Gründer der Zeitschrift.
Neben den Werken, die er in Zusammenarbeit mit
Neyman, J. veröffentlichte und die dazu führten
Pearson, ES ging auch auf die Entwicklung der
Theorie des Hypothesentests ein
Probleme der Qualitätskontrolle und **des Betriebs
Forschung.**

Ausgewählte Artikel von Pearson, ES:

1928 (mit Neyman, J.) Über die Verwendung und
Interpretation bestimmter Testkriterien
zum Zweck statistischer Schlussfolgerungen.
Biometrika 20A, 175–240, S. 263–
295.

1931 Der Signifikanztest für den
Korrelationskoeffizienten. Marmelade. Stat. Assoc.
26, 128–134.

1932 Die prozentualen Grenzen für die Verteilung
von Rangein-Proben aus einer
Normalpopulation *(n* ÿ 100). Biometrie ka 24,
404–417.

1933 Neyman, J. und Pearson, ES On
die Prüfung statistischer Hypothesen in
Beziehung zur Wahrscheinlichkeit a priori. Proz.
Camb. Philos. Soc. 29, 492–510.

1935 Die Anwendung statistischer Methoden
Vorteile für die industrielle Standardisierung und
Qualitätskontrolle. Brit. Stant. 600.
British Standards Institution, London.

WEITERLESEN

Hypothesentest

Pearson, Karl

Geboren in London, Pearson, Karl (1857–1936)
ist bekannt für seine zahlreichen Beiträge zu

Machine Translated by Google

Statistiken. Nach seinem Studium am Kings College in Cambridge wurde er 1885 zum Vorsitzenden der Abteilung für Angewandte Mathematik des University College in London ernannt. Er verbrachte seine gesamte Karriere an dieser Universität, wo er 1911 zum Leiter der Abteilung für Eugenik ernannt wurde. 1901 gründete er mit Hilfe von **Galton, Francis,** die Zeitschrift *Biometrika* und war bis zu seinem Tod deren Chefredakteur
1936.

Im Jahr 1906 begrüßte er für ein Jahr William Sealy in seinem Gosset-Labor , mit dem er Probleme im Zusammenhang mit Proben kleiner Abmessungen löste. Er ging 1933 in den Ruhestand und seine Abteilung an der Universität wurde in zwei Teile geteilt: Die Eugenik-Abteilung wurde **Fisher, Ronald Aylmer,** anvertraut **und die Statistik-** Abteilung seinem eigenen Sohn **Pearson, Egon Shape.**

Ausgewählte Artikel von Pearson, Karl:

1894 1948. Beiträge zur Mathematik
Mathematische Evolutionstheorie. I. S. 1–40. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, Cambridge. Erstveröffentlichung als: Zur Zerlegung asymmetrischer Frequenzkurven, Philos. Trans. Roy.
Soc. London. Ser. A 185, 71–110.

1895 1948. Beiträge zur Mathematik
Mathematische Evolutionstheorie. II. Schrägungsvariation in homogenem Material. In: Karl Pearsons frühe statistische Arbeiten. Cambridge University Press, Cambridge, S. 41–112. Erstveröffentlichung in Philos. Trans. Roy. Soc.
London. Ser. A 186, 343–414.

1896 1948. Mathematische Beiträge zu
die Evolutionstheorie. III. Regression, Vererbung und Panmixie. In: Karl Pearsons frühe statistische Arbeiten.
Cambridge University Press, Cam
Brücke, S. 113–178. Erstveröffentlichung in Philos. Trans. Roy. Soc. London. Ser.
A 187, 253–318 im Philosophical Magazine, 5. Reihe, 50, S. 157–175.

WEITERLESEN

Chi-Quadrat-Verteilung
Korrelationskoeffizient

Prozentsatz

Der Prozentsatz ist der Begriff einer Kennzahl, die es ermöglicht, den Anteil von Einzelpersonen oder statistischen Einheiten mit einem bestimmten Merkmal in einer Sammlung zu beschreiben, bewertet auf der Basis von 100.

MATHEMATISCHE ASPEKTE

Der Prozentsatz wird im Allgemeinen mit *p* bezeichnet, wenn er an einer **Stichprobe** gemessen wird, und mit ÿ, wenn er eine **Population betrifft.**
Lassen Sie es eine Stichprobe von *Größe geben, wenn k* Individuen bestimmte Merkmale aufweisen; Der Prozentsatz der Individuen ergibt sich aus der **Statistik** p:

$$p = \frac{k}{N} \cdot 100\ \% .$$

Die Berechnung des gleichen Prozentsatzes für eine Population der Größe N **ergibt den Parameter ÿ:**

$$ÿ = \frac{k}{N} \cdot 100\ \% .$$

Perzentil

Perzentile sind **Standortmessungen**

in einem Datensatz berechnet. Wir nennen Perzentile die **Werte, die** eine Verteilung durch 100 teilen

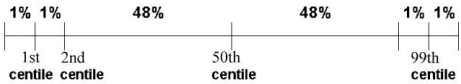
gleiche Teile (jeder Teil enthält die gleiche Anzahl an **Beobachtungen).** Das x-te Perzentil ist

P

Machine Translated by Google

der WertCx , sodass x % der Beobachtungen niedriger und (100ÿx) % der Beobachtungen größer als *Cx sind.* Beispielsweise ist das 20. Perzentil der Wert, der die 20 % der Werte, die darunter liegen, von den 80 % trennt.

die höher sind als es. Wir haben dann 99 Perzentile für eine gegebene Verteilung:



Beachten Sie zum Beispiel:
Perzentil 10 = erstes **Dezil**

Perzentil 25 = erstes **Quartil**
Perzentil 50 = **Median**

Dieser Begriff gehört zur Familie **der Quantile.**

MATHEMATISCHE ASPEKTE

Der Berechnungsprozess ähnelt dem des **Medians, der Quartile** und **der Dezile.** Wenn alle Rohbeobachtungen vorliegen, werden die Perzentile wie folgt berechnet :

1. Ordnen Sie die *n* **Beobachtungen** in Form einer **Häufigkeitsverteilung an.**
2. Die Perzentile entsprechen den **Beobachtungen,** für die sich der relative Wert kumuliert hat

   Die Häufigkeit überschreitet jeweils 1 %, 2 %, 3 %, ..., 98 %, 99 %.

   Einige Autoren schlagen Folgendes für Mula vor, das es ermöglicht, den **Wert** der verschiedenen Perzentile genau zu bestimmen: *Berechnung des j-ten Perzentils:*

   Betrachten Sie *i als* den ganzzahligen Teil von *j·(n+1) und k* als den gebrochenen Und Teil von *j·(n +1)* 100 . 100

   Betrachten Sie *xi* und xi+1 als die Datenwerte **der** Beobachtungen, **die** jeweils an der i-ten und *(i +* 1)-ten Position klassifiziert sind (wenn die *n* Beobachtungen in zunehmendem Maße sortiert werden Bestellung).

   Das j-te Perzentil ist gleich:

$$Cj = xi + k \cdot (xi+1 \text{ ÿ } xi) .$$

Wenn wir **Beobachtungen** haben , die es sind gruppiert in Klassen werden die Perzentile wie folgt ermittelt:

1. Bestimmen Sie die Klasse, in der die gewünschten sind Perzentil ist: •
   1. Perzentil: erste Klasse, für die die relative kumulierte **Häufigkeit** über 1 % liegt.
   - 2. Perzentil: erste Klasse, bei der die relative kumulierte Häufigkeit über 2 % liegt.
   - 99. Perzentil: Erstklasse, bei der die relative kumulierte Häufigkeit über 99 % liegt.
2. Berechnen Sie den **Wert** der Perzentile als Funktion **der Hypothese ,** dass die **Beobachtungen** in jeder Klasse gleichmäßig verteilt sind:

$$\text{Perzentil} = L1 + \frac{(n \cdot q) \text{ ÿ finf fcentile}}{} \cdot c ,$$

Dabei ist L1 die untere Grenze in der Klasse des Perzentils, n die Gesamtzahl der Beobachtungen, q das Zentil, q 99 , q **100** 1 100 für den 1. pro *die* Summe 2 100 für das 2. Perzentil ..., für das *der* —— 99. Perzentil ist finf Häufigkeiten , die niedriger **als** die Perzentilklasse sind, fperzentil die Häufigkeit der Perzentilklasse, und *c* ist die Dimension des **Intervalls** der Perzentilklasse.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Berechnung der Perzentile hat nur für eine **quantitative Variable** ihre wahre Bedeutung, die ihre **Werte** in einem bestimmten Intervall annehmen kann . In der Praxis können Perzentile nur bei einer großen Anzahl **von Beobachtungen** berechnet werden , da diese Berechnung aus der Division besteht

Machine Translated by Google

Satz von Beobachtungen in 100 Teile. Dieser Begriff wird selten angewendet, da in den meisten deskriptiven Analysen die Bestimmung der **Dezile** erfolgt oder sogar die **Quartile** reichen dafür aus Interpretation der Ergebnisse.

BEISPIELE

Betrachten Sie ein Beispiel für die Berechnung von Perzentile auf der **Häufigkeitsverteilung** einer kontinuierlichen **Variablen,** bei der der **Beobachter ationen** sind in Klassen gruppiert.
Die folgende **Häufigkeitstabelle** stellt dar der Gewinn (in Tausend Euro) des Jahres 2000 Bakterien:

Profitieren (Tausende in Euro)	Frei quen cies	Cumu verspätet Frequenzen	Relativ kumulieren ed fr quenz
100–150	160	160	0,08
150–200	200	360	0,18
200–250	240	600	0,30
250–300	280	880	0,44
300–350	400	1280	0,64
350–400	320	1600	0,80
400–450	240	1840	0,92
450–500	160	2000	1,00
Gesamt	2000		

Die Klasse, die das erste Perzentil enthält, ist die Klasse 100–150 [diejenige, für die die relative kumulierte Häufigkeit über 0,01 (1 %) liegt].
Unter Berücksichtigung der Tatsache, dass die **Beobachtungen** in jeder Klasse gleichmäßig verteilt sind, erhalten wir:
für das erste Perzentil der folgende **Wert:**

1. Perzentil

$$= 100 + \left(\frac{2000 \cdot \frac{1}{100} - 0}{160} \right) \cdot 50$$

$$= 106{,}25 .$$

Die Klasse, die das 10. Perzentil enthält, ist die Klasse 150–200. Der Wert des 10. Perzentils ist gleich

10. Perzentil

$$= 150 + \left(\frac{2000 \cdot \frac{10}{100} - 160}{200} \right) \cdot 50$$

$$= 160 .$$

Wir können alle Perzentile im berechnen gleicher Weg. Daraus lässt sich beispielsweise schließen, dass 1 % der 2000 Bäckereien einen Gewinn zwischen 100.000 und 106.250 Euro haben, bzw dass 10 % einen Gewinn zwischen 100.000 und haben 160000 Euro usw.

WEITERLESEN

- **Dezil**
- **Standortmaß**
- **Median**
- **Quantil**
- **Quartil**

P

Permutation

Der Begriff Permutation ist Gegenstand der **kombinatorischen Analyse.** Es bezieht sich auf eine Vereinbarung oder eine Ordnung von *n* Objekten. Theoretisch, wir sollte zwischen dem Fall unterscheiden, in dem *n* Objekte sind vollständig identifiziert und diejenigen, wo sie sind nur teilweise identifiziert.

GESCHICHTE

Siehe **kombinatorische Analyse.**

MATHEMATISCHE ASPEKTE

1. *Anzahl der Permutationen verschiedener Objekte*

Machine Translated by Google

Die Anzahl der möglichen Permutationen von *n* Objekte entsprechen *n* Fakultäten:

Permutationen

$= n! = n \cdot (n - 1) \cdot \ldots \cdot 3 \cdot 2 \cdot 1$.

2. *Anzahl der Permutationen teilweise unterschiedlicher Objekte*

Die Anzahl der möglichen Permutationen von *n* Objekte , darunter n1, n2,..., *nr* nicht unter ihnen unterschieden gleich:

$$\text{Anzahl von Permutationen} = \frac{N!}{n1! \cdot n2! \cdot \ldots \cdot Nr!},$$

wobei n1 + n2 + ... + *nr* = n.

BEISPIELE

Wenn wir drei Objekte A, B und C haben, dann ist das Mögliche Permutationen sind wie folgt:

ABC	
ACB	
BAC	
BCA	
TAXI	
CBA	

Somit haben wir sechs mögliche Permutationen. Ohne sie aufzuzählen, können wir die finden Anzahl der Permutationen unter Verwendung der folgenden Formel:

Anzahl der Permutationen = n! ,

wobei *n* die Anzahl der Objekte ist. Das gibt uns in unserem Beispiel:

Anzahl der Permutationen = 3!

$= 3 \cdot 2 \cdot 1 = 6$.

Stellen Sie sich vor, wir komponieren die Signale, indem wir einige Flags ausrichten. Wenn wir fünf Flaggen unterschiedlicher Farbe haben, erhöht sich die Anzahl der Signale, die wir haben komponieren kann wird sein:

$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Andererseits, wenn unter den Fahnen dort sind zwei rote, zwei weiße und eine schwarze Flagge, wir können komponieren:

$$\frac{5!}{2! \cdot 2! \cdot 1!} = 30$$

verschiedene Signale.

WEITERLESEN

Anordnung
Kombination

Kombinatorische Analyse

Piktogramm

Ein Piktogramm ist ein Symbol, das ein Konzept, einen Gegenstand, einen Ort oder ein Ereignis durch Illustration darstellt. Das Piktogramm wird wie das **Balkendiagramm** verwendet entweder um die **Kategorien** einer **qualitativen Variablen** zu vergleichen oder um Datensätze zu vergleichen , die aus verschiedenen Jahren oder verschiedenen Orten stammen. Piktogramme werden meist in Zeitschriften verwendet Werbegrafiken für Vergleiche und sind von geringem statistischem Interesse, weil sie liefern nur grobe Näherungswerte.

MATHEMATISCHE ASPEKTE

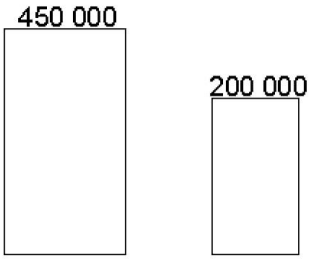
Die Figur, die zur Veranschaulichung der verschiedenen Größen in einem Piktogramm gewählt wird, sollte möglichst in vier Teile teilbar sein, damit sie kann Hälften und Viertel darstellen. Der **Maßstab** wird in Abhängigkeit vom verfügbaren Platz gewählt. Daher ist die Zahl nicht unbedingt gleich zur Einheit.

BEISPIELE

Betrachten Sie die Weizenproduktion in zwei Ländern A und B. Die Produktion von Land A ist entspricht 450.000 Doppelzentnern und die des Landes Bis von 200.000 Doppelzentnern (1 Doppelzentner entspricht bis 100kg).

Machine Translated by Google

Wir erhalten das folgende Piktogramm:



Die Oberfläche jeder Figur ist proportional zu die dargestellte Menge.

WEITERLESEN

Balkendiagramm

Grafische Darstellung

Quantitatives Diagramm

Kuchendiagramm

Das Kreisdiagramm ist eine **Art quantitatives Diagramm.** Es besteht aus einem Kreis, der jeweils in Sektoren unterteilt ist Sektor mit einem Winkel, der proportional ist die dargestellte Größe.

GESCHICHTE

Siehe **grafische Darstellung.**

MATHEMATISCHE ASPEKTE

Um ein Kreisdiagramm zu erstellen, wird die Summe der dargestellten **Häufigkeiten** berechnet und anschließend Die relativen Häufigkeiten, die die verschiedenen Sektoren repräsentieren, werden berechnet. Um diese zu zeichnen Sektoren auf einem Kreis, die relativen **Häufigkeiten** werden wie folgt in Grad umgerechnet Transformation:

$$\ddot{y} = f \cdot 360\ddot{y} \quad ,$$

wobei ÿ den Winkel in der Mitte darstellt und *f* die relative **Häufigkeit.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Kreisdiagramme dienen zur visuellen Darstellung **der Daten, die** die verschiedenen Teile von bilden eine ganze Bevölkerung. Man findet sie häufig in die Medien (Fernsehen, Zeitschriften, Magazine, usw.), wo sie der einfachen Erklärung dienen und synthetische Weise ein Konzept oder eine Situation, die ist allein mit Zahlen schwer zu verstehen.

BEISPIELE

Wir werden die Verteilung des Familienstandes in Australien am 30. Juni 1981 anhand von darstellen Ein Kreisdiagramm. Die **Daten** sind folgende:

Familienstand in Australien am 30. Juni 1981 (in Tausende)

Familienstand	Häufigkeit	Relativ Frequenz
Bachelor	6587.3	0,452
Verheiratet	6836,8	0,469
Geschieden	403,5	0,028
Witwe	748,7	0,051
Gesamt	14576.3	1.000

Quelle: Australian Bureau of Statistics, Australian Pocket Year Book 1984, S. 11

Zunächst werden die relativen **Häufigkeiten** in Grad umgerechnet:

a) ÿ = 0,452 · 360ÿ = 162, 72ÿ
b) ÿ = 0,469 · 360ÿ = 168, 84ÿ
c) ÿ = 0,028 · 360ÿ = 10, 08ÿ
d) ÿ = 0,051 · 360ÿ = 18, 36ÿ

Das Kreisdiagramm wird auf der nächsten Seite angezeigt.

WEITERLESEN

Balkendiagramm

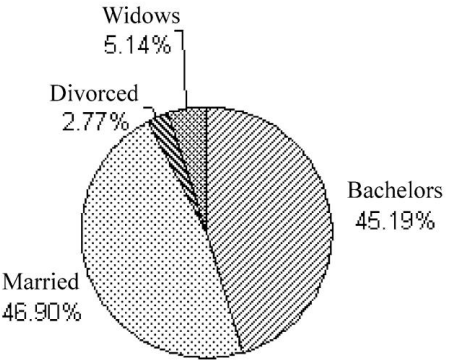
Grafische Darstellung

Piktogramm

Quantitatives Diagramm

P

Machine Translated by Google



Pitman, Edwin James George

Pitman, Edwin James George wurde in geboren
Melbourne (Victoria), Australien im Jahr 1897. Er
ging an der Kensington State School zur Schule
und South Melbourne College. Während seiner
Im letzten Studienjahr erhielt er ein Stipendium für ein
Mathematikstudium in Wyselaskie
und in Dixson sowie am Ormond College. Er erlangte
einen Abschluss in Briefen (ähnlich
zu einem BA) (1921), einem Abschluss in den Naturwissenschaften
(ähnlich einem BS) (1922) und einen Master
Abschluss (1923). Anschließend wurde er zum
vorübergehenden Professor für Mathematik in Canterbury ernannt
College, University of New Zealand (1922–)
1923). Er kehrte nach Australien zurück, wo er
wurde zum Assistenzprofessor am Trinity College ernannt
Ormond Colleges sowie nebenberuflicher Dozent für
Physik an der Melbourne University (1924–
1925). Im Jahr 1926 wurde Pitman, EJG benannt
Professor für Mathematik an der Tasmania University,
wo er bis zu seiner Pensionierung blieb
im Jahr 1962. Er starb 1993 in Kingston (Tas Mania).

Ein wichtiger Beitrag von Pitman, EJG dazu
Die Wahrscheinlichkeitstheorie befasst sich mit der Untersuchung von
das Verhalten einer charakteristischen Funktion in
eine Nachbarschaft von Null.

Ausgewählte Werke von Pitman, Edwin James George:

1937 Signifikanztests, die sein können
auf Proben aus beliebigen Populationen
angewendet. Beilage, J. Roy. Stat.
Soc. Ser. B 4, 119–130.

1937 Signifikanztests, die sein können
auf Proben aus beliebigen Populationen
angewendet werden. II. Der Korrelationskoeffizient
prüfen. Beilage, J. Roy. Stat. Soc.
Ser. B 4, 225–232.

1938 Signifikanztests, die sein können
gelten für Proben aus beliebigen Populationen.
III. Die Varianzanalyse
prüfen. Biometrie 29, 322–335.

1939 Die Schätzung von Standort und Umfang
Parameter. Biometrie 30, 391–
421.

1939 Tests von Hypothesen zu Standort- und
Skalenparametern. Biometrie ka 31, 200–215.

Punktschätzung

Punktschätzung eines **Populationsparameters**
ermöglicht es, einen eindeutigen berechneten **Wert** zu erhalten
aus einer **Probe.** Dieser Wert wird berücksichtigt
die **Schätzung** des unbekannten Parameters.

MATHEMATISCHE ASPEKTE

Betrachten Sie eine **Population ,** in der Merkmal *X* vorhanden ist
studiert; Angenommen, die Form der Verteilung von *X*
ist bekannt, der **Parameter** ÿ, von dem diese Verteilung
abhängt, ist jedoch bekannt
Unbekannt.

Um ÿ zu schätzen, wird eine **Stichprobe** der Größe *n* benötigt
(X1, *X2,...,Xn)* wird aus der **Population** entnommen und
eine Funktion *G(X1, X2,...,Xn)*
wird für eine bestimmte Erkenntnis geschaffen
(x1, *x2,...,xn)* liefert einen eindeutigen **Wert**

Machine Translated by Google

als **Schätzung** des Wertes des **Parameters** ÿ.

DOMÄNEN UND EINSCHRÄNKUNGEN

Im Allgemeinen entspricht die Punktschätzung eines **Parameters** einer **Population** nicht genau dem **Wert** dieses Parameters. Dort Es liegt ein **Stichprobenfehler** vor , der daraus resultiert die Tatsache, dass ein Teil der Bevölkerung gewesen ist weggelassen. Es ist möglich, diesen Fehler zu messen Berechnen der **Varianz** oder **Standardabweichung** des verwendeten **Schätzers**

Bewerten Sie den Parameter der Bevölkerung.

BEISPIELE

Ein Unternehmen, das Glühbirnen herstellt möchte die durchschnittliche Lebensdauer von untersuchen seine Glühbirnen. Es wird eine **Stichprobe** der Größe *n* genommen (X1, *X2,...,Xn)* aus der Lichterzeugung Glühbirnen.

Die Funktion *G(X1, X2,...,Xn)* , die es tun wird Verwenden Sie zur Schätzung des Parameters ÿ, der dem **Mittelwert** ÿ der **Grundgesamtheit** entspricht, wie folgt folgt:

$$X^{-} = \sum_{i=1}^{N} \frac{Xi}{N} .$$

Dieser **Schätzer** X¯ von ÿ ist ein **Zufall variabel** und für eine bestimmte Realisierung (x1, *x2,...,xn)* der **Stichprobe** nimmt einen genauen **Wert** an , der mit x¯ bezeichnet wird. Daher kann beispielsweise für *n* = 5 die folgende Stichprobe sein erhalten (in Stunden):

(812, 1067, 604, 918, 895) .

Der **Schätzer** x¯ wird zu:

$$x^{-} = \frac{812 + 1067 + 604 + 918 + 895}{5}$$

$$= \frac{4296}{5} = 859{,}2 .$$

Der **Wert** 859,2 ist eine Punktschätzung des durchschnittliche Lebensdauer dieser Glühbirnen.

WEITERLESEN

- **Fehler**
- **Einschätzung**
- **Schätzer**
- **Probe**
- **Probenahme**

VERWEISE

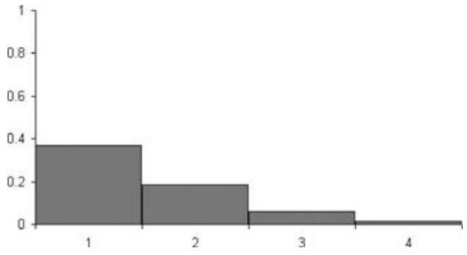
Lehmann, Erich L., Casella, G.: Theorie der Punktschätzung. Springer-Verlag, Neu York (1998)

Poisson-Verteilung

Die Poisson-Verteilung ist eine diskrete **Wahrscheinlichkeitsverteilung.** Es ist besonders nützlich für Phänomene des Zählens in Zeiteinheiten oder Raum.

Die **Zufallsvariable** *X* entspricht die Anzahl der pro Einheit beobachteten Elemente Zeit oder Raum folgen einer Poisson-Verteilung des Parameters ÿ, bezeichnet als P(ÿ), wenn seine **Wahrscheinlichkeitsfunktion** ist:

$$(X = x) = x! \qquad \frac{\exp(\ddot{y}\ddot{y}) \cdot \ddot{y}\, P^{x}}{} .$$



Poisson-Verteilung, ÿ = 1

GESCHICHTE

Im Jahr 1837 veröffentlichte **Poisson, SD** die nach ihm benannte Distribution. Er hat es erreicht

Verteilung unter Berücksichtigung der Grenzen der **Binomialverteilung.**

P

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Der **Erwartungswert** der Poisson-Verteilung ist per Definition:

$$
\begin{aligned}
E[X] &= \sum_{x=0}^{\ddot{y}} x \cdot P(X = x) \\
&= \sum_{x=0}^{\ddot{y}} x \cdot \frac{e^{\ddot{y}} \ddot{y}^{x}}{X!} \\
&= \sum_{x=1}^{\ddot{y}} x \cdot \frac{e^{\ddot{y}} \ddot{y}^{x}}{X!} \\
&= \ddot{y} \cdot e^{\ddot{y}} \sum_{x=1}^{\ddot{y}} \frac{\ddot{y}^{x\ddot{y}1}}{(x \ddot{y} 1)!} \\
&= \ddot{y} \cdot e^{\ddot{y}} \cdot e^{\ddot{y}} \\
&= \ddot{y} .
\end{aligned}
$$

Die **Varianz** von *X* ist gleich:

$$
\begin{aligned}
\mathrm{Var}(X) &= E\,X2^{2} \ddot{y} (E[X])^{2} \\
&= E[X(X \ddot{y} 1) + X] \ddot{y} (E[X])^{2} \\
&= E[X(X \ddot{y} 1)] + E[X] \ddot{y} (E[X])^{2} .
\end{aligned}
$$

Seit

$$
\begin{aligned}
E[X(X \ddot{y} 1)] &= \sum_{x=0}^{\ddot{y}} x(x \ddot{y} 1) \frac{e^{\ddot{y}} \ddot{y}^{x}}{X!} \\
&= \sum_{x=2}^{\ddot{y}} x(x \ddot{y} 1) \frac{e^{\ddot{y}} \cdot \ddot{y}^{x}}{X!} \\
&= e^{\ddot{y}} \cdot \ddot{y}^{2} \cdot \sum_{x=2}^{\ddot{y}} \frac{\ddot{y}^{x\ddot{y}2}}{(x \ddot{y} 2)!} \\
&= e^{\ddot{y}} \cdot \ddot{y}^{2} \cdot e^{\ddot{y}} \\
&= \ddot{y}^{2} ,
\end{aligned}
$$

wir erhalten:

$$
\mathrm{Var}(X) = \ddot{y}^{2} + \ddot{y} \ddot{y} \ddot{y}^{2} = \ddot{y} .
$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Zur Bestimmung wird die Poisson-Verteilung verwendet unter anderem folgende Veranstaltungen:

- Anzahl der von einem radioaktiven Stoff emittierten Partikel.
- Anzahl der von einem Cen aufgezeichneten Telefongespräche ter.
- Anzahl der Unfälle, die einem passieren versicherte Person.
- Anzahl der Ankünfte an einem Schalter.
- Anzahl der Bakterien im Mikroskop Vorbereitung.
- Anzahl der Pflanzen oder Tiere auf einer Oberfläche in der Natur von einem Beobachter bestimmt.

Wenn ÿ gegen Unendlich geht, ist die Poisson Die Verteilung kann durch die Normalverteilung **mit** Mittelwert **ÿ** und **Vari** angenähert werden **ance** ÿ.

BEISPIELE

Eine Sekretärin macht durchschnittlich zwei Fehler pro Seite. Wie hoch ist die Wahrscheinlichkeit zu haben? Drei Fehler auf einer Seite?
Es folgt die Anzahl der Fehler pro Seite,X eine Poisson-Verteilung des **Parameters** ÿ = 2, und seine **Wahrscheinlichkeitsfunktion** ist dann:

$$
\begin{aligned}
(X = x) &= x! \frac{\exp(\ddot{y}\ddot{y}) \cdot \ddot{y}^{X}}{P} \\
&= \frac{\exp(\ddot{y}2) \cdot 2x}{X!} .
\end{aligned}
$$

Die **Wahrscheinlichkeit ,** drei Fehler zu machen Eine Seite ist gleich:

$$
P(X = 3) = \frac{e^{\ddot{y}2 \cdot 23}}{3!} = 0{,}1804 .
$$

WEITERLESEN

Binomialverteilung

Diskrete Wahrscheinlichkeitsverteilung
Normalverteilung

VERWEISE

Poisson, SD: Recherchen zur Wahrscheinlichkeit von Straftaten im Strafrecht

Machine Translated by Google

et en matière Civile. In: Procédés des Allgemeine Regeln für die Wahrscheinlichkeitsrechnung. Bachelier, Imprimeur-Libraire pour les Mathématiques, Paris (1837)

Poisson, Siméon-Denis

Poisson, Siméon-Denis (1781–1840), obwohl er aus einer bescheidenen Provinzfamilie stammte, konnte sein Studium an der Ecole Polytechnique in Paris fortsetzen, wo er später Professor wurde.

Ab 1815 lehrte er auch an der Sorbonne und wurde im selben Jahr in die Akademie der Wissenschaften gewählt.

Poisson interessierte sich für die Forschung in verschiedenen Bereichen Fachgebiete: Mechanik, Physik und Wahrscheinlichkeitstheorie. Im Jahr 1837 veröffentlichte er einen Artikel mit dem Titel „Recherches sur la probabil ité des jugements en matière criminelle et en matière Civile, précédées des règles générales du calcul des probabilités“. In seinen zahlreichen Schriften legte er die Grundlagen einer statistischen Methode für die Sozialwissenschaften dar.

Ausgewählte Werke von Siméon-Denis Poisson:

1824 Über die Wahrscheinlichkeit der Ergebnisse über mehrere Monate hinweg. Connais sance des temps pour l'an 1827, S. 273–302.

1829 Suite du mémoire sur la probabil ité résultat moyen des observations, eingefügt in die Zeitkenntnis des Jahres 1827. Zeitgeschichte für das Jahr 1832, S. 3–22. **1836a** Note sur la loi des grands nombres.

Comptes rendus hebdomadaires des séances de l'Académie des sciences 2: S. 377–382.

1836b Hinweis zur Wahrscheinlichkeitsrechnung. Comptes rendus hebdomadaires des Sitzungen der Académie des Sciences 2: S. 395–400.

1837 Recherche über die Wahrscheinlichkeit von Urteilen im Kriminal- und im Zivilrecht, vor den allgemeinen Regeln der Wahrscheinlichkeitsrechnung. Bachelier, Imprimeur Libraire pour les Mathématiques, Paris.

WEITERLESEN

Poisson-Verteilung

Gepoolte Varianz

Die gepoolte **Varianz** wird verwendet, um den **Wert** der Varianz von zwei oder mehr Populationen zu schätzen, wenn die jeweiligen Varianzen jeder Population unbekannt sind, aber als gleich betrachtet werden können.

MATHEMATISCHE ASPEKTE

Auf der Basis von *k* Stichproben der Dimensionen n1, n2,..., nk wird die gepoolte **Varianz** S2 $_P$ ist eine Schätzung berechnet durch:

$$S^2_P = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{\sum_{i=1}^{k} n_i - k},$$

wobei *xij* die j-te **Beobachtung** der Probe *i* und x¯i ist . ist das **arithmetische Mittel** der Stichprobe i. Wenn man die jeweiligen Varianzen jeder Stichprobe kennt, kann die gepoolte Varianz auch wie folgt definiert werden:

$$S^2_P = \frac{\sum_{i=1}^{k} (n_i - 1) \cdot S^2_i}{\sum_{i=1}^{k} n_i - k},$$

P

Machine Translated by Google

wo $S2_{\bar{}}$ ist die Varianz der Stichprobe i:

$$S2_{\bar{}} = \frac{\sum_{j=1}^{ni} (xij - \bar{x}i.2)}{ni - 1} .$$

BEISPIELE

Betrachten Sie zwei Beispiele von Computern verschiedener Marken, für die wir die Zeit (in) notiert haben Stunden) vor dem ersten Problem:

Marke 1	Marke 2
2800	2800
2700	2600
2850	2400
2650	2700
2700	2600
2800	2500
2900	
3000	

Die Anzahl der Stichproben *k* beträgt 2 und die Dimension der Mustermarken entspricht, bzw:

$$n1 = 8 ,$$
$$n2 = 6 .$$

Die Berechnung der **Varianz** des ersten Probe gibt uns

$$S2_1 = \frac{\sum_{j=1}^{n1} (x1j - \bar{x}1.)^2}{n1 - 1} ,$$

Wo

$$\bar{x}1. = \frac{\sum_{j=1}^{n1} x1j}{n1} = \frac{22400}{8} = 2800$$

$$S2_1 = \frac{95000}{8 - 1} = 13571{,}428 .$$

Für das zweite Beispiel erhalten wir:

$$S2_2 = \frac{\sum_{j=1}^{n2} (x2j - \bar{x}2.)^2}{n2 - 1}$$

mit

$$\bar{x}2. = \frac{\sum_{j=1}^{n2} x2j}{n2} = \frac{15600}{6} = 2600$$

$$S2_2 = \frac{70000}{6 - 1} = 14000 .$$

Wenn wir davon ausgehen, dass die unbekannten Varianzen von zwei Populationenσ_1^2 undσ_2^2 sind also identisch können wir eine **Schätzung** von $\sigma^2 =$ berechnen $\sigma_1^2 = \sigma_2^2$ durch die gepoolte Varianz:

$$S2_P = \frac{\sum_{i=1}^{2} (ni - 1) \cdot S2_{\bar{}}}{\sum_{i=1}^{2} ni - 2}$$

$$= \frac{(8-1)\cdot 13571{,}428 + (6-1)\cdot 14000}{8+6-2}$$

$$= 13750 .$$

WEITERLESEN

Varianz

Bevölkerung

Eine Population ist definiert als eine Sammlung statistischer Einheiten gleicher Art, an deren quantifizierbaren Informationen wir interessiert sind

Die Bevölkerung bildet das Referenzuniversum während der Untersuchung eines bestimmten statistischen Problems.

BEISPIELE

Die Bürger eines Staates, eine Gruppe von Bäumen, Arbeiter in einer Fabrik und die Preise von Konsumgütern bilden alle diskrete Populationen.

Machine Translated by Google

WEITERLESEN

Probe

Häufigkeit

Siehe **Prävalenzrate.**

Prävalenzrate

Die **Prävalenz** einer Krankheit ist die Zahl der zu einem bestimmten Zeitpunkt betroffenen Personen in der statistischen Grundgesamtheit. Diese Vorstellung liegt nahe zum Begriff „Aktie“.

Die Prävalenzrate einer Krankheit ist der Anteil der betroffenen Personen an der Bevölkerung zu einem bestimmten Zeitpunkt.

GESCHICHTE

Farr, William, Pionier in der Verwendung von Statistiken in Epidemiologie und Schöpfer der Konzepte der Sterblichkeitsrate und Frage und Antwort, zeigte, dass die **Prävalenz** einer Krankheit ist äquivalent zum Produkt der **Inzidenz** und Dauer der Erkrankung.

MacMahon, B. und Pugh, TF (1970) veranschaulichten den Zusammenhang zwischen der Inzidenzrate und die Prävalenzrate unter Verwendung von Daten, die zwischen 1948 und 1952 zu akuten Erkrankungen gesammelt wurden Leukämie in der weißen Bevölkerung Brooklyn, New York. Die Inzidenzrate betrug 32,5 Fälle pro Million Einwohner pro Jahr, und die Prävalenzrate betrug 6,7 Fälle pro Million Einwohner. Die Dauer von akute Leukämie musste somit 0,21 Jahre betragen bzw 2,5 Monate:

$$\text{Dauer von Akute Leukämie} = \frac{\text{Prävalenz aß}}{\text{Inzidenzrate}} = \frac{6.7}{32.5}\ \text{Jahr} = 0{,}21\ \text{Jahr}.$$

MATHEMATISCHE ASPEKTE

Die Prävalenzrate ist wie folgt definiert:

$$\text{Prävalenzrate} = \frac{\text{Anzahl der Betroffenen}}{\text{Dimension von Bevölkerung}}.$$

Die Prävalenzrate und die **Inzidenzrate** einer Krankheit sind ungefähr damit verbunden einander durch die mittlere Krankheitsdauer (d. h. die mittlere Überlebensdauer bzw die mittlere Dauer der Kur). Diese Beziehung ist geschrieben:

$$\text{Prävalenzrate} \approx (\text{Inzidenzrate}) \cdot \text{mittlere Dauer der Krankheit}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Prävalenz unterscheidet sich vom **Risiko.** Ein Risiko besteht eine Größe, die einen Vorhersagewert hat, da sie betrifft einen festgelegten Beobachtungszeitraum und enthält daher zeitlich homogene Informationen. Im Gegensatz dazu ist die Prävalenz tangential zur Geschichte einer Krankheit in einer Bevölkerung: Sie kann keinen Hinweis darauf geben das Risiko, sofern es sich zu einem bestimmten Zeitpunkt aggregiert Moment, kürzlich diagnostizierte Fälle und andere in der fernen Vergangenheit. Beachten Sie, dass je größer das mittlere Überleben, wenn die Krankheit unheilbar ist, oder je länger die Genesung, wenn die Krankheit heilbar ist, desto größer ist die Prävalenz Erkrankung.

Beachten Sie abschließend, dass die Prävalenzrate häufig ist bezieht sich nicht auf eine genau definierte Krankheit, sondern auf eine Werteintervall eines biologischen Parameters. Beispielsweise kann die Prävalenzrate einer Hypercholesterinämie definiert werden als der Anteil der Personen mit einem Cholesterinspiegel von mehr als 6,5 mmol/l in einer bestimmten Bevölkerung zu einem bestimmten Zeitpunkt. Somit werden die Perzentile eines biologischen **Parameters** als Maß für den Anteil der Individuen angesehen, für die der Parameter einen Wert annimmt

P

Machine Translated by Google

innerhalb eines bestimmten Intervalls kann als Prävalenzrate interpretiert werden. Beispielsweise die Prävalenzrate relativ zu Werten kleiner als oder gleich dem **Median** (oder Perzentil 50) eines biologischen Parameters ist 50 %.

WEITERLESEN

Zurechenbares Risiko

Vermeidbares Risiko

Ursache und Wirkung in der Epidemiologie

Inzidenzrate

Quoten und Quotenverhältnis

Relatives Risiko

Risiko

VERWEISE

Cornfield, J.: Eine Methode zur Schätzung von Vergleichsraten anhand klinischer Daten. Anwendungen bei Lungen-, Brust- und Gebärmutterhalskrebs. J. Natl. Krebsinst. **11,** 1269–75 (1951)

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Little Brown, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Morabia, A.: L'Épidémiologie Clinique. Ausgaben „Que sais-je?“. Presses Univer Sitaires de France, Paris (1996)

Wahrscheinlichkeit

Wir können die Wahrscheinlichkeit eines **Ereignisses** entweder durch die Verwendung der relativen Häufigkeiten oder durch einen axiomatischen Ansatz definieren.

Beim ersten Ansatz gehen wir davon aus, dass ein **Zufallsexperiment** viele Male wiederholt wird

das gleicheconditions.ForeacheventAdefiniert

Im **Probenraum** definieren wir *nA* als die Häufigkeit, mit der das Ereignis A während der ersten *n* Wiederholungen des **Experiments auftrat.** In diesem Fall ist die Wahrscheinlichkeit des Ereignisses A, bezeichnet mit *P* (A), definiert durch:

$$P(A) = \lim_{n\ddot{y}\ddot{y}} \frac{n/A}{N},$$

Das bedeutet, dass *P* (A) als Grenzwert relativ zur Häufigkeit von Ereignis *A* definiert ist

trat im Verhältnis zur Gesamtzahl der Wiederholungen auf.

Im zweiten Ansatz akzeptieren wir für jedes Ereignis A, dass es eine Wahrscheinlichkeit von A, *P* (A) gibt, die die folgenden drei Axiome erfüllt:

1. 0 ÿ *P (A)* ÿ 1, 2. *P* () = 1, 3. Für jede Folge sich gegenseitig ausschließender Ereignisse A1, *A2,...* (das heißt von Ereignissen *Ai* ÿ *Aj* = ÿ, wenn *i* = j) :

$$P\left(\bigcup_{i=1}^{\ddot{y}} Ai\right) = \sum_{i=1}^{\ddot{y}} P(Ai).$$

GESCHICHTE Die ersten Risikospiele markieren den Beginn der Geschichte der Wahrscheinlichkeitstheorie, und wir können bestätigen, dass sie auf die Entstehung *des Homo sapiens zurückgehen.*

Der Ursprung des Wortes „Gefahr“ bietet weniger Gewissheit. Laut Kendall, Maurice George (1956) wurde es zur Zeit des dritten Kreuzzugs nach Europa gebracht und daraus abgeleitet

vom arabischen Wort „al zhar“, was „Würfel“ bedeutet.

Benzécri, JP (1982) gibt einen Hinweis auf Franc d'Orient, genannt Guillaume de Tyr, nach dem das Wort Hazard mit dem Namen einer Burg in Syrien zusammenhängt. Das

Machine Translated by Google

Die Burg war Gegenstand einer Belagerung, bei der die Angreifer ein Würfelspiel erfanden, das diesen Spielen ihren Namen gab.
In alten Zivilisationen finden sich Spuren von Glücksspielen. Zeichnungen und Gegenstände, die in ägyptischen Gräbern aus der Ersten Dynastie (3500 v. Chr.) gefunden wurden, zeigen, dass diese Spiele bereits zu dieser Zeit gespielt wurden.
Laut dem griechischen Historiker Herodot wurden Gefahrenspiele von Palamedes während der Belagerung Trojas erfunden.
Das Astragalspiel für Hunde war eines der ersten Gefahrenspiele. Der Astragalus ist ein kleiner Fußknochen, der symmetrisch zur vertikalen Achse ist.
Von den Griechen und später von den Römern sehr geschätzt, handelt es sich laut Kendall, MG (1956) möglicherweise um das ursprüngliche Craps-Spiel. Sehr alte Astragale wurden in Ägypten gefunden, aber ältere, aus Terrakotta gefertigte Exemplare wurden im Nordirak entdeckt.
Sie stammen aus dem dritten Jahrtausend v. Chr
Eine weitere, ebenfalls aus Terrakotta stammende Figur aus derselben Zeit wurde in Indien gefunden.
Es gibt eine Fülle von Beweisen dafür
Sie zeigen, wie in der Antike Würfelspiele gespielt wurden. Aber damals dachte niemand daran, die gleiche wahrscheinliche Eigenschaft zu ermitteln, um eine Seite zu gewinnen.
Für David, FN (1955) helfen zwei Hypothesen, diese Lücke zu erklären. Der ersten Hypothese zufolge waren die Würfel im Allgemeinen deformiert. Der zweiten Hypothese zufolge war das Würfeln religiöser Natur. Es war unterwegs, um den Göttern Fragen zu stellen: Das Ergebnis eines bestimmten Würfelwurfs enthüllte die Antwort auf Fragen, die den Göttern gestellt worden waren.

Unter den Glücksspielen haben auch Spielkarten einen sehr alten Ursprung. Wir wissen, dass sie in China, Indien, Arabien und Ägypten verwendet wurden.
Im Westen finden wir den ersten Nachweis von Karten in Venedig im Jahr 1377, in Nürnberg im Jahr 1380, und in Paris im Jahr 1397. Tarotkarten sind am häufigsten alte Spielkarten.
Bis zum 15. Jahrhundert kämpften Kirche und Könige laut Kendall, MG (1956) gegen die Praxis des Würfelspiels.

Als Beweis dafür können wir die Gesetze Ludwigs IX., Eduards III. und Heinrichs VIII. nennen. Kartenspiele wurden in die Liste der illegalen Aktivitäten aufgenommen, und wir finden einen Hinweis auf das 1397 in Paris erlassene Verbot von Kartenspielen. Trotz der Verbote gab es von den Römern bis zur Renaissance durchgehend Glücksspiele

soziale Klassen.
Es überrascht nicht, dass sie den Ursprung mathematischer Arbeiten zur Wahrscheinlichkeit bilden. Im Westen stammen die ersten Schriften zu diesem Thema aus Italien. Sie wurden von Cardano in seiner 1663 posthum veröffentlichten Abhandlung *Liber de Ludo Aleae* und von Galileo Galilei in seiner *Sopra le Scoperte dei Dadi verfasst.* In den Werken von Galileo finden wir das

Prinzip der Gleichwahrscheinlichkeit von Würfelspielen.
Im übrigen Europa interessierten diese Gefahrenprobleme auch Mathematiker.
Huygens aus den Niederlanden veröffentlichte 1657 ein Werk mit dem Titel *De Ratiociniis in Aleae Ludo ,* der laut Holgate, P. (1984) das Interesse an Wahrscheinlichkeiten auf andere Bereiche ausgeweitet hat; Parallel dazu würde die Spieltheorie in Bezug auf die Körperdynamik entwickelt. Die Arbeiten von Huygens haben die Werke zweier anderer bekannter Mathematiker erheblich beeinflusst: **Bernoulli, Jakob** und **de Moivre, Abraham.**

Aber erst in Frankreich nahm die Wahrscheinlichkeitstheorie mit Pascal, Blaise (1623–1662) und de Fermat, Pierre (1601–1665) wirklich Gestalt an. Laut Todhunter, I. (1949) wurde Pascal, B. von einem berühmten Spieler, Gombauld, Antoine, Ritter von Méré, gebeten, das folgende Problem zu lösen: In einem Spiel

P

Machine Translated by Google

Es gibt zwei Gegner. Der erste gewinnt *n* von *m* Spielen, während der Zweite gewinnt. *p* von *m* Spielen. Dies sind die Chancen für einen von ihnen, das Spiel zu gewinnen (vorausgesetzt, er gewinnt den Ersten, der *m* Spiele benötigt).
Pascal,B. kontaktierte deFermat, P., der die Lösung fand. Pascal selbst hat das entdeckt Wiederholungsformel, die das gleiche Ergebnis zeigt. Laut Benzécri, JP (1982) war der Osten zu einer Zeit, als der Westen Zeuge des Untergangs des Römischen Reiches und des Beginns der christlichen Ära war, mit intensiver wissenschaftlicher und künstlerischer Aktivität beschäftigt. Unter diesen Umständen entdeckten östliche Gelehrte wie Omar Khayam wahrscheinlich Wahrscheinlichkeitsregeln.

Es gibt einen Aspekt der Wahrscheinlichkeit, der für Mathematiker von besonderem Interesse ist: die Probleme der **kombinatorischen Analyse.**

MATHEMATISCHE ASPEKTE

Axiomatische Basis der Wahrscheinlichkeit

Wir betrachten ein **Zufallsexperiment** mit einem **Probenraum ,** der *n* Elemente enthält :

$$= \{x1, x2,...,xn\} .$$

Wir können jedem einfachen **Ereignis** *xi* eine Wahrscheinlichkeit *P (xi)* mit den folgenden Eigenschaften zuordnen: 1. Die Wahrscheinlichkeiten *P (xi)* sind nicht negativ:

$$P(xi) \geq 0 \text{ für } i = 1, 2,...,n.$$

2. Die Summe der Wahrscheinlichkeiten von *xi* für *i* von 1 nach *n* ist gleich 1:

$$\sum_{i=1}^{N} P(xi) = 1 .$$

3. Die Wahrscheinlichkeitsfunktion P ist eine Funktion, die jedem einfachen Ereignis eines Zufallsexperiments eine Zahl zwischen 0 und 1 zuweist:

$$P: \rightarrow [0, 1] .$$

Eigenschaften von Wahrscheinlichkeiten

1. Die Wahrscheinlichkeit eines **Stichprobenraums** ist die höchste Wahrscheinlichkeit, die zugeordnet we zu einer Veranstaltung:

$$P() = 1 .$$

2. Die Wahrscheinlichkeit eines unmöglichen Ereignisses ist gleich 0:

$$\text{Wenn } A = \emptyset , \quad \text{dann ist } P(A) = 0 .$$

3. Sei A¯ das **Komplement** von *A* in , Dann die Wahrscheinlichkeit von A¯ ist gleich 1 minus der Wahrscheinlichkeit von A:

$$P(\bar{A}) = 1 - P(A).$$

4. Seien *A* und *B* zwei inkompatible Ereignisse *(A* ∩ *B* = ∅). Die Wahrscheinlichkeit von *A* ∪ *B* entspricht der Summe der Wahrscheinlichkeiten von *A* und B:

$$P(A \cup B) = P(A) + P(B).$$

5. Seien A und B zwei Ereignisse. Die Wahrscheinlichkeit von *A* ∪ *B* ist gleich:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

WEITERLESEN

Bedingte Wahrscheinlichkeit
Kontinuierliche Wahrscheinlichkeitsverteilung
Diskrete Wahrscheinlichkeitsverteilung
Fall
Zufälliges Experiment
Probenraum

VERWEISE

Benzécri, JP: Geschichte und Vorgeschichte der Données-Analyse. Dunod, Paris (1982)

Machine Translated by Google

David, FN: Würfeln und Spielen (eine Anmerkung zu die Geschichte der Wahrscheinlichkeit). Biometrie **42,** 1–15 (1955)

Holgate, P.: Der Einfluss von Huygens' Werk in Dynamics über seinen Beitrag zur Wahrscheinlichkeit. Int. Stat. Rev. **52,** 137–140 (1984)

Kendall, MG: Studien zur Geschichte von Wahrscheinlichkeit und Statistik: II. Die Anfänge einer Wahrscheinlichkeitsrechnung. Biometrie ka **43,** 1–14 (1956)

Stigler, S.: Die Geschichte der Statistik, die Messung der Unsicherheit vorher 1900. Belknap, London (1986)

Todhunter, I.: Eine Geschichte der Mathematik Theorie der Wahrscheinlichkeit. Chelsea Publishing Unternehmen, New York (1949)

Wahrscheinlichkeitsverteilung

Synonym **für Wahrscheinlichkeitsfunktion,** wird aber verwendet viel häufiger als **„Wahrscheinlichkeit".** **„Funktion"** zur Beschreibung der Verteilungsgabe die Wahrscheinlichkeit des Auftretens eines Wertes einer Zufallsvariablen *X* beim Wert von x.

GESCHICHTE

Im 17. Jahrhundert begann die systematische Untersuchung von Problemen im Zusammenhang mit dem Zufall Phänomene begannen. Der bedeutende Physiker Galilei hatte bereits versucht, die **Fehler** zu untersuchen physikalischer Messungen unter Berücksichtigung dieser als zufällig und schätzen ihre **Wahrscheinlichkeit.** In dieser Zeit entstand die Theorie der Versicherungen erschien ebenfalls und basierte auf der Analyse der Gesetze, die zufällige Phänomene beherrschen wie Morbidität, Mortalität, Unfälle usw. Allerdings war es zunächst notwendig, einfachere Phänomene wie Glücksspiele zu untersuchen. Diese sind besonders einfach und übersichtlich gestaltet **Modelle** zufälliger Phänomene, die es ermöglichen einer, der die spezifischen Gesetze beobachtet und studiert die sie beherrschen; darüber hinaus die Möglichkeit von das gleiche **Experiment** viele Male wiederholen ermöglicht eine experimentelle Überprüfung dieser Gesetze.

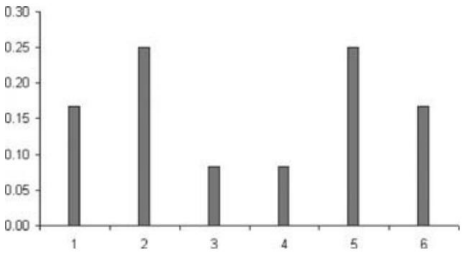
MATHEMATISCHE ASPEKTE

Die Funktion *P(b)* = P(X = b), wobei *b* variiert je nach den möglichen **Werten** des diskrete **Zufallsvariable** X, heißt die Wahrscheinlichkeitsfunktion von X. Da P(X = b)ist Immer positiv oder Null, die Wahrscheinlichkeitsfunktion ist ebenfalls positiv oder Null.
Die Wahrscheinlichkeitsfunktion wird auf einem dargestellt Achsensystem. Die verschiedenen **Werte** *b* von *X* sind als Abszissen aufgetragen, die BilderP(b) als Ordinate. Dargestellt wird die **Wahrscheinlichkeit** *P(b).* durch Rechtecke mit einer Breite gleich Eins und eine Höhe gleich der Wahrscheinlichkeit von b. Seit *X* muss mindestens einen der **Werte** b annehmen, d. h Die Summe von *P(b)* muss gleich 1 sein, d. h dass die Summe der Flächen der Rechtecke ist muss gleich 1 sein.

BEISPIELE

Stellen Sie sich ein **Zufallsexperiment** vor , bei dem ein fester Würfel geworfen wird. Betrachten Sie die **Zufallsvariable** *X,* die der Anzahl der erhaltenen Punkte entspricht. Die Wahrscheinlichkeitsfunktion P(X = b) ist durch die **Wahrscheinlichkeiten** gegeben jedem **Wert** von X zugeordnet :

B	1	2	3	4	5	6
P(b)	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{6}$



P

Machine Translated by Google

WEITERLESEN

Kontinuierliche Wahrscheinlichkeitsverteilung
Dichtefunktion
Diskrete Verteilungsfunktion
Diskrete Wahrscheinlichkeitsverteilung
Verteilungsfunktion
Wahrscheinlichkeit
Wahrscheinlichkeitsfunktion
Zufällige Variable

VERWEISE

Johnson, NL, Kotz, S.: Verteilungen in der Statistik: Diskrete Verteilungen. Wiley, New York (1969)

Johnson, NL, Kotz, S.: Verteilungen in der Statistik: Kontinuierliche univariate Verteilung Butionen, Bde. 1 und 2. Wiley, New York (1970)

Johnson, NL, Kotz, S., Balakrishnan, N.: Discrete Multivariate Distributions, John Wiley (1997)

Rothschild, V., Logothetis, N.: Wahrscheinlichkeitsverteilungen. Wiley, New York (1986)

Wahrscheinlichkeitsfunktion

Die Wahrscheinlichkeitsfunktion einer diskreten **Zufallsvariablen** ist eine assoziierende Funktion Jeder **Wert** dieser Zufallsvariablen wird seiner **Wahrscheinlichkeit zugeordnet.**

Siehe **Wahrscheinlichkeitsverteilung.**

GESCHICHTE Siehe **Wahrscheinlichkeit, Wahrscheinlichkeitsverteilung.**

p-Wert

Der p-Wert ist definiert als die unter der **Nullhypothese** berechnete **Wahrscheinlichkeit** von hav Ist das Ergebnis so extrem wie der beobachtete **Wert** in der **Stichprobe** oder ist die Wahrscheinlichkeit, ein Ergebnis zu erhalten, mindestens so extrem wie ein gegebener Datenpunkt, vorausgesetzt, der Datenpunkt war nur ein Ergebnis des Zufalls?

GESCHICHTE

Der p-Wert wurde 1975 von Gibbens und Pratt eingeführt.

MATHEMATISCHE ASPEKTE

Lassen Sie uns den Fall eines **Hypothesentests** veranschaulichen , der mit dem **Mittelwertschätzer** durchgeführt wird . das gleiche Prinzip gilt für jeden anderen Schätzer; nur die Notationen sind unterschiedlich. Angenommen, wir möchten die folgenden Hypothesen testen:

$$H0 : ÿ = ÿ0$$

$$H1 : ÿ>ÿ0 ,$$

wobei ÿ den Mittelwert einer normalverteilten **Grundgesamtheit** mit einer bekannten **Standardabweichung** ÿ darstellt. Eine **Stichprobe** der Dimension *n* ergibt einen beobachteten Mittelwert x¯. Betrachten Sie den Fall, in dem x¯ ÿ0 überschreitet , und berechnen Sie die **Wahrscheinlichkeit ,** unter der **Nullhypothese** ÿ = ÿ0 eine **Schätzung ÿ**ˆ zu erhalten , die größer oder gleich ¯x ist . Der Wert *p* entspricht dieser Wahrscheinlichkeit. Daher:

$$p = P(ÿ^ ÿ ¯x|ÿ = ÿ0) .$$

Der Standardzufallswert **Z** *gegeben* durch ÿˆ ÿ ÿ0

Z = folgt der Normalverteilung von ÿÿ n

Mittelwert 0 mit **Standardabweichung** 1. Einführung

Wenn wir diese **Variable** in den Ausdruck von p umwandeln, finden wir:

$$= PZ ÿ \quad \frac{x¯ ÿ ÿ0\ p}{\overline{ÿÿ\ n}} \quad ,$$

Machine Translated by Google

was in dieser Form **normal** gelesen werden kann **Tisch.**

Für ein **Signifikanzniveau** ÿ ermöglicht der Vergleich von *p* mit ÿ eine Entscheidung über eine eventuelle Ablehnung der **Nullhypothese.** Wenn: • *p* ÿ ÿ: wir lehnen die Nullhypothese H0 zugunsten der **Alternativhypothese** H1 ab; • *p* > ÿ: Wir lehnen die Nullhypothese nicht ab H0.

Wir können auch die Obergrenze des **Akzeptanzbereichs von H0 berechnen,** wenn wir den Wert kennen Cÿ wie z

$$P(\bar{y} \text{ ÿ } C\ddot{y}|\ddot{y} = \ddot{y}0) = \ddot{y} .$$

Abhängig von der standardmäßigen normalen Zufallsvariablen zÿ wird der Wert Cÿ ausgedrückt durch:

$$C\ddot{y} = \ddot{y}0 + z\ddot{y} \cdot \frac{\ddot{y}}{\sqrt{n}} .$$

BEREICHE UND EINSCHRÄNKUNGEN

Wir verwenden häufig den *p*-Wert, um das Ergebnis eines **Hypothesentests zu verstehen.** Es wird für einen **einseitigen Test** verwendet , wenn die Hypothesen die folgende Form haben:

$$H0 : \ddot{y} = \ddot{y}0$$
$$H1 : \ddot{y} > \ddot{y}0 ,$$

oder

$$H0 : \ddot{y} = \ddot{y}0$$
$$H1 : \ddot{y} < \ddot{y}0 ,$$

oder

$$H0 : \ddot{y} \ddot{y} \ddot{y}0$$
$$H1 : \ddot{y} > \ddot{y}0 ,$$

oder

$$H0 : \ddot{y} \ddot{y} \ddot{y}0$$
$$H1 : \ddot{y} < \ddot{y}0 .$$

Der Wert *p* kann als das kleinste **Signifikanzniveau** interpretiert werden, für das die **Nullhypothese** nicht abgelehnt werden kann.

BEISPIELE

Angenommen, wir möchten einen einseitigen **Hypothesentest** durchführen

$$H0 : \ddot{y} = 30$$
gegen $H1 : \ddot{y} > 30$

für eine normalverteilte **Grundgesamtheit** mit einer **Standardabweichung** ÿ = 8. Eine **Stichprobe** mit der **Größe** *n* = 25 ergibt einen beobachteten **Mittelwert** von 34. Da $\bar{x}$ (= 34) ÿ = 30 deutlich überschreitet, sollten wir versuchen, die **Alternativhypothese** zu akzeptieren , die ÿ > 30 lautet.

Berechnen Sie die **Wahrscheinlichkeit ,** unter der **Nullhypothese** ÿ = 30 einen beobachteten Mittelwert von 34 zu erhalten. Dies ist der *p*- Wert:

$$p = P(\bar{X} \text{ ÿ } 34|\ddot{y} = 30)$$
$$= P\left(\frac{\bar{X} \text{ ÿ } 30 \text{ ÿ}}{\frac{8}{5}} \quad \frac{34 \text{ ÿ } 30}{\frac{8}{5}}\right)$$
$$= P(Z \text{ ÿ } 2{,}5) ,$$

wobei $Z = \frac{\bar{X} \text{ ÿ } \ddot{y}}{\frac{\ddot{y}}{\sqrt{n}}}$ normalverteilt ist mit Mittelwert 0 und Standardabweichung 1; *p* = 0,0062 gemäß der **Normaltabelle.**

Für ein **Signifikanzniveau** ÿ = 0,01 lehnen wir die Nullhypothese für die Alternativhypothese ÿ > 30 ab. Wir können die Obergrenze des **Akzeptanzbereichs** von H0, Cÿ wie z

$$P(\bar{X} \text{ ÿ } C\ddot{y}|\ddot{y} = 30) = 0{,}01$$

8 und finden Sie Cÿ = 30 + 1,64 $\frac{8}{\sqrt{25}}$ = 32,624. 5

P

Machine Translated by Google

Dies bestätigt, dass wir die Nullhypothese ÿ = 30 wirklich ablehnen müssen, wenn der beobachtete Mittelwert 34 beträgt.

WEITERLESEN

- **Akzeptanzregion**
- **Hypothesentest**
- **Normaler Tisch**
- **Signifikanzniveau**

REFERENZEN

Gibbens, RJ, Pratt, JW: P-Werte-Interpretation und -Methodik. Bin. Stat. 29(1), 20–25 (1975)

Machine Translated by Google

QQ-Plot (Quantil-zu-Quantil-Diagramm)

Das QQ-Diagramm oder Quantil-zu-Quantil-Diagramm ist ein Diagramm, das die Konformität zwischen der empirischen Verteilung und der gegebenen theoretischen Verteilung testet.

Eine der Methoden zur Überprüfung der Fehlernormalität eines Regressionsmodells besteht darin, ein QQ-Diagramm der Residuen zu erstellen. Wenn die Punkte auf der Linie *x* = y ausgerichtet sind, sind die Daten normalverteilt.

GESCHICHTE

Die als QQ-Plot bekannte Methode der **grafischen Darstellung** erschien in den frühen 1960er Jahren. Seitdem ist es zu einem notwendigen Werkzeug bei der Analyse von Daten und/oder Residuen geworden, da es sehr viele Informationen bietet und leicht zu interpretieren ist.

MATHEMATISCHE ASPEKTE

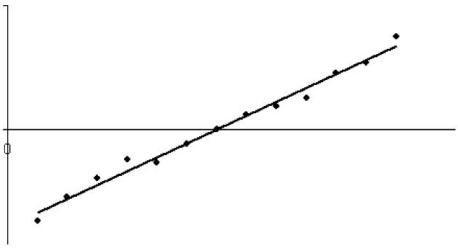
Um ein QQ-Diagramm zu erstellen, folgen wir zwei Schritten.

1. Ordnen Sie die Daten x1, *x2,...,xn* in aufsteigender Reihenfolge an:

   x[1] ÿ x[2] ÿ ... ÿ *x[n]* .

2. Ordnen Sie jedem Datenpunkt *x[i]* das i/(n+ **1)-Quantil** *qi* der Standardnormalverteilung zu. Tragen Sie in einem Diagramm als Ordinate die geordneten Daten *xi* und als Abszisse die Quantile *qi ein.*

Wenn die Variablen die gleiche Verteilung haben, ist die grafische Darstellung zwischen den Quantilen der ersten Variablen relativ zu den Quantilen der zweiten Verteilung eine Linie mit der Steigung 1. Wenn also die Daten xi normalverteilt sind, müssen die Punkte im Diagramm *dies* tun nahezu auf der Gleichungsgeraden *xi = qi ausgerichtet sein.*



BEREICHE UND EINSCHRÄNKUNGEN

Das QQ-Diagramm wird verwendet, um zu überprüfen, ob Daten einer bestimmten Verteilung folgen oder ob zwei gegebene Datensätze dieselbe Verteilung haben. Wenn

Die Verteilungen sind gleich, der Graph ist eine Linie. Je weiter das erhaltene Ergebnis entfernt ist

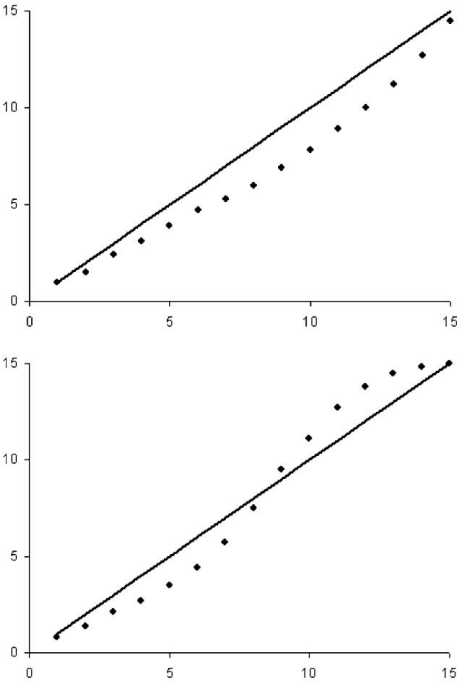
Je größer die 45ÿ- Diagonale, desto weiter ist die empirische Verteilung von der theoretischen. Der

Extrempunkte weisen eine größere Variabilität auf als diejenigen im Zentrum der Verteilung. Daher

Ein U-förmiges Diagramm bedeutet, dass eine Verteilung relativ zu einer anderen verzerrt ist. Ein S-förmiges Diagramm zeigt an, dass die Verteilungen einen größeren Einfluss von Extremwerten auf eine andere Verteilung darstellen (Long Tail).

Machine Translated by Google

Eine häufige Verwendung der QQ-Diagramme ist wann man möchte das Verhalten von Residuen kennen während einer einfachen oder **multiplen linearen Regression ,** genauer gesagt, um zu wissen, ob ihre Verteilung normal ist.



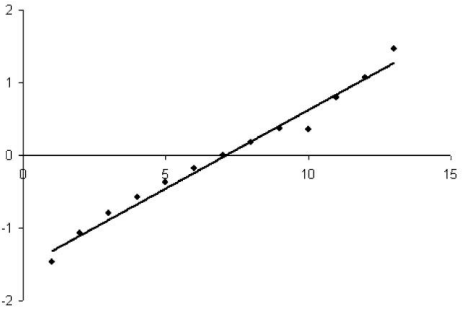
BEISPIELE

Die folgende Tabelle enthält die Residuen (in aufsteigender Ordnung) einer linearen Regression mit 13 Beobachtungen. Wir assoziieren sie mit dem Quantilesqi wie oben beschrieben.Das erhaltene

Das QQ-Diagramm wird durch die folgende Abbildung dargestellt. Wir sehen das mit Ausnahme von zwei Punkten, die es sich wahrscheinlich Ausreißer, die Punkte sind gut ausgerichtet.

Tabelle: Residuen und Quantile

ei	Qi
1 ÿ1,62 ÿ1,47	
2 ÿ0,66 ÿ1,07	
3 ÿ0,62 ÿ0,79	
4 ÿ0,37 ÿ0,57	
5 ÿ0,17 ÿ0,37	
6 ÿ0,12 ÿ0,18	
7	0,04	0,00
8	0,18	0,18
9	0,34	0,37
10	0,42	0,57
11	0,64	0,79
12	1.33	1.07
13	2,54	1,47



WEITERLESEN

- **Analyse von Residuen**
- **Datenanalyse**
- **Explorative Datenanalyse**
- **Multiple lineare Regression**
- **Wahrscheinlichkeitsverteilung**
- **Quantil**
- **Restwert**
- **Einfache lineare Regression**

VERWEISE

Chambers, JM, Cleveland, WS, Klein er, B., Tukey, PA: Grafische Methoden für Datenanalyse. Wadsworth, Belmont, Kalifornien (1983)

Hoaglin, DC, Mosteller, F., Tukey, JW (Hrsg.): Understanding Robust and Explorative Datenanalyse. Wiley, Neu York (1983)

Machine Translated by Google

Wilk, MB, Ganadesikan, R.: Wahrscheinlichkeit
Darstellungsmethoden für die Datenanalyse.
Biometrie **55,** 1–17 (1968)

Qualitative kategoriale Variable

Eine qualitative kategoriale **Variable** ist eine Variable mit Modalitäten in Form von cate

gories, wie zum Beispiel „Mann“ und
„Frau“ der Variable „Geschlecht“; oder die Kategorien
„Rot“, „Orange“, „Grün“, „Blau“, „in Digo“ und „Violett“
der Variablen „Farbe“.
Die Modalitäten der qualitativen Kategorisierung
Variable kann auf einem Nominalwert dargestellt werden
Skala oder auf einer Ordinalskala.

Ein Beispiel für eine qualitative kategoriale Variable
mit einer Ordinalskala ist die qualitative Berufsvariable
mit Kategorien
„qualifiziert“, „halbqualifiziert“ und „nicht qualifiziert“.

WEITERLESEN

Kategorie
Variable

Quantil

Quantile messen die Position (oder den **Mittelpunkt) . Tendenz)** und versuchen Sie nicht unbedingt, das Zentrum einer Beobachtungsverteilung zu bestimmen
sondern um eine bestimmte Position zu beschreiben.
Dieser Begriff ist eine Erweiterung des Konzepts
des **Medians** (der eine Verteilung teilt).
der Beobachtung in zwei Teile). Die am häufigsten
verwendeten Quantile sind:

- Quartile, die eine Sammlung von trennen
  Beobachtungen in vier Teile,
- Dezile, die eine Sammlung von trennen
  Beobachtungen in zehn Teile,
- Zentile, die eine Sammlung von trennen
  Beobachtungen in hundert Teile.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Berechnung von Quantilen ist sinnvoll
nur für eine **quantitative Variable ,** die das kann
Nehmen Sie Werte in einem bestimmten **Intervall an.**

Das Konzept des Quantils bezeichnet die Aufteilung
einer Beobachtungsverteilung in
eine beliebige Anzahl von Teilen. Notiere dass der
Je größer die Anzahl der Beobachtungen, desto mehr
Anspruchsvoller kann die Trennung der Verteilung
sein.

Quartile können im Allgemeinen für jede Verteilung
verwendet werden. Die Berechnung von Dezilen und
damit auch von Zentilen erfordert eine relativ große
Anzahl von Beobachtungen, um eine gültige
Interpretation zu erhalten.

WEITERLESEN

Dezil
Standortmaß
Median
Perzentil
Quartil

Quantitatives Diagramm

Quantitative Diagramme werden zur Darstellung und Darstellung verwendet
Fassen Sie die kommenden numerischen Informationen zusammen
aus dem Studium einer **kategorialen Quantität Variable.**

Die am häufigsten verwendeten Arten quantitativer
Diagramme sind:

- **Balkendiagramm**
- **Piktogramm**
- **Kreisdiagramm**

GESCHICHTE

Beniger, JR und Robyn, JL (1978) gehen zurück
die historische Entwicklung der quantitativen
Karten seit dem 17. Jahrhundert. Schmid, CF
und Schmid, SE (1979) diskutieren die zahlreichen
Formen quantitativer Graphen.

Machine Translated by Google

WEITERLESEN

Balkendiagramm

Grafische Darstellung

Piktogramm

Kuchendiagramm

VERWEISE

Beniger, JR, Robyn, DL: Quantitative Grafiken in der Statistik: eine kurze Geschichte. Bin. Stat. **32,** 1–11 (1978)

Schmid, CF, Schmid, SE: Handbook of Graphic Presentation, 2. Aufl. Wiley, New York (1979)

Quantitative Variable

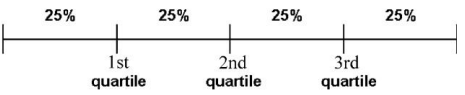
Eine quantitative Variable ist eine **Variable** mit numerischen Modalitäten. Beispielsweise sind Gewicht, Größe, Alter, Geschwindigkeit und Zeit quantitative Variablen. Wir unterscheiden diskrete Variablen (z. B. Anzahl der Kinder pro Familie) von kontinuierlichen Variablen (z. B. Länge eines Sprungs).

WEITERLESEN

Variable

Quartil

Quartile sind Ortsmaße einer Beobachtungsverteilung. Quartile unterteilen eine Verteilung in vier Teile. Somit gibt es für eine gegebene Verteilung drei Quartile.
Zwischen jedem Quartil finden wir 25 % der Gesamtbeobachtungen:



Beachten Sie, dass das zweite Quartil dem Median entspricht .

MATHEMATISCHE ASPEKTE

Die Berechnung des Quartils erfolgt ähnlich wie die des **Medians.** Wenn wir alle Beobachter haben

Die Berechnung der Quartile erfolgt wie folgt: 1. Die *n* Beobachtungen müssen in Form einer **Häufigkeitsverteilung angeordnet werden.**

2. Quartile entsprechen Beobachtungen, deren relative kumulierte **Häufigkeit** 25 %, 50 % und 75 % übersteigt. Einige Autoren schlagen die folgende Formel vor, die es ermöglicht, den **Wert** verschiedener Quartile präzise zu bestimmen : Berechnung des j-ten Quartils: Sei *i* der ganzzahlige Teil von *j·(n+1)* der Bruchteil von *j·(n+1)* $\frac{}{4}$ und *k* $\frac{}{4}$.

Seien *xi* und xi+1 die Werte der Beobachtungen jeweils an der i-ten und *(i* + 1)-ten Position (wenn die Beobachtungen in aufsteigender Reihenfolge angeordnet sind).
Das j-te Quartil ist

$$Qj = xi + k \cdot (xi+1 \text{ ÿ } xi) .$$

Wenn wir Beobachtungen in Klassen gruppieren, werden Quartile wie folgt bestimmt: 1. Bestimmen Sie die Klasse, in der sich das Quartil befindet gefunden:

- *Erstes Quartil:* Klasse, für die die relative kumulative Häufigkeit 25 % übersteigt.
- *Zweites Quartil:* Klasse, für die die relative kumulative Häufigkeit 50 % übersteigt.
- *Drittes Quartil:* Klasse, für die die relative kumulative Häufigkeit 75 % übersteigt.

2. Berechnen Sie den Wert des Quartils in Abhängigkeit von der Annahme, dass die Beobachtungen in jeder Klasse gleichmäßig verteilt sind. das j-te Quartil *Qj* ist:

$$Qj = Lj + \frac{n \cdot \frac{j}{4} \text{ ÿ finf } fj}{} \cdot cj ,$$

Machine Translated by Google

wobei *Lj* die untere Grenze der Klasse von ist
Quartil *Qj, n* ist die Gesamtzahl der Beobachtungen,
finf ist die Summe der Häufigkeiten
niedriger als die Klasse des Quartils ist *fj*
Häufigkeit der Quartilklasse *Qj* und
cj ist die Größe des **Intervalls** der Klasse von
Quartil *Qj*.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Berechnung von Quartilen ist sinnvoll
nur für eine **quantitative Variable ,** die das kann
Nehmen Sie Werte in einem bestimmten **Intervall an.**

Das Quartil ähnelt dem **Median.** Es ist
basiert ebenfalls auf dem Beobachtungsrang, nicht auf Werten.
Ein **Ausreißer** wird nur einen geringen Einfluss haben
auf den Quartilwerten.

BEISPIELE

Nehmen wir zunächst ein Beispiel, wo wir haben
zehn Beobachtungen *(n* = 10):

1244555679

Trotz der Tatsache, dass Quartile für eine kleine
Anzahl von Beobachtungen nicht berechnet werden können
(Wir sollten bei der Interpretation sehr vorsichtig sein),
wir werden diesen Fall untersuchen, um dies zu tun
die Prinzipien hinter der Berechnung verstehen.

Das erste Quartil Q1 liegt an der Position = 2,75. Q1
$\frac{n+1}{4}$ beträgt drei Viertel der **Entfernung** zwischen
der zweiten und dritten Observa
tionen (die wir x2 und x3 nennen werden) . Wir können
Berechnen Sie Q1 wie folgt:

$$Q1 = x2 + 0{,}75 \cdot (x3 \ddot{y} x2)$$
$$= 2 + 0{,}75 \cdot (4 \ddot{y} 2)$$
$$= 3{,}5 .$$

Das zweite Quartil Q2 (das den **Median darstellt)** ist
gefunden an der Position $2 \cdot \frac{n+1}{4}$ (oder $\frac{n+1}{2}$), welche
entspricht für unser Beispiel 5,5.

$$Q2 = x5 + 0{,}5 \cdot (x6 \ddot{y} x5)$$
$$= 5 + 0{,}5 \cdot (5 \ddot{y} 5)$$
$$= 5 .$$

An der Stelle befindet sich das dritte Quartil Q3
$3 \cdot \frac{n+1}{4} = 8{,}25$. Q3 entspricht:

$$Q3 = x8 + 0{,}25 \cdot (x9 \ddot{y} x8)$$
$$= 6 + 0{,}25 \cdot (7 \ddot{y} 6)$$
$$= 6{,}25 .$$

Die Werte 3,5, 5 und 6,25 trennen die
Beobachtungen in vier gleiche Teile.
Betrachten wir für das zweite Beispiel das
Die folgende **Häufigkeitstabelle** stellt die dar
Anzahl der Kinder pro Familie in 200 Familien:

Wert (Nummer von Kinder)	Relativ (Nummer von Familien)	Relativ häufig cy	Kumuliert Frequenz
0	6	0,03	0,03
1	38	0,19	0,22
2	50	0,25	0,47
3	54	0,27	0,74
4	42	0,21	0,95
5	8	0,04	0,99
6	2	0,01	1,00
Gesamt	200	1,00	

Das erste Quartil entspricht diesen Beobachtungen
die eine relative kumulative **Häufigkeit** haben
über 25 %, was zwei entspricht
Kinder (weil die relative kumulative
Die Frequenz für zwei Kinder beginnt bei 22
auf 47 %, davon 25 %). Der Zweite
Quartil entspricht drei Kindern, weil das
Die relative kumulierte Häufigkeit für drei Kinder liegt
zwischen 47 und 74 %, einschließlich

Machine Translated by Google

50 %. Das dritte Quartil entspricht vier Kindern weil die relative kumulierte Häufigkeit für vier Kinder geht von 74 auf 95 %, was umfasst 75 %.

Die Quartile 2, 3 und 4 unterteilen die 200 Familien in Viertel. Wir können 50 der 200 gruppieren Familien ins erste Quartal mit null, eins, oder zwei Kinder, 50 im zweiten Quartal mit zwei oder drei Kindern, 50 ins dritte Viertel mit drei oder vier Kindern und 50 mit vier, fünf oder sechs ins vierte Viertel Kinder.

Betrachten Sie nun ein Beispiel zur Berechnung von Quartilen aus der **Häufigkeit Verteilung** einer kontinuierlichen **Variablenwo**

Die Beobachtungen werden in Klassen gruppiert. Die folgende Häufigkeitstabelle stellt die dar Gewinn (in Tausend Franken) von 100 Filialen:

Profitieren (Tausend von Franken)	Cumu verspätet frei quenz	Cumu verspätet frei quenz	Relativ Frequenz
100–200	10	10	0,1
200–300	20	30	0,3
300–400	40	70	0,7
400–500	30	100	1,0
Gesamt	100		

Die Klasse, die das erste Quartil enthält, ist die Klasse 200–300 (diejenige mit einer relativen kumulativen Häufigkeit von 25 %).

Unter Berücksichtigung der Tatsache, dass die Beobachtungen in jeder Klasse gleichmäßig verteilt sind, erhalten wir:

für das erste Quartil die folgenden **Werte:**

1. Quartil

$$= 200 + \left(\frac{100 \cdot \frac{1}{4} \; \ddot{y} \; 10}{20} \right) \cdot 100$$

$$= 275 .$$

Die Klasse, die das zweite Quartil enthält, ist die Klasse 300–400. Der **Wert** der Sekunde Quartil ist gleich:

2. Quartil

$$= 300 + \left(\frac{100 \cdot \frac{2}{4} \; \ddot{y} \; 30}{40} \right) \cdot 100$$

$$= 350 .$$

Die Klasse, die das dritte Quartil enthält, ist die Klasse 400–500. Der Wert des dritten Quartils entspricht:

3. Quartil

$$= 400 + \left(\frac{100 \cdot \frac{3}{4} \; \ddot{y} \; 70}{30} \right) \cdot 100$$

$$= 416,66 .$$

Wir können daraus schließen, dass 25 der 100 Geschäfte Gewinne zwischen 100.000 und 275.000 haben Franken, 25 haben einen Gewinn zwischen 275000 und 350000 Franken, 25 haben Gewinne zwischen 350000 und 416066 Franken und 25 haben Profite zwischen 416066 und 500000 Franken.

WEITERLESEN

- **Dezil**
- **Standortmaß**
- **Median**
- **Perzentil**
- **Quantil**

Quetelet, Adolphe

Quetelet, Adolphe wurde 1796 in geboren Gent, Belgien. Im Jahr 1823 wurde der Minister für Die öffentliche Bildung bat ihn, die Aufsicht zu übernehmen der Bau einer Sternwarte für die

Machine Translated by Google

Stadt Brüssel. Zu diesem Zweck wurde er nach Paris geschickt, um Astronomie zu studieren und Pläne für den Bau der Sternwarte auszuarbeiten. In Frankreich, Bouvard, wurde Alexis wegen seiner Ausbildung angeklagt. Quetelet traf Adolphe unter anderem **Poisson, Siméon-Denis** und **de Laplace, Pierre Simon,** die ihn von der Bedeutung probabilistischer Berechnungen überzeugten. In den nächsten vier Jahren unternahm er weite Reisen, um verschiedene Observatorien zu besuchen und Instrumente zu sammeln. Er schrieb populäre Werke über die Wissenschaft: *Astronomie élémentaire, Astronomie populaire et Positions de physique* und *Instructions populaires sur le calcul des probabilités* im Jahr 1828. Darüber hinaus gründete er zusammen mit Garnier (seinem Doktorvater) eine Zeitschrift: Correspondance mathématique et physique. 1828 wurde er zum Astronomen der Brüsseler Sternwarte ernannt und stieg schnell zum Direktor der Sternwarte auf, eine Position, die er für die nächsten 25 Jahre innehatte.
Im Jahr 1836 wurde er Lehrer der Fürsten Ernst und Albert von Sachsen-Coburg und Gotha, was ihm ermöglichte, seinem Interesse an der Berechnung von Wahrscheinlichkeiten nachzugehen. Seine Lektionen für die Jungen wurden 1846 unter dem Titel *Lettres à SAR le Herzog von Sachsen-Cobourg und Gotha;* veröffentlicht Sie befassten sich mit der Wahrscheinlichkeitstheorie in ihrer Anwendung auf Ethik- und Politikwissenschaften. Quetelet starb 1874.

Hauptwerke von Quetelet, Adolphe:

1826 Astronomie élémentaire. Malther, Paris.

1828 Körperpositionen. Brüssel.

1828 Beliebte Anleitung zur Wahrscheinlichkeitsrechnung.

1846 Briefe an SAR an den Herzog von Sachsen-Cobourg und Gotha über die Theorie des probabilités, appliquée aux sci ences morales et politiques, Brüssel.

Quotenstrichprobenerhebung

Bei der Quotenstichprobe handelt es sich um eine nicht zufällige **Stichprobenmethode .** Eine **Stichprobe** wird so ausgewählt, dass sie ein möglichst bevölkerungsnahes Bild wiedergibt . Die Quotenmethode basiert auf der bekannten Verteilung einer **Bevölkerung** nach bestimmten Merkmalen (Geschlecht, Alter, sozioökonomische Klasse usw.).

GESCHICHTE

Siehe **Probenahme.**

MATHEMATISCHE ASPEKTE

Stellen Sie sich eine **Population** vor , die aus *N* Individuen besteht. Die Verteilung der Bevölkerung hat bekanntermaßen die Charakteristika Teristik *A, B, . . . , Z* mit den jeweiligen Modalitäten A1, *A2,...,Aa;* B1, *B2,...,Bb; ...;* Z1, *Z2,...,Zz,* die gewünschte **Stichprobe** erhält man durch Multiplikation der Anzahl der verschiedenen Modalitäten aller Kontrollmerkmale durch die Sondierungsrate **q ,** also den **Prozentsatz** der Bevölkerung, den man sondieren möchte.

Die **Stichprobe** hat die Größe $n = N \cdot q$ und die folgende Verteilung:

$$q \cdot XA1 \ , \ q \cdot XA2 \ ,...,q \cdot XAa \ , \ q \cdot XB1 \ , \ q \cdot XB2 \ , \ ... \ ,q \cdot XBb \ , \ ... \ ,q \cdot XZz \ ,$$

wobei *X...* die Anzahl der verschiedenen Modalitäten für alle Kontrollmerkmale darstellt.

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Ein Kontrollmerkmal, das als Grundlage für verwendet wird

Die Verteilung der **Probe** muss dem gehorchen

Folgende Regeln gelten: (1) Es muss in engem

Zusammenhang mit den untersuchten Variablen stehen. (2) Es

muss eine bekannte Verteilung für das Ganze haben

Bevölkerung.

Häufig verwendete Kontrollmerkmale sind Geschlecht,

Alter, sozioprofessionelle Schicht, Region usw.

Die Vorteile der Quotenstichprobe **sind** die

Folgendes:

- Im Gegensatz zur Zufallsstichprobe erfordert die Quotenstichprobe **nicht** das Vorhandensein von eine **fundierte** Grundlage.
- Die Kosten sind viel geringer als bei Zufall Probenahme.

Die Hauptnachteile sind:

- Die Quotenmethode verfügt nicht über eine ausreichende theoretische Grundlage. Es basiert auf der folgenden **Hypothese:** eine richtige Verteilung der Kontrollmerkmale gewährleistet die Repräsentativität der Verteilung von die beobachteten Merkmale.
- Die Quotenmethode kann das nicht berechnen Genauigkeit der erhaltenen **Schätzungen** aus der **Probe.** Denn die Ermittler wählen die zu befragenden Personen aus Es ist unmöglich zu wissen, mit welcher **Wahrscheinlichkeit** ein bestimmtes Individuum der **Bevölkerung** Teil der Stichprobe sein könnte. Es ist daher unmöglich, die Berechnung anzuwenden Wahrscheinlichkeiten, die im Fall des Zufalls Durch Stichproben kann jeder Schätzung ein Maß für den möglichen **Fehler** zugeordnet werden begangen worden.

BEISPIELE

Angenommen, wir wollen die Marktmöglichkeiten für ein Konsumprodukt in der Stadt X untersuchen. Die gewählten Kontrollvariablen sind Geschlecht, Alter, und sozioprofessionelle Klasse. **Die Bevölkerung** Verteilung (in Tausend Personen) ist die Folgendes:

Geschlecht	Alter	Sozioprofessionelle K	
Männlich 162	0–20 70	Häuptling	19.25
Weiblich 188	21–60 192,5	Unabhängig Dent Prof.	14.7
	ÿ 61 87,5	Führungskräfte 79,1 Arbeiter 60,9	
		Unem eingesetzt	176.05
Insgesamt 350	Insgesamt 350	Insgesamt	350

Wir wählen eine Sounding-Rate, die dem $\frac{1}{300}$. Der Sample entspricht, und haben daher die folgende Größe und Vertrieb:

$$n = N \cdot q = 350000 \cdot = 1167 . \frac{1}{300}$$

Geschlecht	Alter	Sozioprofessionelle K	
Männlich	540 0–20	233 Chef	64
Weiblich 627	21–60 642	Unabhängig Dent Prof.	49
	ÿ 61	292 Manager 264 Arbeiter 203	
		Unem eingesetzt	587
Gesamt 1167	Gesamt 1167	Gesamt	1167

Die Auswahl der Personen erfolgt dann durch **einfache Zufallsstichprobe** in jeder Kategorie.

WEITERLESEN

- Einschätzung
- Einfache Zufallsstichprobe
- Probenahme

VERWEISE

Sukhatme, PV: Stichprobentheorie von Umfragen mit Anwendungen. Iowa State University Press, Ames, IA (1954)

Machine Translated by Google

Zufälliges Experiment

Ein **Experiment ,** bei dem das Ergebnis nicht im Voraus vorhersehbar ist, wird als Zufallsexperiment bezeichnet.

Ein Zufallsexperiment kann wie folgt charakterisiert werden:

1. Es ist möglich, die Menge aller möglichen Ergebnisse (den sogenannten **Stichprobenraum** des Zufallsexperiments) zu beschreiben.
2. Es ist nicht möglich, das Ergebnis mit Sicherheit vorherzusagen.
3. Es ist möglich, jedes mögliche Ergebnis mit einer **Eintrittswahrscheinlichkeit** zu verknüpfen .

BEREICHE UND GRENZEN

Die Zufallsphänomene, für die **Wahrscheinlichkeiten** gelten, können in vielen verschiedenen Bereichen angetroffen werden, beispielsweise in den Wirtschaftswissenschaften, den Sozialwissenschaften, der Physik, der Medizin, der Psychologie usw.

BEISPIELE

Hier sind einige Beispiele für Zufallsexperimente, die durch ihre Wirkung und durch die Menge der möglichen Ergebnisse (ihren **Stichprobenraum) gekennzeichnet sind.**

Aktion	Probenraum
Eine Münze werfen	= *{Kopf, Zahl}*
Einen Würfel werfen	= {1, *2, 3, 4, 5,* 6}
Mit einem einen Ball aus einem Behälter ziehen grüner Ball, ein blauer Ball und ein gelber Ball	= *{Blau, Grün, Gelb}*
Niederschlag in einen bestimmten Zeitraum	= Satz von nichtnegative reelle Zahlen (unendliche nichtabzählbare Menge)

WEITERLESEN

Fall

Wahrscheinlichkeit

Probenraum

Zufallszahl

Eine Zufallszahl nennt man die Realisierung von eine gleichmäßig verteilte Zufallsvariable im Intervall [0, 1].

DOMÄNEN UND EINSCHRÄNKUNGEN

Wenn die oben auferlegten Bedingungen nicht strikt überprüft werden, d. h. wenn die Verteilung am Ort der Probenahme nicht genau ist,

Machine Translated by Google

Da die **Gleichverteilung** und die aufeinanderfolgenden Ziehungen nicht unbedingt unabhängig voneinander sind, sagen wir, dass die Menge der zufällig erhaltenen Zahlen eine Reihe von Pseudozufallszahlen bildet.

BEISPIELE

Ein Behälter fasst 10 Bälle von 0 bis 9. Wenn unsere Ziel ist es, eine Zufallszahl zwischen 0 und 1 mit *m* Dezimalstellen zu erhalten. Wir müssen *m* Kugeln mit Ersetzung zeichnen und versuchen, eine Anzahl von *m* Ziffern zu bilden . Dann betrachten wir diese Ziehungen als ob es nur eine Zahl gäbe und dividieren sie durch *10 m.*

Als Beispiel nehmen wir an, dass wir nacheinander die folgenden Kugeln ziehen: 8, 7, 5, 4 und 7. Die entsprechende Zufallszahl ist 0,87547.

WEITERLESEN

- **Generierung von Zufallszahlen**
- **Gleichmäßige Verteilung**

Zufallszahlengenerierung

Siehe **Generierung von Zufallszahlen.**

Zufällige Variable

Eine **Variable ,** deren Wert durch das Ergebnis eines **Zufallsexperiments** bestimmt wird , wird Zufallsvariable genannt.

MATHEMATISCHE ASPEKTE

Eine Zufallsvariable wird im Allgemeinen durch einen der letzten Buchstaben des Alphabets (in Großbuchstaben) bezeichnet. Es handelt sich um eine reellwertige Funktion, die im **Probenraum** definiert ist . Mit anderen Worten, eine echte ZufallsvariableX ist eine Abbildung von auf:

$$X : ÿ \qquad .$$

Wir verweisen auf:

- Diskrete Zufallsvariablen, wenn die Menge aller möglichen Werte für die Zufallsvariable entweder endlich oder unendlich ist fähig;
- Kontinuierliche Zufallsvariablen, wenn die Menge aller möglichen Werte für die Zufallsvariable unendlich überzählbar ist.

BEISPIELE

Beispiele für diskrete Zufallsvariablen

Wenn wir zwei Würfel gleichzeitig würfeln, sind die 36 möglichen Ergebnisse die folgenden:

$$= \{(1, 1),(1, 2),...,(1, 6),$$
$$(2, 1),(2, 2),...,(2, 6),..., (6, 1),(6,$$
$$2),...,(6, 6)\} .$$

Die Zufallsvariable X = „Summe zweier Würfel“ ist eine diskrete Zufallsvariable. Es nimmt seine Werte in der endlichen Menge $E = \{2, 3, 4,..., 11, 12\}$ an, d. h. die Summe zweier Würfel kann *2, 3, 4,...* oder 12 sein Beispiel: *X* ist eine vollständig positive Zufallsvariable und die Operation stammt aus E.

Es ist möglich , den verschiedenen Werten der Zufallsvariablen *X* eine **Wahrscheinlichkeit** zuzuordnen nehmen kann.

Der Wert „Summe gleich 2“ ergibt sich dann nur durch das Ergebnis (1, 1). Andererseits wird der Wert „Summe gleich 8“ durch die Ergebnisse (2, 6), (3, 5), (4, 4), (5, 3) und (6, 2) erhalten.

Da es 36 Ergebnisse mit gleicher Wahrscheinlichkeit enthält, können wir behaupten, dass die „Chance“ oder die „Wahrscheinlichkeit“, eine „Summe gleich 1 2“ zu erhalten, 36 beträgt ; Die Wahrscheinlichkeit, eine „Summe“ zu erhalten gleich 8“ ist 5 usw. 36 ,

Beispiel einer kontinuierlichen Zufallsvariablen

Solche Zufallsvariablen charakterisieren mea

Sichere wie zum Beispiel der Zeitraum

Machine Translated by Google

eines im Intervall definierten Telefongesprächs
]0, ÿ[, Windrichtung definiert im Intervall [0, 360[usw.

WEITERLESEN

Dichtefunktion
Wahrscheinlichkeit
Zufälliges Experiment
Probenraum
Wert

Randomisierung

Randomisierung ist ein Verfahren, mit dem
Die Behandlungen werden nach dem Zufallsprinzip den
Versuchseinheiten zugeordnet. Es wird verwendet, um Subjektivität zu vermeiden
Behandlungsbeeinträchtigungen. Es ermöglicht die Beseitigung
ein systematischer Fehler, der durch nicht kontrollierbare Fehler verursacht wird
Faktoren, die auch bei wiederholten Experimenten bestehen
bleiben. Um dies zu vermeiden, kann Randomisierung verwendet werden
jegliche Voreingenommenheit, die die Daten beschädigen könnte.
Schließlich ist Randomisierung notwendig, wenn wir das anstreben
Schätzen Sie den experimentellen Fehler ab.

GESCHICHTE

Im September **1919Fisher, Ronald Aylmer**
begann seine Karriere an der Rothamsted Experimental
Station, wo landwirtschaftliche Forschung betrieben wurde
stattfand. Er begann mit historischen Daten und
berücksichtigte dann Daten
aus landwirtschaftlichen Versuchen, für die er die
Varianzanalyse entwickelte . Diese Versuche
veranlassten ihn, experimentelle Designs zu entwickeln.
Er entwickelte hocheffiziente experimentelle Designs
unter Verwendung der Prinzipien der Randomisierung,
des lateinischen Quadrats, des randomisierten Blocks und
faktorielle Vereinbarungen.
Alle diese Methoden finden sich in seiner Arbeit wieder
Statistische Methoden für wissenschaftliche Mitarbeiter
(1925). 1926 erschien sein Artikel über die Anordnung
seiner landwirtschaftlichen Experimente
das Buch *The Design of Experiments (1935).*

WEITERLESEN

Vollständig zufälliges Design
Versuchsplanung
Experiment
Versuchseinheit
Behandlung

VERWEISE

Fisher, RA: Statistische Methoden für
Forschungsmitarbeiter. Oliver & Boyd, Edinburgh
(1925)

Fisher, RA: Die Anordnung von Feldexperimenten. J.
Landwirtschaftsministerium. **33,** 503–513
(1926)

Fisher, RA: Das Design von Experimenten.
Oliver & Boyd, Edinburgh (1935)

Fisher, RA: Statistische Methoden, experimentelles
Design und wissenschaftliche Schlussfolgerungen.
Eine Neuauflage von Statistical Methods
für wissenschaftliche Mitarbeiter: „The Design of
Experimente“ und „Statistische Methoden
und wissenschaftliche Schlussfolgerung.“ Bennett, JH,
Yates, F. (Hrsg.) Oxford University Press,
Oxford (1990)

Randomisiertes Blockdesign

R

Ein randomisiertes Blockdesign ist ein **experimentelles**
Design, bei dem die experimentellen Einheiten vorhanden sind
befinden sich in Gruppen, die Blöcke genannt werden. Die Behandlungen
werden dem Experiment zufällig zugeordnet
Einheiten in jedem Block. Bei allen Behandlungen
erscheinen mindestens einmal in jedem Block, den wir haben
ein vollständig zufälliges Blockdesign. Ansonsten liegt
eine unvollständige Randomisierung vor
Blockdesign.
Diese Art von Design dient der Minimierung
die Auswirkungen systematischer Fehler. Konzentriert
sich der Experimentator ausschließlich auf die
Unterschiede zwischen den Behandlungen, sind die Wirkungen bedingt

Machine Translated by Google

zu Variationen zwischen den verschiedenen Blöcken sollten beseitigt werden.

GESCHICHTE

Siehe **experimentelles Design.**

BEISPIELE

Ein Bauer besitzt fünf Grundstücke, auf denen er Mais anbauen möchte. Er will laufen ein Experiment, da er zwei Arten von Mais und zwei Arten von Düngemitteln hat. Darüber hinaus weiß er, dass seine Parzellen in Bezug auf die Sonneneinstrahlung recht heterogen sind und es daher zu einem systematischen Fehler kommen könnte, wenn die Sonneneinstrahlung tatsächlich vorhanden ist erleichtern den Maisanbau.
Der Bauer teilt das Land in fünf Blöcke und ordnet jedem Block nach dem Zufallsprinzip die folgenden vier Behandlungen zu:

A: Mais 1 mit Dünger 1
B: Mais 1 mit Dünger 2
C: Mais 2 mit Dünger 1
D: Mais 2 mit Dünger 2

Er beseitigt damit die Ursache des Systemfehlers **durch** den **Faktor** Sonne. Indem er in jedem Block nach dem Zufallsprinzip dasselbe Verfahren anwendet, bevorzugt er weder eine spezielle Maissorte gegenüber den anderen noch eine spezielle Art von Düngemittel gegenüber den anderen. Die folgende Tabelle zeigt die zufällige Anordnung der vier Behandlungen in den fünf Blöcken:

Block 1	*DCAB*
Block 2	*CBAD*
Block 3	*ABCD*
Block 4	*ACBD*
Block 5	*DCAB*

Eine **Varianzanalyse** ermöglicht es, nach Entfernung von Blockeffekten Rückschlüsse darauf zu ziehen, ob ein signifikanter Unterschied zwischen den Behandlungen besteht.

WEITERLESEN

- **Varianzanalyse**
- **Versuchsplanung**

Bereich

Die Reichweite ist das am einfachsten zu berechnende **Maß für die Streuung .**
Betrachten Sie eine Reihe von **Beobachtungen** relativ zu eine **quantitative Variable** X. Der Bereich ist definiert als die Differenz zwischen dem höchsten beobachteten Wert und dem niedrigsten beobachteten Wert Wert.

MATHEMATISCHE ASPEKTE

Betrachten Sie eine Reihe von **Beobachtungen** relativ zu eine **quantitative Variable** X.
Wenn wir mit Xmax den **Wert** der höchsten Beobachtung in einer Reihe von **Beobachtungen** und mit Xmin den niedrigsten Wert bezeichnen , dann ist der Bereich gegeben durch:

$$\text{Bereich} = X_{max} - X_{min} .$$

Wenn **Beobachtungen** in Klassen gruppiert werden, entspricht der Bereich der Differenz zwischen der Mitte der beiden Extremklassen
es. Sei $\bar{y}_1$ der Mittelpunkt der ersten Klasse und $\bar{y}_k$ der Mittelpunkt der letzten Klasse. Der Bereich ist gleich:

$$\text{Bereich} = \bar{y}_k - \bar{y}_1 .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Spannweite ist ein **Streuungsmaß ,** das nicht sehr häufig verwendet wird, da es nur einen Hinweis auf die Spanne der **Werte** gibt , aber keine Informationen über die Verteilung innerhalb dieser Spanne liefert.

Dennoch hat es den großen Vorteil, dass es einfach zu berechnen ist. Allerdings sollte man bei der Interpretation sehr vorsichtig sein.

Machine Translated by Google

Die Reichweite ist ein **Maß für die Streuung ,** das nur die beiden **Extremwerte** berücksichtigt . Diese Methode kann daher zu falschen Interpretationen führen, wenn die Anzahl der Kinder in sieben Familien 0, 0, 1, 1, 2, 3 und 6 beträgt, die Spanne 6 beträgt, während alle Familien bis auf eine zwischen 0 und 3 Kinder haben. In diesem

In diesem Fall ist die Variabilität relativ gering, während die Spanne hoch ist.

Die Verwendung des Bereichs zur Messung der Streuung (oder Variabilität) stellt eine weitere Unannehmlichkeit dar. Sie kann nur steigen, wenn die Anzahl der **Beobachtungen** zunimmt. Das ist ein Nachteil weil im Idealfall ein Maß für die Variabilität unabhängig von der Anzahl der **Beobachtungen sein sollte.**

Dabei muss man äußerste Vorsicht walten lassen Interpretation eines Bereichs, da es Ausreißer geben kann . Diese haben einen direkten Einfluss auf die **Wert** des Bereichs.

BEISPIELE

Hier ist das Alter von zwei Gruppen von Menschen, die an einem Abendkurs teilnehmen:

Gruppe 1:	33 34 34 37 40 40 40 43 44 45
Gruppe 2:	22 25 31 40 40 41 44 45 50 52

Das **arithmetische Mittel** $\bar{x}$ dieser beiden Mengen von **Beobachtungen** ist identisch:

$$\bar{x} = \frac{390}{10} = 39 .$$

Auch wenn das **arithmetische Mittel** des Alters dieser beiden Gruppen mit 39 Jahren gleich ist, ist die Streuung um den Mittelwert sehr unterschiedlich. Die Spanne für die erste Gruppe beträgt 12 (45 ÿ 33), während die Spanne für die zweite Gruppe 52 ÿ 22 = 30 Jahre beträgt.

Beachten Sie, dass der Bereich nur ein indikatives Maß für die Variabilität der **Werte darstellt;** Es gibt nicht die Form der Verteilung innerhalb dieser Spanne an.

WEITERLESEN

- **Interquartilbereich**
- **Maß der Streuung**

VERWEISE

Anderson, TW, Sclove, SL: Eine Einführung in die statistische Analyse von Daten. Houghton Mifflin, Boston (1978)

Rang einer Matrix

Der Rang einer **Matrix** ist die maximale Anzahl Anzahl linear unabhängiger Zeilen oder Spalten.

MATHEMATISCHE ASPEKTE

Sei A eine $m \times n$ **-Matrix.** Wir definieren es Rang, bezeichnet mit rA, als maximale Anzahl linear unabhängiger Zeilen oder Spalten. Insbesondere:

$$rA \ddot{y} \min (m, n) .$$

Wenn$rA = \min (m, n)$, dann sagen wir, dass Matrix A den vollen Rang hat.

Wenn der Rang einer quadratischen Matrix mit Ordnungen gleich n ist, ist ihre **Determinante** ungleich Null; mit anderen Worten, diese Matrix ist invertierbar.

BEISPIELE

Sei A eine 2 × 3-Matrix:

$$A = \begin{matrix} \ddot{y}732 \\ 104 \end{matrix} .$$

Sein Rang ist kleiner oder gleich min(2, 3). Die Anzahl der linear unabhängigen Zeilen oder Spalten beträgt 2. Folglich ist $rA = 2$.

WEITERLESEN

- **Bestimmend**
- **Matrix**

Machine Translated by Google

Rao, Calyampudi Radhakrishna

Rao, Radhakrishna wurde 1920 geboren
Hadagali, Karnata, Indien. Er erhielt seine
Ph.D. 1948 an der Universität Cambridge.
Seit 1989 erhielt er die Ehrendoktorwürde zahlreicher
Universitäten in der Umgebung
Welt, insbesondere die Universität Neuenburg,
in der Schweiz.

Sein Beitrag zur Entwicklung der statistischen Theorie
und ihrer Anwendungen ist vergleichbar mit den Werken
von **Fisher, Ronald Aylmer,**
Neyman, Jerzy und andere wichtige moderne
Statistiker. Zu seinen Beiträgen zählen die
Am wichtigsten sind das Fisher-Rao-Theorem,
das Rao-Blackwell-Theorem, das Cramer-Theorem
Rao-Ungleichung und Raos U-Test. Er ist der
Mitautor von 12 Büchern und mehr als 250
wissenschaftliche Artikel. Sein Buch *Statistics and*
Truth: Putting Chance to Work befasst sich mit
Grundlegende Logik und statistische Anwendungen und
wurde in viele Sprachen übersetzt
Messgeräte.

Ausgewählte Werke und Veröffentlichungen von Rao, Calyampudi Radhakrishna:

1973 Lineare statistische Inferenz und ihre
Bewerbungen, 2. Aufl. Wiley, Neu
York.

1989 (mit Shanbhag, DN) Weitere Erweiterungen von
Choquet–Deny und Deny
Theoreme mit Anwendungen in der
Charakterisierungstheorie. QJ Math., 40,
333–350.

1992 (mit Zhao, LC) Lineare Darstellung von R-
Schätzungen in linearen Modellen.
Dürfen. J. Stat., 220:359–368.

1997 Statistik und Wahrheit: Chance setzen
to Work, 2. Aufl. Weltwissenschaftlich,
Singapur.

1999 (mit Toutenburg, H.) Lineare Modelle. Kleinste
Quadrate und Alternativen,
2. Aufl. Springer-Reihe in Statistik.
Springer,BerlinHeidelbergNeu
York.

WEITERLESEN

Gauß-Markov-Theorem

Verallgemeinerte Umkehrung

Regressionsanalyse

Die Regressionsanalyse ist eine Technik, die es
ermöglicht, die **Beziehung** zu untersuchen und zu messen
zwischen zwei oder mehr Variablen. Beginnend mit
In **einer Stichprobe** registrierte Daten versucht die
Regressionsanalyse, eine Schätzung von Amathe zu ermitteln
Matische Beziehung zwischen zwei oder mehr Vari
fähig. Ziel ist es, den Wert abzuschätzen
eine Variable als Funktion einer oder mehrerer anderer
äh Variablen. Die geschätzte Variable wird aufgerufen
die **abhängige Variable** und wird üblicherweise mit Y
bezeichnet. Im Gegensatz dazu sind die Variablen
die die Variationen in *Y* erklären, werden aufgerufen
unabhängige Variablen, und sie werden bezeichnet
von X.
Wenn Y nur von oneX abhängt, haben wir eine einfache
Regressionsanalyse, aber wenn *Y* abhängt
auf mehr als einer unabhängigen Variablen, wir
haben mehrere Regressionsanalysen. Wenn die
Beziehung zwischen der abhängigen und der
unabhängigen Variable linear ist, dann liegt eine lineare
Regressionsanalyse vor.

GESCHICHTE

Der Pionier der linearen Regressionsanalyse,
Boscovich, Roger Joseph, ein Astronom
war neben seiner Tätigkeit als Arzt einer der ersten
eine Methode zur Bestimmung der Koeffizienten einer
Regressionsgeraden zu finden. Um eine Linie zu erhalten

Machine Translated by Google

Indem man so nah wie möglich an allen Beobachtungen vorbeigeht, sollten zwei Parameter bestimmt werden und zwar so, dass zwei Bedingungen von Boscovich, RJ eingehalten werden:

1. Die Summe der Abweichungen muss gleich sein null.
2. Die Summe der Abweichungen als Absolutwert muss minimal sein.

Boscovich, RJ hat diese Bedingungen geschaffen um 1755–1757. Er verwendete eine Methode von geometrische Auflösung, die er wann anwendete Messung der Länge von fünf Erdteilen Meridiane.

Im Jahr 1789 **de Laplace, Pierre Simon,** in seinem Arbeit *Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pend ule,* nahm die beiden von auferlegten Bedingungen an Boscovich, RJ und etablierte eine algebraische Methode zur Auflösung von Boscovich

Algorithmus.

Zwei weitere Mathematiker, **Gauss, Carl Friedrich** und **Legendre, Adrien Marie,** scheinen es entdeckt zu haben, ohne es zu bestätigen miteinander, die Methode der **kleinsten Quadrate .** Es scheint, dass Gauss, CF es verwendet hat seit 1795. Aber es ist Legendre, AM, der veröffentlichte 1805 die Methode der kleinsten Quadrate in seinem Werk *Nouvelles méthodes pour la détermination des orbites des comètes,* wo wir den Anhang *Sur la méth ode des moindres carrés finden.* Eine Kontroverse über den wahren Urheber der Entdeckung, die 1809 zwischen den beiden Männern ausbrach. Gauss, CF veröffentlichte seine Methode unter Berufung auf Verweise auf seine bisherigen Werke. Beachten Sie das im selben Punkt, ein Amerikaner namens Adrian, Robert, der keine Kenntnis von der Arbeit von Gauß und Léendre hatte, arbeitete auch mit Least zusammen Quadrate.

Das 18. Jahrhundert prägte die Entwicklung der Methode der kleinsten Quadrate. Nach Nach Stigler, S. (1986) ist diese Entwicklung stark mit drei Problemen verbunden aus dieser Zeit. Das erste betraf die mathematische Darstellung und Bestimmung der Bewegungen des Mondes. Der

Zweitens ging es um die Erklärung der säkularen Ungleichheit, die in der Bewegung von beobachtet wurde Jupiter und Saturn. Der dritte schließlich betraf die Bestimmung der Form des Erde.

Laut demselben Autor findet man in der Werke von **Galton, Francis** der Ursprung des Begriff lineare Regression. Seine Forschung konzentrierte sich zur Vererbung: Der Begriff der Regression entstand aus einer Analyse der Vererbung innerhalb der Familie Mitglieder.

Im Jahr 1875 führte Galton, CF, ein **Experiment** mit kleinen, in sieben Erbsen verteilten Erbsen durch Gruppen nach Gewicht. Für jede Gruppe nahm er zehn Samen und bat sieben Freunde, sie zu kultivieren vate sie. Er kultivierte somit 490 Keime.

Er entdeckte, dass jede nach Gewicht klassifizierte Samengruppe einer Normalverteilung folgte und die Kurven, die auf unterschiedlichen Gewichten zentriert waren, in einer gleichermaßen normalen Verteilung verteilt waren; die Variabilität von Die verschiedenen Gruppen waren identisch. Galton, F. sprach von „Reversion“ (in der Biologie bezeichnet „Reversion“ die Rückkehr zu einem primitiven Typus). Es war eine lineare Umkehrung Galton, F., war Zeuge, da die sieben Samengruppen normal verteilt waren, im Vergleich nicht zum Gewicht ihrer Eltern, aber auf einen Wert nahe der Gesamtbevölkerung bedeuten.

Die menschliche Vererbung war auch ein Forschungsthema für Galton, F., der versuchte, die Vererbung zu definieren Verhältnis zwischen der Länge der Elternsamen und das ihrer Nachkommen. 1889 veröffentlichte er sein Werk unter dem *Titel Natural Inher itance.* Auf diesem Gebiet nutzte er ab 1885 das Begriff Regression, und in seinen letzten Werken landete er schließlich beim Konzept der **Korrelation.**

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Siehe **LAD-Regression, multiple lineare Regression** und **einfache lineare Regression.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Ziel der Regressionsanalyse besteht nicht nur darin, die Beziehung zwischen der **abhängigen Variablen** und der/den **unabhängigen Variable(n) zu bestimmen.** Ziel der Regressionsanalyse ist es auch, die Verlässlichkeit von Schätzungen und damit die Verlässlichkeit der erhaltenen Vorhersagen festzustellen.

Mit der Regressionsanalyse lässt sich außerdem prüfen, ob die Ergebnisse statistisch sind signifikant ist und ob die **Beziehung** zwischen den Variablen real oder nur scheinbar ist.

Die Regressionsanalyse findet vielfältige Anwendungsmöglichkeiten in allen wissenschaftlichen Bereichen. Tatsächlich können wir von dem Moment an, in dem wir eine Beziehung zwischen Variablen bestimmen, zukünftige Werte vorhersagen, vorausgesetzt, dass die Bedingungen unverändert bleiben und immer ein **Fehler vorliegt.** Üblicherweise überprüfen wir die Gültigkeit einer **Regressionsanalyse** durch eine Residuenanalyse.

BEISPIELE

Siehe **lineare Regression (multiple oder einfach), LAD-Regression** und **Ridge.**

WEITERLESEN

Analyse von Residuen
Varianzanalyse
Bestimmtheitsmaß
Hutmatrix
Kleinsten Quadrate
Hebelpunkt
Multiple lineare Regression
Normale Gleichungen
Restwert
Einfache lineare Regression

VERWEISE

Eisenhart, C.: Boscovich und die Kombination von Beobachtungen. In: Kendall, M., Plackett, RL (Hrsg.) Studies in the History of Statistics and Probability, vol. II. Grif fin, London (1977)

Galton, F.: Natürliche Vererbung. Macmillan, London (1889)

Gauss, CF: Theoria Motus Corporum Coelestium. Werke, 7 (1809)

Laplace, PS von: Auf den Gradmessern der Mittagszeit und auf den Pendellängen. Geschichte der Royal Academy of Inscriptions et Belles Lettres, mit den Memoiren der littérature tirées des

Registriert diese Akademie. Paris (1789)

Legendre, AM: Neue Methoden zur Bestimmung der Umlaufbahnen von Kometen.

Courcier, Paris (1805)

Plackett, RL: Studien zur Geschichte der Wahrscheinlichkeit und Statistik. In: Kendall, M., Plackett, RL (Hrsg.) Die Entdeckung der Methode der kleinsten Quadrate. Bd. II. Griffin, London (1977)

Stigler, S.: Die Geschichte der Statistik, die Messung der Unsicherheit vor 1900. Belknap, London (1986)

Ablehnungsbereich

Der Ablehnungsbereich ist das in der **Stichprobenverteilung** der untersuchten **Statistik** gemessene **Intervall ,** das in einem **Hypothesentest** zur Ablehnung der **Nullhypothese** H0 führt. Der Ablehnungsbereich ist **komplementär** zum **Akzeptanzbereich** und mit einer **Wahrscheinlichkeit** ÿ verknüpft, die als **Signifikanzniveau** des Tests oder **Typ-I-Fehler bezeichnet wird.**

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Betrachten Sie einen **Hypothesentest** für einen **Parameter** ÿ. Wir verwenden die **Statistik** *T* , um zu schätzen

Passen Sie diesen Parameter an.

Fall 1: Einseitiger Test (rechtes Ende)

Die Hypothese lautet wie folgt:

$$H0 : ÿ = ÿ0$$

$$H1 : ÿ>ÿ0 .$$

wobei ÿ0 der voreingestellte Wert für den Parameter ÿ ist.

Bei einem einseitigen Test (Test für den rechten Schwanz) entspricht der Ablehnungsbereich einem auf der rechten Seite des **kritischen Werts** begrenzten **Intervall :**

Ablehnungsbereich = [kritischer Wert; ÿ[, und

Akzeptanzbereich =]ÿÿ; kritischer Wert[,

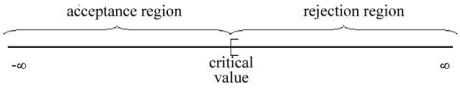
vorausgesetzt, dass der kritische Wert gleich ist:

$$ÿT + zÿ \cdot ÿT ,$$

Dabei ist *ÿT* der **Mittelwert** der **Stichprobenverteilung** für die Statistik T, *ÿT* der **Standardfehler** für die Statistik T und ÿ das **Signifikanzniveau** des Tests.

Der Wert zÿ wird aus **statistischen Tabellen** der Verteilung T ermittelt. Wenn der Wert von *T*

Der in der **Stichprobe** berechnete Wert liegt im Ablehnungsbereich, d. h. wenn *T* größer oder gleich dem kritischen Wert ist, muss die **Nullhypothese** H0 zugunsten der **Alternativhypothese** H1 abgelehnt werden. Andernfalls kann die Nullhypothese H0 nicht abgelehnt werden.



Fall 2: Einseitiger Test (linkes Ende)

Die Hypothese lautet wie folgt:

$$H0 : ÿ = ÿ0$$

$$H1 : ÿ<ÿ0 .$$

Bei einem einseitigen Test (Test für den rechten Schwanz) wird der Ablehnungsbereich auf der linken Seite durch den kritischen Wert begrenzt:

Ablehnungsbereich =]ÿÿ; kritischer Wert] ,

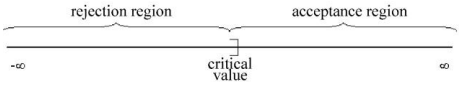
und der Akzeptanzbereich ist:

Akzeptanzbereich =]kritischer Wert; ÿ[,

vorausgesetzt, der kritische Wert ist gleich:

$$ÿT ÿ zÿ \cdot ÿT .$$

Wenn der in der **Stichprobe** berechnete Wert von *T* im Ablehnungsbereich liegt, das heißt, wenn *T* kleiner oder gleich dem kritischen Wert ist, muss die **Nullhypothese** H0 zugunsten der **Alternativhypothese** H1 verworfen werden. Andernfalls kann die Nullhypothese H0 nicht abgelehnt werden.



Fall 3: Zweiseitiger Test

Die Hypothese lautet wie folgt:

$$H0 : ÿ = ÿ0$$

$$H1 : ÿ = ÿ0$$

Bei einem **zweiseitigen Test** wird der Ablehnungsbereich in zwei Intervalle unterteilt:

Ablehnungsbereich =\ unterer kritischer Wert;

oberer kritischer Wert

und der Akzeptanzbereich:

Akzeptanzbereich = unterer kritischer Wert; oberer

kritischer Wert ,

wobei die kritischen Werte gleich sind:

$$\frac{}{2} \cdot ÿT \text{ und } ÿT + z ÿ \; ÿT ÿ \; z ÿ_{\frac{}{2}} \cdot ÿT .$$

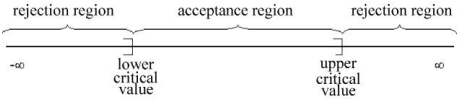
R

Machine Translated by Google

Wenn der in der **Stichprobe** berechnete Wert von *T* liegt im Ablehnungsbereich, das heißt, wenn *T* größer ist größer oder gleich dem oberen kritischen Wert orifT ist kleiner oder gleich dem unteren kritischen Wert ist, muss die **Nullhypothese** H0 verworfen werden zugunsten der **Alternativhypothese** H1.

Andernfalls kann die Nullhypothese H0 nicht gelten abgelehnt.

Die folgende Abbildung veranschaulicht die Akzeptanz- und Ablehnungsbereiche für die oben genannten Fälle.



DOMÄNEN UND EINSCHRÄNKUNGEN

Bei einem **einseitigen Test** beträgt der Ablehnungsbereich eindeutig und wird erkannt:

- Auf der rechten Seite der **Probenverteilung für** eine einseitige Prüfung im rechten Ende.
- Auf der linken Seite der Stichprobenverteilung für einen einseitigen Test im linken Ende.

Im Gegensatz dazu ist der Ablehnungsbereich geteilt im **zweiseitigen Test in zwei Teile zerlegt,** die befinden sich ganz links und rechts der Stichprobenverteilung. An jeden einzelnen von Diesen Regionen ist eine **Wahrscheinlichkeit** zugeordnet $\frac{\ddot{y}}{2}$, wobei ÿ das Signifikanzniveau des Tests ist.

BEISPIELE

Ein Unternehmen produziert Stahlseile. Bezogen auf eine Stichprobe von $n = 100$ Einheiten, die verifiziert werden soll ob der Durchmesser der Kabel 0,9 cm beträgt.

Die **Standardabweichung** der Grundgesamtheit beträgt bekanntermaßen gleich 0,05 cm.

Der **Hypothesentest** ist in diesem Fall ein **zweiseitiger Test.** Die Hypothese lautet wie folgt:

Nullhypothese H0 : ÿ = *0,9*

Alternativhypothese H1 : ÿ = *0,9*

Für ein **Signifikanzniveau** von ÿ = 5 % gilt: Wert $z\,\ddot{y}_{\frac{}{2}}$ der **Normaltabelle** beträgt 1,96. Die kritischen Werte sind dann:

$$\ddot{y}x^{-} \pm z\,\ddot{y}_{\frac{}{2}} \cdot \ddot{y}x^{-};$$

wobei ÿx¯ der **Mittelwert** der Stichprobenverteilung für die Mittelwerte ist:

$$\ddot{y}x^{-} = \ddot{y} = 0{,}9$$

und $\ddot{y}x^{-}$ ist der **Standardfehler** des Mittelwerts:

$$\ddot{y}x^{-} = \frac{\ddot{y}}{\ddot{y}n} = \frac{0{,}05}{\ddot{y}100} = 0{,}005$$

$$\ddot{y}\ \ddot{y}x^{-} \pm z\,\ddot{y}_{\frac{}{2}}\ddot{y}x^{-} = 0{,}9 \pm 1{,}96 \cdot 0{,}005$$

$$= 0{,}9 \pm 0{,}0098\ .$$

Das **Intervall]0,8902;** 0,9098[entspricht zum **Akzeptanzbereich** der Nullhypothese.

Der Ablehnungsbereich ist der Komplementärbereich Intervall und gliedert sich in zwei Teile:

$$]\ddot{y}\ddot{y};\ 0{,}8902] \text{ und } [0{,}9098;\ \ddot{y}[\ .$$

Wenn der erhaltene Stichprobenmittelwert darin liegt Intervall muss die **Nullhypothese** H0 sein zugunsten der **Alternativhypothese** H1 abgelehnt.

WEITERLESEN

Akzeptanzregion
Kritischer Wert
Hypothesentest
Signifikanzniveau

VERWEISE

Bickel, RD: Auswahl eines optimalen Ablehnungsbereichs für Mehrfachtests. Arbeiten Artikel, Büro für Biostatistik und Bioinformatik, Medical College of Georgia (2002)

Machine Translated by Google

Beziehung

Der Begriff der Beziehung drückt die Beziehung aus
das zwischen zwei Zufallsvariablen existiert.

Es ist einer der wichtigsten Begriffe in
Statistiken. Es erscheint in Konzepten wie **Klassifikation,**
Korrelation, Abhängigkeit, Regressionsanalyse,
Zeitreihen usw.

WEITERLESEN

Einstufung

Korrelationskoeffizient

Abhängigkeit
Modell

Regressionsanalyse
Zeitfolgen

Relatives Risiko

Das **relative Risiko** wird als Verhältnis definiert
zwischen dem **Risiko** im Verhältnis zu den einzelnen Personen
die einem Faktor ausgesetzt sind. Es misst die
relative Wirkung einer bestimmten Ursache, also der
Grad des Zusammenhangs zwischen einer Krankheit
und seine Ursache.

GESCHICHTE

Siehe **Risiko.**

MATHEMATISCHE ASPEKTE

Das relative Risiko ist also ein Wahrscheinlichkeitsverhältnis
geschätzt von:

$$\text{relatives Risiko} = \frac{\text{Gefahr der Exposition}}{\text{Risiko der Nichtexposition}} = \frac{\text{Inzidenzrate für exponierte Personen}}{\text{Inzidenzrate für Nichtexponierte}}.$$

Beachten Sie, dass das relative Risiko ein Maß für ist
eine Einheit aus, da sie als das Verhältnis von zwei definiert ist
Größen, die in derselben Einheit gemessen werden.

DOMÄNEN UND EINSCHRÄNKUNGEN

Das relative Risiko drückt das Risiko aus
Einheiten, die einem Faktor als Vielfaches davon ausgesetzt sind
Risiko, dass die Einheiten diesem Faktor nicht ausgesetzt sind:

Gefahr der exponierten Einheiten
= Risiko nicht exponierter Einheiten
· relatives Risiko .

Wir wollen den Zusammenhang zwischen einem Risiko untersuchen
Faktor und eine Krankheit. Wenn kein Ass vorhanden ist
Ziation sollte das relative Risiko theoretisch sein
gleich 1 sein. Bei Exposition gegenüber dem Risikofaktor
erhöht die Zahl der Krankheitsfälle,
dann wird das relative Risiko größer als 1.
Wenn andererseits die Exposition gegenüber dem Faktor
verringert die Zahl der Krankheitsfälle,
dann liegt das relative Risiko zwischen 0 und 1.

Es ist einfach, ein **Konfidenzintervall** $P = (1 ÿ ÿ)$ % für
das relative Risiko zu berechnen . Wir betrügen
sider hier den folgenden allgemeinen Fall:
Das Konfidenzintervall wird im berechnet
logarithmisch skaliert und anschließend zurücktransformiert
in die arithmetische Skala. Der Kleinbuchstabe let

Weitere Informationen finden Sie in der folgenden Tabelle:

	Krankheit	Nicht krankheitsbedingt	insgesa
Ausgesetzt	A	$n ÿ a$	N
Nicht belichtet	C	$m ÿ cm$	

Das Konfidenzintervall $P = (1 ÿ ÿ)$ % für
Das relative Risiko RR ist dann gegeben durch:

$$\exp \ln (RR) \pm zÿ \sqrt{\frac{1}{A} \quad \frac{1}{N} + \frac{1}{C} \quad \frac{1}{M}},$$

wobei „exp“ der Exponent für den Ausdruck ist
in Klammern und zÿ der **Wert** im Normalen
Tisch. Es wird angenommen, dass a, n, c und m sind
groß genug. Wenn das Konfidenzintervall für
Das relative Risiko umfasst dann den Wert 1,0
Wir kommen zu dem Schluss, dass der Belichtungsfaktor dies nicht tut
spielen dabei keine statistisch signifikante Rolle
Wahrscheinlichkeit, an der Krankheit zu erkranken.

R

Machine Translated by Google

BEISPIELE

Wir verwenden hier Beispiel 1 aus dem folgenden Abschnitt **Zurechenbares Risiko.** Das relative Brustrisiko

Krebs im Zusammenhang mit Rauchexposition entspricht dem **Risiko** (oder **der Inzidenzrate)** für jede Expositionshöhe dividiert durch das Risiko (bzw Inzidenzrate) für Nichtexponierte:

Gruppe	Inzidenzrate (100000 Jahre) (I_A)	Relativ Risiko (I_A / I_0)
Nicht belichtet	57,0	1,0
Passivraucher 126.2		2.2
Aktive Raucher 138,1		2.4

Das relative Risiko beträgt 2,2 für Passivraucher und 2,4 für Aktivraucher. Verglichen zum Risiko des Nicht-Ausgesetzten, zum Risiko für Aktive Raucher und Passivraucher ist jetzt 2,4- und 2,2-mal größer. Es gibt ein ungefähres bescheidene Assoziation, die die Tatsache unterstreicht dass Brustkrebs auch von Faktoren abhängt außer Rauchen. Das relative Risiko liefert somit eine teilweise Beschreibung der Beziehung zwischen Rauchen und Brustkrebs. Allein Es erlaubt keine eindeutigen Schlussfolgerungen und weist darauf hin, dass Rauchen wahrscheinlich ist verantwortlich für 50 % der Brustkrebsfälle in der Bevölkerung.

WEITERLESEN

- **Zurechenbares Risiko**
- **Vermeidbares Risiko**
- **Ursache und Wirkung in der Epidemiologie**
- **Inzidenzrate**
- **Quoten und Quotenverhältnis**
- **Prävalenzrate**
- **Risiko**

VERWEISE

Cornfield, J.: Eine Methode zur Schätzung von Vergleichsraten anhand klinischer Daten. Appli Kationen zu Lungen-, Brustkrebs und Gebärmutterhals. J. Natl. Krebsinst. **11,** 1269–75 (1951)

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Kleiner Brauner, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Morabia, A.: L'Épidémiologie Clinique. Ausgaben „Que sais-je?“. Presses Univer Sitaires de France, Paris (1996)

Resampling

Bei der Verwendung einer **Probe** zur Schätzung eines bestimmten Werts **Parameter** bezüglich **der Bevölkerung** von aus dem die Stichprobe gezogen wird, können wir verwenden dieselbe Probe mehr als einmal, um unsere zu verbessern statistische Analyse und Schlussfolgerung. Dies geschieht durch Probenahme innerhalb der Ausgangsprobe. Dies ist die Idee hinter Resampling-Methoden. Mithilfe von Resampling können wir den Bias berechnen oder die Standardabweichung einer Schätzung Parameter, oder wir können auch Schlussfolgerungen ziehen aus Konstruktionskonfidenzintervall und Testen von Hypothesen für den geschätzten Parameter. Zu den Techniken, die Resample nutzen, zählen beispielsweise **Bootstrap, Jackknife** und Cross Validation. Ihre Besonderes Interesse gilt nichtparametrischen Techniken, weil:

- Die genaue Verteilung der Daten ist nicht gegeben bekannt und
- Die Daten stammen aus einem nichtlinearen **Design,** was sehr schwer zu modellieren ist.

Andererseits besteht ein damit verbundenes Problem Verfahren entstehen, wenn die Probe dies nicht tut

Machine Translated by Google

Wir repräsentieren die Bevölkerung und laufen Gefahr, völlig falsche Schlussfolgerungen zu ziehen.

GESCHICHTE Die Klappmessermethode wurde in den 1950er Jahren eingeführt, um die **Verzerrung** einer **Schätzung zu verringern.** In den 1960er und 1970er Jahren wurde das Resampling häufig **eingesetzt ,** um die Varianz zu schätzen .
Die **Bootstrap-** Methode wurde Ende der 1970er Jahre von Efron, B. erfunden und seitdem von zahlreichen Forschern in verschiedenen Bereichen weiterentwickelt.

MATHEMATISCHE ASPEKTE

Siehe **Bootstrap, Jackknife.**

WEITERLESEN

- **Bootstrap**
- **Jackknife-Methode**
- **Monte-Carlo-Methode**
- **Randomisierung**

LITERATUR

Edington, ES: Randomisierungstests. Marcel Dekker, New York (1980)

Efron, B.: Das Klappmesser, der Bootstrap und andere Resampling-Pläne, CBMS Monograph. Nr. 38. SIAM, Philadelphia (1982)

Fisher, RA: Das Design von Experimenten. Oliver & Boyd, Edinburgh (1935)

Pitman, EJG: Signifikanztests, die auf Proben aus beliebigen Populationen angewendet werden können. Beilage, J. Roy. Stat. Soc. 4, 119–130 (1937)

Restwert

Residuen werden als Differenz definiert zwischen den beobachteten Werten und dem Schätzwert gepaarte (angepasste) Werte eines Regressionsmodells . Sie stellen dar, was durch die Regressionsgleichung nicht erklärt wird.

GESCHICHTE

Siehe **Fehler.**

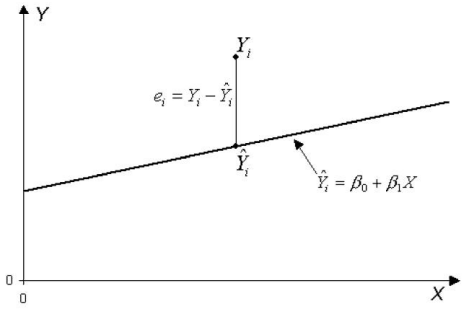
MATHEMATISCHE ASPEKTE

Die Residuen *ei* sind gegeben durch:

$$ei = Yi \ddot{y} Y\hat{}i , \qquad i = 1,...,n,$$

Dabei bezeichnet *Yi* eine **Beobachtung** und Y^i seinen angepassten **Wert,** der durch Anpassen eines Regressionsmodells erhalten wird .

Die folgende Grafik veranschaulicht die Residuen in einer **einfachen linearen Regression:**



BEISPIELE

Siehe **Restanalyse.**

WEITERLESEN

- **Analyse von Residuen**
- **Fehler**
- **Hutmatrix**
- **Hebelpunkt**
- **Multiple lineare Regression**
- **Regressionsanalyse**
- **Einfache lineare Regression**

R

Machine Translated by Google

Reis, Stuart Arthur

Rice, Stuart Arthur (1889–1969) studierte an der University of Washington (1912 erhielt er seinen Bachelor of Science und 1915 seinen Master). Seine akademische Arbeit widmete sich der Entwicklung einer Methodik in den Sozialwissenschaften.

In den frühen 1930er Jahren trat Rice, SA, der Regierung von Franklin D. Roosevelt bei. Er wurde stellvertretender Direktor der Volkszählung Büro von 1933 bis Anfang der 1950er Jahre. Sein Hauptbeitrag zur Statistik besteht in der Modernisierung des Einsatzes von Statistiken in der Regierung. Er förderte die Entwicklung von COGSIS (Committee on Government Statistics and Information Services) und arbeitete an der Adaption moderner Stichprobentechniken sowie der mathematischen Statistik in Bundesbehörden. 1955 gründete er Stuart A. Rice Associates (später Surveys & Research Corporation), eine statistische Beratungsagentur für den öffentlichen und privaten Sektor.

In den 1960er Jahren ging er in den Ruhestand.

WEITERLESEN

Volkszählung

Offizielle Statistiken

Ridge-Regression

Das Hauptziel der Ridge-Methode besteht darin, den mittleren quadratischen Fehler der Schätzungen zu minimieren, d. h. Verzerrung und Varianz zu kompensieren. Indem wir Kollinearität als ein Problem der Qua-Singularität von **W** W betrachten, wobei **W die** Matrix der erklärenden Variablen

Singularitäten.

GESCHICHTE

Die Ridge-Regression wurde 1962 von Hoerl, AE, eingeführt, der feststellte, dass die Existenz einer Korrelation zwischen erklärenden Variablen bei der Anwendung der Methode der kleinsten Quadrate zu Fehlern bei der Schätzung führen kann. Als Alternative entwickelte er die Ridge-Regression, die es ermöglicht, Schätzungen zu berechnen, die verzerrt sind, aber zumindest eine geringere Varianz aufweisen Quadratschätzungen. Während die Methode der kleinsten Quadrate unverzerrte Schätzungen liefert, minimiert die Ridge-Methode den **mittleren quadratischen Fehler** der Schätzungen.

MATHEMATISCHE ASPEKTE

Betrachten wir das **multiple lineare Regressionsmodell :**

$$\mathbf{Y} = \mathbf{X}\ddot{y} + \ddot{y} \quad ,$$

wobei $\mathbf{Y} = (y1,...,yn)$ ein $(n \times 1)$-Vektor der Beobachtungen im Zusammenhang mit dem ist **abhängige Variable** $(n$ Beobachtungen), $\ddot{y} = \ddot{y}0,...,$ ypy1 ist der zu schätzende (p×1) Parametervektor, $\ddot{y} = (\ddot{y}1,..., \ddot{y}n)$ ist der $(n \times 1)$ Fehlervektor und

$$\mathbf{X} = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} 1\ X11 \ ...\ X1(p\ddot{y}1) \\ \vdots \quad \vdots \qquad\qquad \vdots \\ 1\ Xn1 \ ...\ Xn(p\ddot{y}1) \end{matrix} \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix}$$

ist eine $(n \times p)$ **-Matrix** mit unabhängigen Variablen.

Die **Schätzung** von **ÿ** mittels der Methode der **kleinsten Quadrate** ergibt sich aus:

$$\ddot{y} = \ddot{y}\ 0,...,\ \ddot{y}\ p\ddot{y}1 = \mathbf{XX} \qquad \ddot{y}1 \quad \mathbf{XY} \quad .$$

Betrachten Sie die standardisierte Datenmatrix :

$$\mathbf{W} = \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} \begin{matrix} Xs_{11} & Xs_{12} & \cdots\ Xs\ 1p\ddot{y}1 \\ Xs_{21} & Xs_{22} & \cdots\ Xs\ 2p\ddot{y}1 \\ \vdots & \vdots & \ldots \quad \vdots \\ s\ n1 & Xs_{n2} & \cdots\ Xs\ np\ddot{y}1 \end{matrix} \begin{matrix} \ddot{y} \\ \ddot{y} \\ \ddot{y} \\ \ddot{y} \end{matrix} .$$

Machine Translated by Google

Daher wird das Modell zu:

$$\mathbf{Y} = \ddot{y} + \mathbf{W}\ddot{y} + \ddot{y} \quad ,$$

Wo

$$\ddot{y} = (\ddot{y}0,\ldots, \ddot{y}0) = \ddot{y}s_{0,\ldots,\ddot{y}s0} \quad \text{Und}$$

$$\ddot{y} = \ddot{y}1,\ldots, \ddot{y}p\ddot{y}1 = \ddot{y}s_{1,\ldots,\ddot{y}s\ddot{p}y1} .$$

Die Schätzungen der kleinsten Quadrate sind durch $\ddot{y}0 = \ddot{y}$ gegeben $\overset{s}{0} = \overline{Y}$ und durch:

$$\ddot{y} = \ddot{y}1,\ldots, \ddot{y}p\ddot{y}1$$

$$= \ddot{y}\ {}^{s}_{1,\ldots,}\ \ddot{y}^{s}_{p\ddot{y}1}$$

$$= (\mathbf{W}\ W)\ \ddot{y}1\mathbf{W}\ \mathbf{Y} \quad .$$

Diese Schätzungen sind unvoreingenommen, das heißt:

$$E\ \ddot{y}\ j = \ddot{y}j\ ,\ j = 0,\ldots,p\ \ddot{y}\ 1\ E\ \ddot{y}j = \ddot{y}j\ ,\ j \quad ,$$

$$= 0,\ldots,p\ \ddot{y}\ 1 \quad .$$

Unter allen möglichen linearen unverzerrten Schätzungen sind es die Schätzungen mit der minimalen **Varianz.** Wir wissen das

$$\text{Var}\ \ddot{\mathbf{y}} = \ddot{y}2\ \mathbf{XX}^{\ddot{y}1} \quad ,$$

$$\text{Var}(\ddot{y}\ ) = \ddot{y}2\ \mathbf{W}\ W\ddot{y}1 \quad ,$$

wobei $\ddot{y}2$ die Fehlervarianz ist. Beachten Sie, dass **WW** gleich dem $(n\ \ddot{y}\ 1)$-fachen der Korrelationsmatrix für die unabhängigen Variablen ist. Folglich entsteht im günstigsten Fall keine Kollinearität, da **WW** ein Vielfaches von I ist. Um eine Matrix **WW** möglichst nahe zu haben Im günstigsten Fall ersetzen wir **WW** durch **WW** + **kI,** wobei k ein positiver Skalar ist.

Die Ridge-Regression ersetzt die Schätzungen der kleinsten Quadrate

$$\ddot{\mathbf{y}} = \mathbf{W}\ W\ddot{y}1\ \mathbf{WY}$$

durch die Ridge-Schätzungen

$$\ddot{y}R = \mathbf{WW} + \mathbf{kI}^{\ddot{y}1}\ \mathbf{WY}$$

$$= \ddot{y}R1\ ,\ldots,\ \ddot{y}R(p\ddot{y}1) .$$

Es gibt immer einen **Wert** für k , für den der **gesamte mittlere quadratische Fehler** der Ridge-Schätzungen kleiner ist als der gesamte mittlere quadratische Fehler der Kleinste-Quadrate-Schätzungen. 1975 $(p\ddot{y}1)\ddot{y}2$ Hoerl et al. schlug vor, dass $k = \frac{\text{haben}}{\ddot{y}\ \ddot{y}}$ einen niedrigeren mittleren quadratischen Gesamtfehler ergibt, wobei $(p\ \ddot{y}\ 1)$ die Anzahl der erklärenden Variablen ist. Um k zu berechnen, müssen wir ŷ und $\ddot{\mathbf{y}}$ durch s und $\ddot{\mathbf{y}}$ schätzen . Dann Verwenden der Ridge-Regression mit $\frac{\text{wird}\ \ddot{\mathbf{y}}\ \text{durch}\ \ddot{y}R\ (p\ddot{y}1)\ddot{y}}{\ddot{\mathbf{y}}\ \ddot{y}}$. $k =$ Beachten Sie jedoch, dass das verwendete k eine Zufallsvariable ist, da es von den Daten abhängt.

DOMÄNEN UND EINSCHRÄNKUNGEN

Eigenschaften von Ridge-Schätzungen

Ridge-Schätzungen sind verzerrt, da die erwarteten Werte $E\ \ddot{y}Rj$ der Schätzungen $\ddot{y}Rj$ in liegen der **Vektor:**

$$E\ \ddot{y}Rj = \mathbf{WW} + \mathbf{kI}^{\ddot{y}1}\ \mathbf{W}\ W\ddot{\mathbf{y}} = \ddot{\mathbf{y}}\ ,$$

$$\ddot{y}k > 0 .$$

Beachten Sie, dass der Wert $k = 0$ den Schätzungen der kleinsten Quadrate entspricht. Ridge schätzt, dass einige Verzerrungen berücksichtigt werden, um die Varianz zu verringern ance der Schätzungen. Wir haben:

$$\text{Var}(\ddot{y}R) = \ddot{\mathbf{y}}\mathbf{2VR}\ ,$$

mit:

$$VR = \mathbf{WW} + \mathbf{kI}^{\ddot{y}1}\ \mathbf{WWWW} + \mathbf{kI}^{\ddot{y}1} .$$

Beachten Sie, dass die Diagonalelemente der Matrix VR kleiner sind als die Diagonalelemente, die **W** $W\ddot{y}1$ entsprechen , was diesen Grat impliziert Schätzungen weisen eine geringere Varianz auf als die geringsten Quadratschätzungen.

Daher sind Ridge-Schätzungen weniger variabel als Schätzungen nach der Methode der kleinsten Quadrate. Ist ihre Voreingenommenheit akzeptabel? Um einen Kompromiss zwischen den zu messe

R

Machine Translated by Google

Bias und Varianz einer Schätzung berechnen wir im Allgemeinen ihren mittleren quadratischen Fehler, bezeichnet vonMSE. Für eine Schätzung ÿ von ÿ ist dies definiert als:

$$MSE\ \ddot{y} = E\ \ddot{y}\ \ddot{y}\ \ddot{y}^2$$

$$= \text{Var}\ \ddot{y} + E\ \ddot{y}\ \ddot{y}\ \ddot{y}^2 .$$

Entsprechend definieren wir den Gesamtmittelwert quadrierter Fehler eines geschätzten Vektors um die Summe der mittleren quadratischen Fehler davon Komponenten. Wir haben:

$$TMSE\ (\ddot{y}R) = \sum_{j=1}^{p\ddot{y}1} MSE\ \ddot{y}Rj$$

$$= \sum_{j=1}^{p\ddot{y}1} E\ \ddot{y}Rj\ \ddot{y}\ \ddot{y}j^2$$

$$= \sum_{j=1}^{p\ddot{y}1} \text{Var}\ \ddot{y}Rj + E\ \ddot{y}Rj\ \ddot{y}\ \ddot{y}j^2$$

$$= (p\ \ddot{y}\ 1)\ \ddot{y}2\text{Trace}\ (VR) + \sum_{j=1}^{p\ddot{y}1} E\ \ddot{y}Rj\ \ddot{y}\ \ddot{y}j^2 .$$

Diese Methode wird als adaptive Ridge-Regression bezeichnet. Die Eigenschaften des Grates werden geschätzt gilt nur, wenn *k* eine Konstante ist, was nicht der Fall ist Dies gilt für die adaptive Ridge-Regression. Nichtsdestotrotz deuten die Eigenschaften der Kammschätzungen auf adaptive Kammschätzungen hin sind gute Schätzungen. Tatsächlich haben statistische Studien gezeigt, dass für viele Datensätze ein Die adaptive Ridge-Schätzung ist besser mittlerer quadratischer Fehler als die Schätzung der kleinsten Quadrate.

BEISPIELE

Im folgenden Beispiel werden 13 Proben von Zement wurden abgebunden. Für jeden ist der pro Prozent der vier chemischen Inhaltsstoffe wurde gemessen und ist unten tabellarisch aufgeführt. Der Außerdem wurde die abgegebene Wärmemenge gemessen.

Das Ziel bestand darin, zu definieren, wie die Mengen xi1, xi2, xi3 und *xi4* beeinflussen *yi,* die Wärmemenge ausgesendet.

Tabelle: Abgegebene Wärme

Proben	Zutat				Hitze
	1234				
ich	Xi 1	Xi 2	Xi 3	Xi 4	Yi
1	7	26	6	60	78,5
2	1	29	15	52	74,3
3	11	56	8	20	104.3
4	11	31	8	47	87,6
5	7	52	6	33	95,9
6	11	55	9	22	109.2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93.1
10	21	47	4	26	115,9
11	1	40	23	34	83,9
12	11	66	9	12	113.3
13	10 68		8	12	109.4

wobei *Yi* die von der Probe i entwickelte Wärme ist , *Xi1* ist die Menge der Zutat 1 in Probe i, *Xi2* ist die Menge der Zutat 2 in Probe i, *Xi3* ist die Menge der Zutat 3 in Probe i, *undXi4* ist die Menge der Zutat 4 in Probe i.

Wir führen zunächst eine multiplelineare Regression durch. Das lineare Regressionsmodell ist:

$$yi = \ddot{y}0 + \ddot{y}1Xi1 + \ddot{y}2Xi2 + \ddot{y}3Xi3 + \ddot{y}4Xi4 + \ddot{y}i .$$

Wir erhalten folgende Ergebnisse:

Machine Translated by Google

Variable	Koeffizient Std. Entwickler		tc
Konstante	62.4100 70.0700		0,89
X1	1.5511	0,7448	2.08
X2	0,5102	0,7238	0,70
X3	0,1019	0,7547	0,14
X4	ÿ0,1441	0,7091 *ÿ0,20*	

Der durch **standardisierte Daten** ermittelte geschätzte Vektor ist:

$$\ddot{y} = (9{,}12;\ 7{,}94;\ 0{,}65;\ \ddot{y}2{,}41)\ ,$$

mit einer Standardabweichung von 4,381, 11,26, 4,834 bzw. 11,87. Die Schätzung für *s* von ÿ ist *s* = 2,446. Wir folgen einem Bergrücken Analyse mit:

$$k = \frac{(p\ \ddot{y}\ 1)\ \ddot{y}2}{\ddot{y}\ \ddot{y}} = \frac{4 \cdot (2.446)^{2}}{(9.12)\ 2{+}(7.94)\ 2{+}(0.65)\ 2{+}(2.41)\ 2} = 0{,}157\ .$$

Dies ist eine adaptive Ridge-Regression seit *k* hängt von den Daten ab. Wir verwenden jedoch den Begriff Gratregression. Wir erhalten:

$$YR = (7{,}64;\ 4{,}67;\ \ddot{y}0{,}91;\ \ddot{y}5{,}84)\ ,$$

mit einer Standardabweichung von 1,189, 1,377, 1,169 bzw. 1,391. Wir haben das geschätzte Werte:

$$Yi = 95{,}4 + 7{,}64Xs_{i1} + 4{,}67Xs_{i2} \ddot{y}\ 0{,}91Xs_{i3} \ddot{y}\ 5{,}84Xs_{i4}\ .$$

Der geschätzte Wert für die Konstante gilt nicht weichen von der Schätzung der kleinsten Quadrate ab.

WEITERLESEN

- Kollinearität
- Mittlere quadratische Fehler
- Simulation
- Standardisierte Daten

VERWEISE

HoerlA.E.:Anwendung von RidgeAnalysisto Regressionsprobleme. Chem. Ing. Prog. 58(3), 54–59 (1962)

Hoerl AE, Kennard RW: Ridge Regression: Bias Estimation for Nonorthogo nal Problems. Technometrie **12,** 55–67 (1970)

Hoerl AE, Kennard RW, Baldwin KF: Ridge-Regression: einige Simulationen. Komm. Stat., Bd. 4, S. 105–123 (1975)

Risiko

In der Epidemiologie entspricht Risiko dem Wahrscheinlichkeit, dass eine Krankheit in einer Bevölkerung in einem bestimmten Zeitintervall auftritt. Das Risiko ist für beide Populationen gleichermaßen definiert oder eine homogene Untergruppe von Individuen. Die Homogenität wird hier in dem Sinne verwendet dass Personen einem bestimmten Faktor ausgesetzt sind, der im Verdacht steht, die Krankheit zu verursachen studiert. Der Begriff des Risikos steht in engem Zusammenhang mit dem Begriff der **Inzidenz** (die Inzidenz entspricht der Häufigkeit einer Krankheit). Risiko zu seiner Wahrscheinlichkeit).

R

GESCHICHTE

Im Jahr 1951 stellte Cornfield, Jerome, klar fest zum ersten Mal den Begriff des Risikos. Das ergab sich aus einer Folgestudie, in der das Risiko bestand Faktoren wurden für zwei Gruppen verglichen, eine bestehend aus kranken Individuen und dem anderen von gesunden Menschen. Seitdem besteht Risiko wurde in epidemiologischen Studien häufig verwendet ein Maß für den Zusammenhang zwischen einer Krankheit und die damit verbundenen Risikofaktoren.

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Das Risiko ist die Wahrscheinlichkeit der Erkrankung
in einem festgelegten Zeitraum in einer Gruppe erscheinen
von Personen, die einem Risikofaktor ausgesetzt sind
lernen. Wir schätzen das Risiko anhand des Verhältnisses:

$$\text{Risiko} = \frac{\text{Vorfallfälle (in der exponierten Bevölkerung)}}{\text{Bevölkerung gefährdet}}$$

im angegebenen Zeitraum. Der Nominator entspricht der
Zahl der Neuerkrankungen
der Krankheit während des festgelegten Zeitraums in
die untersuchte Bevölkerung, während der Nenner der Anzahl
der Personen entspricht
Risiko im gleichen Zeitraum.

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Risikobegriff ist intuitiv ein Wahrscheinlichkeitsmaß, das
für eine gegebene Situation definiert wird
Bevölkerung und einem bestimmten Zeitraum.
Versicherungsunternehmen nutzen es, um das vorherzusagen
Häufigkeit von Ereignissen wie Tod, Krankheit,
usw. in einer definierten Population und nach Faktoren wie
Geschlecht, Alter, Gewicht usw.
Risiko kann als Wahrscheinlichkeit interpretiert werden
dass eine Person an einer Krankheit erkrankt, vorausgesetzt,
sie war einer bestimmten Krankheit ausgesetzt
Risikofaktor. Wir schreiben dem Einzelnen das zu
Mittelwert des Risikos für die einzelnen Personen
seiner Gruppe. Versicherungsunternehmen nutzen Risiko
für jede Bevölkerung, um den Beitrag jedes einzelnen
Versicherten zu ermitteln. Ärzte
Verwenden Sie das Risiko, um die Wahrscheinlichkeit abzuschätzen, dass einer
einer ihrer Patienten entwickelt eine bestimmte Krankheit,
eine Komplikation hat oder sich von der Krankheit erholt
Leichtigkeit.

Es ist wichtig zu beachten, dass der Begriff des Risikos
ist immer für einen bestimmten Zeitraum definiert:
das Risiko in 2, 3 oder 10 Jahren.
Das minimale Risiko beträgt 0 %, das maximale Risiko
100%.

BEISPIELE

Nehmen wir an, dass am 1. Januar 1992 eine Gruppe von
1000 60-Jährige werden ausgewählt und 10 Fälle von
Herzinfarkten werden während dieser Zeit beobachtet
Jahr. Das Herzinfarktrisiko für diese Bevölkerungsgruppe
der 60-jährigen Männer entspricht 10/1000 = 1 %
in einem Jahr.
Wir betrachten hier einen Fall von Brustkrebs
im Laufe von 2 Jahren in einer Population von
100.000 Frauen, verteilt auf drei Gruppen
Abhängig von der Höhe der Exposition gegenüber dem
Faktor „Rauch“. Der folgende Notfall
Die Tabelle veranschaulicht dieses Beispiel:

Gruppe	Anzahl Fälle (Vorfälle)	Anzahl Einzelpersonen
Nicht belichtet	40	35100
Passivraucher	140	55500
Aktive Raucher	164	59400
Gesamt	344	150000

Das Risiko für jede Gruppe wird berechnet als:
das Verhältnis der Anzahl der geteilten Fälle
durch die Gesamtzahl der Individuen in jedem
Gruppe. Zum Beispiel das Risiko für das Passive
Raucher beträgt 140/55500 = 0,002523 =
252,3/100000 in 2 Jahren. Wir haben die folgende Tabelle:

Gruppennummer von	Fälle (Vorfälle)	Nummer von Indien Bilder	Risiko in zwei Teile Jahre (100000)
Nonex gestellt	40	35100	114,0
Passiv Raucher	140	55500	252,3
Aktiv Raucher	164	59400	276,1
Gesamt	344	150000	229,3

Der Begriff des Risikos impliziert den Begriff der Vorhersage:
Unter den Frauen einer bestimmten Bevölkerung

Machine Translated by Google

Welcher Anteil davon wird sich entwickeln?
Brustkrebs in den nächsten 2 Jahren? Von dem
In der obigen Tabelle stellen wir fest, dass 229,3/100000 möglich sind
diese Krankheit in den nächsten 2 Jahren entwickeln.

WEITERLESEN

- Zurechenbares Risiko
- Vermeidbares Risiko
- Ursache und Wirkung in der Epidemiologie
- Inzidenzrate
- Quoten und Quotenverhältnis
- Prävalenzrate
- Relatives Risiko

VERWEISE

Cornfield, J.: Eine Methode zur Schätzung von Vergleichsraten anhand klinischer Daten. Anwendungen bei Lungen-, Brust- und Krebserkrankungen Gebärmutterhals. J. Natl. Krebsinst. **11,** 1269–75 (1951)

Lilienfeld, AM, Lilienfeld, DE: Grundlagen der Epidemiologie, 2. Aufl. Claren Don, Oxford (1980)

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Kleiner Brauner, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996)

Morabia, A.: L'Épidémiologie Clinique. Ausgaben „Que sais-je?“. Presses Univer Sitaires de France, Paris (1996)

Robuste Schätzung

Eine **Schätzung** heißt robust, wenn die **Schätzung**
Die verwendeten **Materialien** sind unempfindlich
Ausreißer und kleine Abweichungen von idealisierten
(Modell-)Annahmen. Zu diesen Schätzmethoden gehört
die Maximum-Likelihood-Methode
Typschätzung, bekannt als M-Estimates, lineare
Kombination von Ordnungsstatistiken, bekannt als
L-Schätzungen und auf statistischen Rängen basierende
Methoden, R-Schätzungen. Ausreißer sind wahrscheinlich
Fehler, die auf eine schlechte Anzeige zurückzuführen sind.
eine schlechte Aufnahme oder eine andere Ursache
in die experimentelle Umgebung. Das **arithmetische**
Mittel und andere Schätzer des **kleinsten Wertes**
Quadrate reagieren sehr empfindlich auf solche Ausreißer.
Deshalb ist es in diesen Fällen vorzuziehen
greifen auf robustere Schätzer zurück.

GESCHICHTE

Der Begriff robust, ebenso wie das Feld dazu
mit dem es zusammenhängt, wurde 1953 eingeführt
von Box, GEP Der Begriff wurde erkannt als
erst Mitte der 1960er Jahre ein Bereich der Statistik.
Dennoch ist es wichtig, dies zu beachten
Konzept ist nicht neu; im späten 19. Jahrhundert,
Mehrere Wissenschaftler hatten bereits eine klare Vorstellung davon
dieses Konzept. Tatsächlich die erste mathematische
Arbeite offenbar an robusten Schätzungen der Daten
zurück bis 1818 mit **Laplace, PS** in seinem Werk
Die zweite Ergänzung zur Theorie und zur
Wahrscheinlichkeitsanalyse . Die Verteilung von
Der **Median** ist dieser Arbeit zu entnehmen.
Es ist aus dem Jahr 1964, mit dem Artikel von Huber, PJ,
dass dieser Bereich ein separater Bereich von wurde
Statistiken , die als robuste Statistiken bekannt sind.

MATHEMATISCHE ASPEKTE

Betrachten Sie das folgende **Modell:**

$$\mathbf{Y} = \mathbf{X} \cdot \ddot{\mathbf{y}} + \ddot{\mathbf{y}} \quad ,$$

Dabei ist **Y** der $(n \times 1)$-Vektor der Beobachtungen relativ
zur **abhängigen Variablen**
(n Beobachtungen), $\ddot{\mathbf{y}}$ ist der $(p \times 1)$-Vektor von
die zu schätzenden Parameter und $\ddot{\mathbf{y}}$ ist der
$(n \times 1)$ Fehlervektor.
Es gibt verschiedene Methoden zur robusten Schätzung
mation, die es erlaubt, den **Vektor** abzuschätzen
der **Parameter** ÿ.

R

Machine Translated by Google

Einige der Methoden sind:

• L1- **Schätzung:**

$$\text{Minimieren} \sum_{i=1} |ÿi| .$$

• Allgemeiner ausgedrückt wird das Konzept der M-Schätzer, das auf der Idee basiert, das Fehlerquadrat zu ersetzen **(Methode der kleinsten Quadrate),** durch eine Funktion *ÿ(i)* definiert , wobei ÿ eine symmetrische Funktion mit einem eindeutigen Minimum in Null ist :

$$\text{Minimieren} \sum_{i=1}^{N} ÿ\ (ÿi) .$$

Im Jahr 1964 schlug Huber, PJ die folgende Funktion vor:

$$ÿ\ (x) = \begin{cases} ÿ\ \dfrac{x2}{2}, & \text{wenn } |x| ÿ\ k \\ k \cdot |x| ÿ\ \dfrac{k2}{2}, & \text{wenn } |x| > k \end{cases},$$

wobei *k* eine positive Konstante ist.
Beachten Sie, dass für *ÿ(x)* = |x| die L1 -Schätzer erhalten werden.

WEITERLESEN

Einschätzung

Schätzer

L1 -Schätzung

Maximale Wahrscheinlichkeit

Ausreißer

VERWEISE

Box, GEP: Nichtnormalität und Varianztests. Biometrie **40,** 318–335 (1953)

Huber, PJ: Robuste Schätzung eines Standortparameters. Ann. Mathematik. Stat. **35,** 73–101 (1964)

Huber, PJ: Robuste Statistik. Wiley, New York (1981)

Huber, PJ: Robustheit: Wo stehen wir jetzt? Student, Bd. 1, nein. 2, S. 75–86 (1995)

Laplace, PS de: Zweite Ergänzung zur analytischen Theorie der Wahrscheinlichkeiten. Courcier, Paris (1818). S. 531–580 im vollständigen Werk von Laplace, Bd. 7. Gauthier-Villars, Paris, 1886

Machine Translated by Google

Probe

Eine Stichprobe ist eine Teilmenge einer **Population ,** an der statistische Studien durchgeführt werden Rückschlüsse auf die **Bevölkerung ziehen.**

BEISPIELE

Eine Hausfrau bereitet eine Suppe zu. Um festzustellen, ob die Suppe genug Salz enthält, probiert sie einen Löffel davon. Abhängig von dem, was sie an der hier durch einen Löffel dargestellten Probe beobachtet hat, kann sie entscheiden, ob es notwendig ist, der Suppe Salz hinzuzufügen, was die untersuchte **Population symbolisiert.**

WEITERLESEN

- **Tschebyscheffs Ungleichung Einschätzung**
- **Bevölkerung**
- **Probengröße**
- **Probenahme**

Probengröße

Die Stichprobengröße ist die Anzahl der zu einer **Stichprobe gehörenden Personen oder Elemente.**

GESCHICHTE

Nordin, JA (1944) behandelt das Problem der **Schätzung** der Stichprobengröße im Rahmen einer Entscheidungstheorie durch Ein relatives Beispiel für die potenziellen Verkäufe in einem Markt, in dem ein Verkäufer präsent sein möchte. Cornfield, J. (1951) veranschaulicht auf präzise Weise die Schätzung der Stichprobengröße zur Bildung von Klassen mit bekannten Anteilen. Sittig, J. (1951) erörtert die Wahl der Stichprobengröße für ein typisches wirtschaftliches Problem unter Berücksichtigung der Inspektion und der Kosten, die für fehlerhafte Teile in einer akzeptierten Charge und für korrekte Teile in einer abgelehnten Charge anfallen. Zahlreiche Studien wurden auch von Cox, DR (1952) durchgeführt, der sich von den Arbeiten von Stein, C. (1945) inspirieren ließ, um die Stichprobengröße optimal zu bestimmen.

MATHEMATISCHE ASPEKTE

Sei *X1,...,Xn* eine ersatzlose Zufallsstichprobe aus einer **Population von** N *Beobachtungen* mit einem unbekannten **Mittelwert** ÿ.
Wir gehen davon aus, dass die Stichprobenverteilung der Statistik einer Normalverteilung folgt.

Die **Tschebyscheff-Ungleichung** hängt von der Anzahl *n* der Beobachtungen ab und ermöglicht dies

Schätzen Sie den Mittelwert ÿ anhand des empirischen Mittelwerts der erhaltenen Werte *Xn.*

Wenn *N* groß genug ist, kann die Größe *n* der Stichprobe mithilfe der Tcheby-Chev-Ungleichung bestimmt werden, sodass die Wahrscheinlichkeit, dass $\overline{Xn}$ im Intervall [ÿ ÿ *ÿ,* ÿ + ÿ] (ÿ > 0) enthalten wäre, auf den erforderlichen Wert festgelegt wird Präzision) kann so groß sein, wie wir wollen. Wenn wir anrufen

Machine Translated by Google

ÿ das **Signifikanzniveau,** wir erhalten:

$$P\,(|x^{-}\ \ddot{y}\ \ddot{y}|\ \ddot{y}\ \ddot{y}) = \ddot{y} \text{ oder}$$

$$P\,(|x^{-}\ \ddot{y}\ \ddot{y}| < \ddot{y})\ = 1\ \ddot{y}\ \ddot{y} \quad ,$$

was wir auch schreiben können als:

$$P\ \ddot{y}z\ddot{y}/2\ \ddot{y} \quad \frac{x^{-}\ \ddot{y}\ \ddot{y}}{\ddot{y}x^{-}} \quad \ddot{y}\ z\ddot{y}/2 = 1\ \ddot{y}\ \ddot{y} \quad ,$$

wobei ÿx¯ der **Standardfehler** des Mittelwerts ist, wird durch ÿÿ $\frac{}{n}$, wobei ÿ der **Standard** ist **Abweichung** der Grundgesamtheit und zÿ/2den mit dem **Konfi** verbundenen kritischen Standardwert geschätzt **Denzniveau** 1 ÿ ÿ.

Wir konstruieren das **Konfidenzintervall** mit die Hilfe der **Normaltabelle** für das Konfidenzniveau 1 ÿ ÿ:

$$2 \cdot \ddot{y}n\ \ddot{y}n \quad \frac{\ddot{y}}{} \quad \ddot{y}\ {}^{-}x\ \ddot{y}\ \ddot{y}\ \ddot{y}\ z\ddot{y}/2 \cdot \ddot{y}z\ddot{y}/ \quad \frac{\ddot{y}}{} \cdot$$

Da die **Normalverteilung** symmetrisch ist, Uns interessiert nur die Gleichheit auf der Rechts. Es zeigt an, *dass zÿ/2* · ÿÿ $\frac{}{n}$ ist der Beste Wert, den x¯ ÿ ÿ annehmen kann. Diese Grenze gilt auch gegeben durch die gewünschte Präzision. Wovon:

$$\ddot{y} = z\ddot{y}/2 \cdot \ddot{y}n \quad \frac{\ddot{y}}{} \quad ,$$

das ist:

$$n = z\ddot{y}/2 \cdot \quad \frac{\ddot{y}}{\ddot{y}} \quad ^{2} \quad .$$

Wenn die gewünschte Präzision ausgedrückt werden kann in **Prozentsatz** relativ zum Mittelwertÿ, oder k%, wir dann habe:

$$\ddot{y} = k \cdot \ddot{y}$$

$$\text{und } n = \quad \frac{z\ddot{y}}{k} \cdot \frac{\ddot{y}}{\ddot{y}} \quad ^{2} \quad ,$$

Wo $\frac{\ddot{y}}{\ddot{y}}$ wird als **Varia-Koeffizient** bezeichnet **tion.** Es misst die relative Streuung von die **untersuchte Variable.** In Fällen, in denen **der Sam**

Da **die Pling-Verteilung** nicht normalverteilt ist, sind diese Formeln näherungsweise gültig wenn die Stichprobengröße n viel größer ist als *(n* ÿ 30).

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Bestimmung der Stichprobengröße erfolgt ein wichtiger Schritt für die Organisation einer statistischen Untersuchung. Es beinhaltet nicht nur die komplexen Berechnungen, sondern auch die Berücksichtigung Berücksichtigung der unterschiedlichen Anforderungen der Untersuchung selbst. Somit ist die gewünschte Präzision von Die Untersuchungsergebnisse sind zu berücksichtigen. Generell gilt: Je größer die Stichprobe, desto größer die Präzision. Die Kosten der Untersuchung stellen eine weitere wichtige Einschränkung dar. Der Je kleiner das Budget, desto eingeschränkter Stichprobengröße. Darüber hinaus ist die Verfügbarkeit von andere Ressourcen, wie etwa die Existenz von Daten aus einer **Volkszählung** oder einem System von stammend Anfragenden oder die Dauer der Anfrage, wie sowie das abzudeckende Gebiet, haben können eine dramatische Auswirkung auf die Struktur der Probe.

Die Volkszählungstheorie legt großen Wert auf Insbesondere Proben mit „optimaler“ Dimension beim Optimieren (Maximieren oder Minimieren) einer Zielfunktion, bei der die Prinzipien mathematisch formuliert sind. Darüber hinaus ist es wichtig zu betonen, dass dies bei der optimalen Stichprobendimension im Allgemeinen nicht der Fall ist hängen (oder nur geringfügig) von der Größe ab der ursprünglichen **Bevölkerung.**

Auch im Bereich der Versuchsplanung stellt sich das Problem der optimalen Stichprobengröße. In diesem Fall sollte man das Minimale finden Anzahl der zu berücksichtigenden Beobachtungen um einen signifikanten Unterschied zu erkennen (falls vorhanden). eins) zwischen den verschiedenen Behandlungen wie z ebenso viele Medikamente oder Herstellungsprozesse.

BEISPIELE

Eine Fabrik beschäftigt mehrere tausend Arbeiter. Basierend auf einem früheren Experiment, dem Der Forscher weiß, dass wöchentliche Gehälter gelten normalverteilt mit einer **Standardabweichung** von 40 CHF. Der Forscher möchte abschätzen,

Machine Translated by Google

Ermitteln Sie den mittleren Wochenlohn mit einer Genauigkeit von 10 CHF und einem **Konfidenzniveau** von 99 %.

Wie groß sollte die Stichprobengröße sein?

Für ein Konfidenzniveau von 99 % (ÿ = 0,01 = 1 ÿ 0,99) beträgt der *z0,01*- **Wert ,** den wir in der **Normaltabelle** finden , 2,575.

Die direkte Anwendung der Formel ergibt uns:

$$n \ddot{y} z\ddot{y} \cdot \frac{\ddot{y}}{\ddot{y}}^2 ,$$

$$n \ddot{y} 2{,}575 \cdot \frac{}{40}\,102 ,$$

$$n \ddot{y} 106{,}09 .$$

Wir erreichen die gewünschte Präzision, wenn die Stichprobengröße größer oder gleich 107 Arbeiten ist ers.

WEITERLESEN

Bevölkerung

Probe

VERWEISE

Cochran, WG: Sampling Techniques, 2. Aufl. Wiley, New York (1963)

Cornfield, J.: Methoden zur Stichprobenziehung menschlicher Populationen. Die Bestimmung der Stichprobengröße. Amer. J. Publ. Gesundheit **41,** 654–661 (1951)

Cox, DR: Schätzung durch Doppelstichprobe. Biometrie **39,** 217–227 (1952)

Harnett, DL, Murphy, JL: Einführung in die statistische Analyse. Addison-Wesley, Reading, MA (1975)

Nordin, JA: Bestimmung der Stichprobengröße. Marmelade. Stat. Assoc. **39,** 497–506 (1944)

Sittig, J.: Die wirtschaftliche Wahl des Probenahmesystems bei der Annahmeprobenahme. Stier. Int. Stat. Inst. 33(V), 51–84 (1952)

Stein, C.: Ein Zwei-Stichproben-Test für eine lineare Hypothese, deren Potenz unabhängig von ist die Varianz. Ann. Mathematik. Stat. **16,** 243–258 (1945)

Probenraum

Der Stichprobenraum eines **Zufallsexperiments** ist die Menge aller möglichen Ergebnisse des **Experiments und** wird normalerweise mit bezeichnet.

GESCHICHTE

Siehe **Wahrscheinlichkeit.**

MATHEMATISCHE ASPEKTE

Ein Beispielraum kann sein:

• Endlich (die Ergebnisse sind zählbar); •

Unendlich zählbar (es ist möglich, jedes mögliche Ergebnis einer positiven ganzen Zahl zuzuordnen, sodass jedes mögliche Ergebnis eine andere Zahl hat; die Anzahl der möglichen Ergebnisse ist jedoch unendlich);

• Unendlich nicht zählbar (es ist unmöglich, alle möglichen Ergebnisse aufzuzählen).

Der Probenraum eines **Zufallsexperiments** wird auch als sicheres **Ereignis bezeichnet.** Die **Wahrscheinlichkeit** des sicheren Ereignisses ist gleich 1:

$$P() = 1 \quad .$$

BEISPIELE

Endlicher Probenraum:

- Eine Münze werfen

= {Kopf, Zahl}

- Einen Würfel werfen

= {1, *2, 3, 4, 5,* 6}

Unendlich abzählbarer Probenraum:

– So oft wie nötig eine Münze werfen, um „Kopf“ zu erhalten *(H* = Kopf und *T* = Zahl)

= *{H, TH, TTH, TTTH, TTTTH, TTTTTH, TTTTTTH,...}* .

S

Machine Translated by Google

Unendlicher nichtabzählbarer Probenraum:

– Ort der **Koordinaten** des Aufpralls
einer Kugel auf einer Platte von 10 × 10 cm. In
In diesem Fall ist der Probenraum zusammengesetzt
einer unendlichen, nicht abzählbaren Menge von Farben
dinates, weil die Koordinaten mea sind
auf einer **Skala** reeller Zahlen bestimmt (es ist
Es ist unmöglich, eine erschöpfende Liste davon zu geben
mögliche Ergebnisse).

WEITERLESEN

Fall

Wahrscheinlichkeit

Zufälliges Experiment

VERWEISE

Bailey, DE: Wahrscheinlichkeit und Statistik: Modelle
für die Forschung. Wiley, New York (1971)

Probenahme

Im Allgemeinen bedeutet der Begriff „Stichprobe“ die Auswahl eines Teils einer **Population,** die sogenannte **Stichprobe,** die Untersuchung bestimmter Merkmale *x* dieser Stichprobe und die Tatsache **Rückschlüsse** auf die Bevölkerung zu ziehen . Die Stichprobe stellt auch die Reihe von Operationen dar, die erforderlich sind, um **eine Stichprobe** aus einer bestimmten **Grundgesamtheit zu erhalten.**

GESCHICHTE

Das Konzept der Repräsentativität einer **Stichprobe** ist sehr neu.
Auch wenn die ersten Versuche zur **Extrapolation** erst im 13. Jahrhundert aufkamen
in Frankreich, wo die Zahl der Wohnungen lag
Zur Schätzung der **Bevölkerungszahl** dient die **Volkszählung** blieb bis dahin beliebter als Sampling
das 19. Jahrhundert.
Das Prinzip der Probenahme mit oder ohne
Ersatz erschien zum ersten Mal in
das Werk „De ratiociniis in Aleae Ludo“,
veröffentlicht im Jahr 1657 vom niederländischen
Wissenschaftler Huy gens, Christiaan (1629–1695).
Es war im Jahr 1895, in Bern, während des Kongresses
des Internationalen Instituts für Statistik,
dass Kiaer, Anders Nicolai ein Memo mit dem Titel
„Observations et expériences con cernant des
dénombrements représentatifs“ vorlegte,
Kiaer, als Direktor des Zentralbüros von
Statistiken des Königreichs Norwegen am
Zeit, verglich in seiner Präsentation die Struktur einer
Stichprobe mit der Struktur einer **Bevölkerung,** die
durch **eine Volkszählung ermittelt wurde.** Er war also
Definition des Begriffs Kontrolle a posteriori und für
Zum ersten Mal wurde der Begriff „Vertreter“ verwendet .
Kiaer, AN stieß auf den Widerstand von
fast alle Kongressteilnehmer, die dabei waren
überzeugt, dass die vollständige Aufzählung war
der einzige Weg, weiterzumachen.
Dennoch bestätigte Kiaer, AN 1897 den Begriff der
Repräsentativität und
erhielt die Unterstützung der Konferenz von
Skandinavische Statistiker. Ihm wurde gegeben
die Möglichkeit, seinen Standpunkt zu verteidigen
während mehrerer Konferenzen: St.-Petersburg
(1897), Stockholm (1899), Budapest (1901),
und Berlin (1903).
Mit den Werken von wurde ein neuer Schritt getan
März, Lucien über die Rolle des Zufalls
während der Probenahme; er war der erste, der sich entwickelte
die Idee des probabilistischen Samplings, auch genannt
Stichproben.
Andere Statistiker haben sich daran beteiligt
Problem. Von Bortkiewicz, L., ein Professor aus Berlin,
schlug die Berechnung von **Wahrscheinlichkeiten**
vor , um die Abweichung zu testen
zwischen den Verteilungen einer **Stichprobe** und
die Gesamtpopulation **anhand** der Schlüsselvariablen.
Bowley, AL (1906) entwickelte einen Zufall
Stichprobenverfahren namens Stratifizierung.
Laut Desrosières, A. (1988) war er es
Ich interessiere mich auch für den Begriff des **Vertrauens**

Machine Translated by Google

Die ersten **Berechnungen** legte er 1906 dem Royal Statistical vor
Gesellschaft.
Das Jahr 1925 markierte einen neuen Schritt in der Geschichte des Samplings. Es ist tatsächlich das Jahr des Kongress des Internationalen Instituts für Statistik in Rom, bei dem Zufallsstichproben (choicedonebychance) vorgestellt wurden
und gezielte Probenahme (vernünftige Wahl).
Von nun an würde das Problem nie mehr auftreten noch einmal in Bezug auf die Wahl zwischen **Probenahme** und vollständige Aufzählung, aber es kam es in Bezug auf verschiedene Stichproben zu Abweichungen Methoden.

Von den verschiedenen Einsatzmöglichkeiten des Samplings die meisten in den Vereinigten Staaten häufig verwendet wurde
diejenige, die von der verschiedenen öffentlichen Meinung gemacht wird Institute und insbesondere vorwahlbezogene Untersuchungen.
Der dritte Tag im November 1936, der Tag
Die Ergebnisse der Präsidentschaftswahlen waren Veröffentlichung ist einer der wichtigsten Tage in dieser Hinsicht. Für diese Umfrage sagte der Readers Digest den Sieg von Landon voraus,
wohingegen die Meinungsforscher Crossley, Roper und Gallup dies ohne gegenseitige Rücksprache bekannt gaben der Sieg seines Gegners Roosevelt. Roo Sevelt wurde gewählt. Der Fehler des Readers Digest bestand darin, seine Vorhersage auf das zu stützen
Das Ergebnis der Wahl ergab sich aus einer Telefonumfrage, doch damals besaßen nur reiche Leute ein Telefon, daher war die Stichprobe voreingenommen!
Das Jahr 1938 markiert in Frankreich die Geburtsstunde des IFOP (Institut Français d'Opinion).
Publique, das französische Institut für Öffentlichkeit Meinung) erstellt von Stoetzel, Jean.
Es ist interessant, ein wenig in die Vergangenheit zu reisen um die Rolle hervorzuheben, die russische Statistiker bei der Entwicklung der Stichprobenziehung gespielt haben Techniken. Tatsächlich waren sie bereits im 19. Jahrhundert bekannt und wurden in der Landwirtschaft verwendet Land. Laut Tassi, Ph. (1988),
Chuprov, AI (1842–1908) war einer der Vorläufer des modernen Sampling und von 1910 verwendete sein Sohn Chuprov, AA, zufällig Stichprobe, bezogen auf Cluster-Stichprobe, **geschichtete Stichprobe** mit und ohne Ersatz und unter dem Einfluss von
Markov, A., befürwortete die Verwendung von **Wahrscheinlichkeiten** in **der Statistik.** Auch wenn der Russe Die Schule sei erst sehr spät für ihre Stichprobenarbeiten bekannt geworden, so die Statistiker dieser Schule Offenbar wusste ich schon lange von diesen Techniken vor denen Westeuropas.
Beachten Sie, dass Kovalevsky, AG im Jahr 1924 hatte bereits intensiv an **Umfragen** von gearbeitet Schichtung und die optimale Allokation durch stratum, während dieses Ergebnis erst 10 Jahre später von Neyman, J. wiederentdeckt wurde.
Schließlich wäre diese Geschichte unvollständig ohne die Beiträge von zu erwähnen
Neyman, Jerzy zum Sampling. Berücksichtigt einer der Begründer der statistischen Theorie
Bei den **Umfragen** befürwortete er die Zufallsbefragung Probenahme auf gezielter Probenahme, und er arbeitete an Stichproben mit oder ohne Überweisung, **geschichteter Stichprobe** und **Cluster Probenahme.** Er nutzte bereits **Schätzungen** nach Quotienten, Differenzen und Regressionen, Techniken, die weiterentwickelt werden sollten
später von Horvitz, DG, Cochran, WG,
Hansen, MM, Hurwitz, WN und Madow, WG

S

MATHEMATISCHE ASPEKTE

Es gibt verschiedene Probenahmemethoden. Sie kann in Zufallsmethoden und unterteilt werden nichtzufällige Methoden.

Das am weitesten verbreitete nichtzufällige Verfahren ist **Quotenstichprobe.**
Die Zufallsmethoden (oder probabilistische Stichproben, manchmal auch statistische Stichproben genannt) basieren auf einem probabilistischen Mechanismus eine **Stichprobe** aus einer **Population** bilden .

Machine Translated by Google

Die in der **Stichprobe** beobachteten **Variablen** sind **Zufallsvariablen.** Aus diesen Variablen entsprechende Werte der **Grundgesamtheit** und der Fehler aufgrund der Stichprobe können berechnet werden.

Die wichtigsten Zufallsmethoden sind:

• **Einfache Zufallsstichprobe** • **Stratifizierte Stichprobe** • **Systematische Stichprobe** • **Cluster-Stichprobe**

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Ziel der Stichprobenerhebung besteht darin, genügend Informationen bereitzustellen, um **Rückschlüsse** auf den Charakter zu ziehen Eigenschaften einer **Population.** Daher ist es notwendig, eine **Probe** auszuwählen , die diese Eigenschaften möglichst genau wiedergibt.

Es wird versucht, den potenziellen **Fehler** durch die Stichprobenentnahme so weit wie möglich zu reduzieren. Aber selbst wenn dieser Stichprobenfehler unvermeidlich ist, wird er toleriert, da er durch die Vorteile der Stichprobenmethode weitgehend kompensiert wird. Diese sind:

• *Kostenreduzierung* Angesichts aller Kosten, die dabei anfallen können Während der **Datenerhebung** können sie offensichtlich durch Stichproben reduziert werden. • *Zeitersparnis* Die Geschwindigkeit der Entscheidungsfindung kann nur gewährleistet werden, wenn die Studie auf eine **Stichprobe beschränkt ist.** • *Erweiterte Möglichkeiten* In bestimmten Bereichen ist eine vollständige Analyse einer gesamten **Population** unmöglich; Die Wahl besteht also darin, entweder Stichproben zu verwenden oder die Studie auf Eis zu legen. • *Präzise Ergebnisse* Denn eines der Ziele der Stichprobenziehung besteht darin, ein **repräsentatives** Bild der Merkmale einer größeren Population wiederherzustellen Proben würden keine wesentlich genaueren Ergebnisse bringen.

Nur Zufallsstichprobenverfahren können den **Fehler** aufgrund der Stichprobe bewerten, da sie auf der Berechnung von **Wahrscheinlichkeiten basieren.**

WEITERLESEN

Cluster-Sampling
Datensammlung
Einschätzung
Schätzer
Quotenstrichprobenerhebung
Einfache Zufallsstichprobe
Geschichtete Stichprobe
Systematische Probenahme

REFERENZEN

Bowley, AL: Ansprache des Präsidenten an die Wirtschaftsabteilung des British Asso Ziation. J. Roy. Stat. Soc. **69,** 540–558 (1906)

Delanius, T.: Elemente der Stichprobenerhebung. SAREC, Stockholm (1985)

Desrosières, A. Die Partei für alles: Allgemeiner Kommentar? Schätzung und Sondagen. Fünf Beiträge zur Statistikgeschichte. Economica, Paris (1988)

Foody, WN, Hedagat, AS: Zur Theorie und Anwendung von BIB-Designs mit wiederholten Blöcken.Annals of Statistics,V=S 932–945 (1977)

Hedayat, AS, Sinha, BK: Design und Inferenz bei der Stichprobenziehung endlicher Populationen. Wiley, New York (1991)

Tassi, Ph.: Ausführlich zum Teil: ein Blick auf die Geschichte der Entwicklung von Sondagen. Schätzungen und Untersuchungen, fünf Beiträge zur Geschichte der Statistik. Economica, Paris (1988)

Machine Translated by Google

Stichprobenverteilung

Eine Stichprobenverteilung ist eine Verteilung einer **Statistik** *T(X1,...,Xn),* wobei *T(.)* ist eine Funktion der beobachteten **Werte** des **Laufs dom-Variablen** *X1,...,Xn.*

GESCHICHTE

Siehe **Probenahme.**

MATHEMATISCHE ASPEKTE

Betrachten Sie eine **Population** der Größe *N* und von **Parameter** ÿ und ÿ (wobei ÿ die Unbekannte ist). **Mittelwert** der Grundgesamtheit und ÿ sein **Standard Abweichung).** Wir nehmen eine **Stichprobe** der Größe *n* aus diese **Bevölkerung.** Die Anzahl der möglichen Stichproben der Größe *n* ergeben sich aus der Anzahl **Kombinationen** von *n* Objekten unter N:

$$k = Cn_N = \frac{N!}{(N \ddot{y} n)! \cdot n!} .$$

Jede Stichprobe kann beispielsweise durch ihren Stichprobenmittelwert charakterisiert werden, der wie folgt bezeichnet wird:

$$\bar{x}1, \bar{x}2,..., \bar{x}k .$$

Die Menge dieser *k* Mittel bildet die Stichprobe Verteilung der Mittel.

Wir werden die folgenden Notationen übernehmen Charakterisieren Sie den **Mittelwert** und **die Standardabweichung ation** der Stichprobenverteilung der bedeutet:

Mittelwert: $\ddot{y}\bar{x}$ und Standardabweichung: $\ddot{y}\bar{x}$.

Mittelwert der Stichprobenverteilung der Mittelwerte

Betrachten Sie den unabhängigen **Zufall** *X1,...,Xn* **Variablen ,** die gemäß einer Verteilung mit einem Mittelwert ÿ und einer **Varianz** ÿ2 verteilt sind. Von Definition des **arithmetischen Mittels,** des Mittelwerts der Stichprobenverteilung der Mittel ist gleich:

$$\ddot{y}\bar{x} = \frac{\sum_{i=1}^{k} \bar{x}i}{k} ,$$

Dabei ist *k* die Anzahl der **Stichproben** der Größe *n* dass es möglich ist, sich aus der **Bevölkerung zu bilden.**

Der **erwartete Wert** des **Mittelwerts** der Stichprobenverteilung der Mittel ist gleich dem Mittelwert der **Grundgesamtheit** ÿ:

$$E[\ddot{y}\bar{x}] = \ddot{y} .$$

Weil:

$$E[\ddot{y}\bar{x}] = E\left[\frac{\sum_{i=1}^{k} \bar{x}i}{k}\right] = \frac{1}{k} \cdot E\left[\sum_{i=1}^{k} \bar{x}i\right]$$

$$= \frac{1}{k} \cdot \sum_{i=1}^{k} E[\bar{x}i]$$

$$= \frac{1}{k} \cdot \sum_{i=1}^{k} E\left[\frac{\sum_{j=1}^{N} xj}{N}\right]$$

$$= \frac{1}{k} \cdot \sum_{i=1}^{k} \ddot{y}\frac{1}{N} \cdot E\left[\ddot{y}_{\ddot{y}j=1}^{N} xj\ddot{y}\ddot{y}\right]$$

$$= \frac{1}{k} \cdot \sum_{i=1}^{k} \ddot{y}\frac{1}{N} \cdot \sum_{j=1}^{N} E[xj]\ddot{y}\ddot{y}$$

$$= \frac{1}{k} \cdot \sum_{i=1}^{k} \frac{1}{N} \cdot n \cdot \ddot{y}$$

$$= \frac{1}{k} \cdot k \cdot \ddot{y} = \ddot{y} .$$

Standardabweichung der Stichprobe Verteilung der Mittel

Per Definition ist die Standardabweichung der Stichprobenverteilung der Mittelwerte ebenfalls

S

Machine Translated by Google

wird als **Standardfehler** bezeichnet und ist gleich:

$$\ÿx^- = \frac{\sqrt{\frac{\sum_{i=1}^{k} (x^-_i \ÿ \ÿx^-)^2}{k}}}{}$$

Ebenso wie ein Zusammenhang zwischen dem **Mittelwert** der Stichprobenverteilung der Mittelwerte und dem Mittelwert der **Grundgesamtheit besteht,** besteht auch ein Zusammenhang zwischen dem **Standardfehler** und der **Standardabweichung** der Grundgesamtheit:

- Im Falle einer erschöpfenden **Stichprobe** (ohne Ersatz) einer endlichen **Grundgesamtheit** ist der **Standardfehler** des Mittelwerts gegeben durch:

$$\ÿx^- = \frac{\ÿ}{\ÿn} \cdot \sqrt{\frac{N \ÿ n}{N \ÿ 1}},$$

  wobei ÿ die **Standardabweichung** des ist **Bevölkerung.**

- Wenn die **Grundgesamtheit** unendlich ist oder wenn die untersuchte **Stichprobe** nicht erschöpfend ist (mit Ersatz), ist der **Standardfehler** des **Mittelwerts** gleich der **Standardabweichung** der **Grundgesamtheit** dividiert durch die Quadratwurzel der **Stichprobengröße :**

$$\ÿx^- = \frac{\ÿ}{\sqrt{\ÿn}}.$$

Merkmale der Stichprobenverteilung der Mittelwerte

- Wenn groß genug *(n* ÿ 30), folgt die Stichprobenverteilung der **Mittelwerte** unabhängig von der **Bevölkerungsverteilung** ungefähr einer **Normalverteilung .** • Wenn die **Grundgesamtheit** gemäß einer **Normalverteilung verteilt ist,** folgt die Stichprobenverteilung der **Mittelwerte** immer einer Normalverteilung, unabhängig von der Größe der **Stichprobe.**
- Die Stichprobenverteilung des **Mittelwerts** ist immer symmetrisch um seinen Mittelwert.

Ähnliche Überlegungen können angestellt werden andere **Statistiken** als das **arithmetische Mittel.** Betrachten Sie die Verteilung der **Varianzen** erhalten aus allen möglichen **Stichproben** der Größe n, entnommen aus einer Normalpopulation **mit** einer Varianz ÿ2. Diese Verteilung wird als Stichprobenverteilung der Varianzen bezeichnet. Eines der Merkmale dieser Verteilung ist, dass sie nur nicht negative **Werte** annehmen kann , wobei die **Varianz** s2 durch eine Quadratsumme definiert wird. Die Stichprobenverteilung der Varianzen hängt durch die folgende Beziehung mit der **Chi-Quadrat-Verteilung** zusam

$$^2 s = \frac{\ÿ2\ÿ2}{v} = \frac{\ÿ2\ÿ2}{n \ÿ 1},$$

wobei ÿ2 eine nach einer **Chi-Quadrat-Verteilung** verteilte **Zufallsvariable** mit *v = n* ÿ 1 **Freiheitsgraden ist.**

Den **Erwartungswert** und die **Varianz** einer Chi-Quadrat-Verteilung kennen :

$$E\ÿ2 = v \quad ,$$

$$\text{Var}\ÿ2 = 2v \quad ,$$

Wir können dieselben Merkmale für die Verteilung *von s2* bestimmen , wenn die **Stichproben** aus einer unendlichen Grundgesamtheit stammen .

Mittelwert der Stichprobenverteilung von Abweichungen

Der **Mittelwert** der Verteilung von s2 ist gleich der **Varianz** der **Grundgesamtheit:**

$$\ÿs2 = E\,s^2 = E\left(\frac{\ÿ2\ÿ2}{v}\right)$$

$$= \frac{\ÿ2}{v} \cdot E\ÿ2$$

(weil ÿ2 /*v* eine Konstante ist)

Machine Translated by Google

$$= \frac{\ddot{y}2}{v} \cdot v$$

$$= \ddot{y}2 \quad .$$

Varianz der Stichprobenverteilung von Varianzen

Die **Varianz** der Verteilung von s2 hängt nicht nur von der Varianz der ab **Population** ÿ2, sondern auch von der Größe der **Stichproben** $n = v + 1$:

$$\ddot{y}2_{s2} = \text{Var } s^2 = \text{Var } \frac{\ddot{y}2\ddot{y}2}{v}$$

$$= \frac{\ddot{y}4}{v2} \cdot \text{Var } \ddot{y}2$$

$$= \frac{\ddot{y}4}{v2} \cdot 2v$$

$$= 2 \cdot \frac{\ddot{y}4}{v} \quad .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Untersuchung der Stichprobenverteilung von Eine **Statistik** ermöglicht es, die Nähe zu beurteilen der an einer einzelnen **Probe** gemessenen Statistik mit einem unbekannten **Parameter** der **Bevölkerung.** Es ermöglicht daher auch die Angabe der **Fehlermargen** der **Schätzer,** berechnet aus der Probe selbst.

Der Begriff der Stichprobenverteilung ist daher an der Basis des Baus von **Konfidenzintervalle,** die den Grad angeben der Genauigkeit eines **Schätzers,** sowie die Durchführung von **Hypothesentests** auf dem unbekannte **Parameter** einer **Population.**

BEISPIELE

Anhand eines Beispiels zeigen wir den Zusammenhang, der zwischen dem Mittelwert einer Stichprobe besteht Verteilung und der Mittelwert einer **Population** und die Beziehung zwischen dem **Standardfehler** und die **Standardabweichung** dieser Population.

Stellen Sie sich eine **Population** von fünf Geschäften vor, deren Durchschnittspreis für ein bestimmtes Produkt, das wir wollen wissen. Die folgende Tabelle gibt den Preis an des Produkts für jedes Geschäft, das dazu gehört Bevölkerung:

Shop-Nr.	Preis
1	30,50
2	32.00
3	37,50
4	30.00
5	33.00

Der **Mittelwert** der **Grundgesamtheit** ist gleich:

$$\ddot{y} = \frac{\sum_{i=1}^{5} xi}{5}$$

$$= \frac{30{,}50 + 32 + 37{,}50 + 30 + 33}{5}$$

$$= 32{,}60 \quad .$$

Durch die Auswahl von **Stichproben** der Größe 3 können wir die verschiedenen Stichprobenmittel bestimmen, die dies können Ergebnis einer zufälligen Auswahl der Stichprobe. Wir erhalten:

$$k = C3_5 = \frac{5!}{3! \cdot 2!} = 10 \text{ mögliche Proben.}$$

Die folgende Tabelle stellt die verschiedenen **Proben** und ihre jeweilige Probenahme dar bedeutet:

Anzahl der Filialen in der Stichprobe		Probenahme bedeutet x_i
1.	1, 2, 3	33.333..
2.	1, 2, 4	30.833..
3.	1, 2, 5	31.833..
4.	1, 3, 4	32.666..
5.	1, 3, 5	33.666..
6.	1, 4, 5	31.166..

S

Machine Translated by Google

Anzahl der Filialen in der Stichprobe		Probenahme bedeutet x_i
7.	2, 3, 4	33.166..
8.	2, 3, 5	34.166..
9.	2, 4, 5	31.666..
10.	3, 4, 5	33.500
	Gesamt	326.000

Aus den **Daten** in dieser Tabelle können wir berechnen der **Mittelwert** der Stichprobenverteilung der bedeutet:

$$\ddot{y}x^{-} = \frac{\sum_{i=1}^{k} x^{-}i}{k} = \frac{326{,}00}{10} = 32{,}60 .$$

In diesem Beispiel stellen wir fest, dass der **Mittelwert** von Die Stichprobenverteilung ist gleich dem Mittelwert der **Bevölkerung:**

$$\ddot{y}x^{-} = \ddot{y} .$$

Wir können auf die gleiche Weise vorgehen, um das zu überprüfen Beziehung zwischen dem **Standardfehler** und dem **Standardabweichung** der **Grundgesamtheit.** Per Definition ist die **Standardabweichung** der **Bevölkerung** wird wie folgt berechnet:

$$\ddot{y} = \sqrt{\frac{\sum_{i=1}^{N} (xi \ddot{y} \ddot{y})2}{N}} .$$

Nach den **Daten** der ersten Tabelle, wir erhalten:

$$\ddot{y} = \sqrt{\frac{35{,}70}{5}} = 2{,}6721 .$$

Mit der gleichen Definition können wir die berechnen **Standardabweichung** der Stichprobenverteilung der Mittelwerte:

$$\ddot{y}x^{-} = \sqrt{\frac{\sum_{i=1}^{k} (x^{-}i \ddot{y} \ddot{y}x^{-})2}{k}}$$

$$= \sqrt{\frac{11.899}{10}} = 1{,}0908 .$$

In der folgenden Tabelle werden die Details aufgeführt Berechnungen:

Anzahl der Geschäfte im Muster		Probenahme bedeutet x_i	$(x_i \ddot{y} \ddot{y} \bar{x})2$
1.	1, 2, 3	33.333..	0,538
2.	1, 2, 4	30.833..	3.121
3.	1, 2, 5	31.833..	0,588
4.	1, 3, 4	32.666..	0,004
5.	1, 3, 5	33.666..	1.138
6.	1, 4, 5	31.166..	2.054
7.	2, 3, 4	33.166..	0,321
8.	2, 3, 5	34.166..	2.454
9.	2, 4, 5	31.666..	0,871
10.	3, 4, 5	33.500	0,810
	Gesamt	326.000	11.899

Nach den erhaltenen Ergebnissen:

$$\ddot{y} = 2{,}6721 \text{ und } \ddot{y}x^{-} = 1{,}0908 ,$$

Wir können die folgende Gleichheit überprüfen:

$$\ddot{y}x^{-} = \frac{\ddot{y}}{\sqrt{\ddot{y}n}} \cdot \sqrt{\frac{N\ddot{y}\,n}{N\ddot{y}\,1}}$$

weil $\ddot{y}\ddot{y}\,n \sqrt{\ddot{y}n} \cdot \sqrt{\frac{N\ddot{y}n}{N\ddot{y}1}} = \frac{2.6721}{\ddot{y}\,\bar{3}} \cdot \sqrt{\frac{5\ddot{y}3}{5\ddot{y}1}} =$ 1.0908. Diese Ergebnisse lassen eine Schlussfolgerung zu das durch Kenntnis der **Parameter** ŷ und ÿ der **Bevölkerung** können wir die entsprechenden Merkmale der Stichprobe bewerten Verteilung der **Mittel.** Im Gegenteil, wir kann auch die Parameter der Bevölkerung bestimmen, indem er die Merkmale der Bevölkerung kennt Stichprobenverteilung.

WEITERLESEN

- **Konfidenzintervall**
- **Einschätzung**
- **Hypothesentest**
- **Standart Fehler**

Machine Translated by Google

Streudiagramm

Durch Transkribieren wird ein Streudiagramm erstellt die **Daten** einer **Probe** auf eine Grafik. In **In** zwei Dimensionen kann ein Streudiagramm *n* Beobachtungspaare *(xi;* yi) darstellen , *i* = *1, 2,...,n.*



MATHEMATISCHE ASPEKTE

Durch Platzieren der Paare entsteht ein Streudiagramm der beobachteten **Werte** in einem **Achsensystem .** Für zwei **Variablen** *X* und Y sind die Paare von entsprechende Punkte (x1, y1), (x2, y2), ..., *(xn, yn)* werden in einem rechteckigen **Achsensystem** platziert tem.

Normalerweise wird **die abhängige Variable Y** aufgetragen auf der vertikalen **Achse (Ordinate)** und die **unabhängige Variable** *X* auf der horizontalen Achse **(Abszisse).**

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Streudiagramm ist ein sehr nützliches und leistungsstarkes Werkzeug, das häufig in der **Regressionsanalyse verwendet wird.** Jeder Punkt repräsentiert ein Paar beobachteter **Werte** der **abhängigen Variablen** und der **unabhängige Variable.** Damit kann grafisch ermittelt werden, ob eine **Beziehung** besteht zwischen zwei **Variablen,** bevor Sie eine auswählen entsprechendes **Modell.** Darüber hinaus sind diese Streudiagramme bei der **Residuenanalyse** sehr nützlich da sie es ermöglichen, das **Modell** zu überprüfen ob es angemessen ist oder nicht.

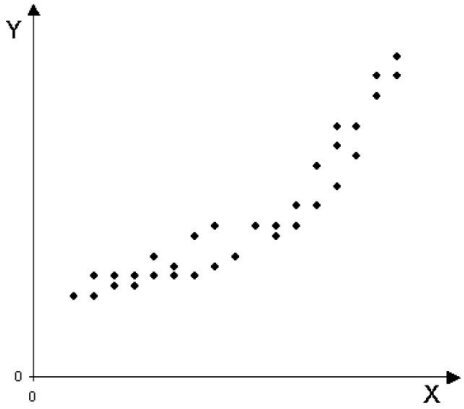
BEISPIELE

Betrachten Sie die folgenden zwei Beispiele von Scat-Terplots, die **Beobachtungspaare** darstellen *(xi, yi)* relativ zu zwei **quantitativen Variablen** *X* und Y:

In diesem Beispiel scheint die Verteilung der Beobachtungen auf einen linearen **Zusammenhang** hinzuweisen zwischen den beiden **Variablen:**

Die Verteilung der **Beobachtungen** im

Die folgende Grafik scheint auf eine nichtlineare Beziehung zwischen *X* und Y hinzuweisen:



WEITERLESEN

- **Analyse von Residuen**
- **Regressionsanalyse**

S

Scheffé, Henry

Scheffé, Henry wurde 1907 in New geboren York. Scheffé, Henrys Doktorarbeit „The Asymptotic Solutions of Cer tain Linear Differential Equations in Which der Koeffizient des Parameters kann sein a Zero“ wurde von Langer, Rudolph betreut E. Sofort begann er seine Karriere als Universitätsprofessor für Mathematik an der University of Wisconsin, dann an der Oregon State Universität sowie am Reed College außerhalb

Machine Translated by Google

von Portland. 1941 wechselte er an die Princeton-Universität, wo ein Statistikteam aufgebaut worden war. Als Universitätsstudent begann für ihn eine zweite Karriere Universitätsprofessor für Statistik. Im akademischen Jahr 1944–1945 war er an der Fakultät der Syracuse University und von 1946 bis 1948 an der University of California in Los Angeles tätig.
Er verließ Los Angeles, um an die Columbia University zu gehen, wo er Vorsitzender der Statistikabteilung wurde. Nach fünf Jahren an der Columbia University ging Scheffé 1953 als Professor für Statistik nach Berkeley.

Die Forschung von Scheffé, H. befasste sich hauptsächlich mit der Varianzanalyse. Eine seiner wichtigsten Arbeiten erschien 1953 über die S-Methode des Simultankonfidenzinterviews

Werte für schätzbare Funktionen in einem Unterraum des Parameterraums. Er untersuchte auch Paarvergleiche und gemischte Modelle. 1958 und erneut 1963 veröffentlichte er Artikel über Gemischexperimente und schrieb 1973 weiter Kalibrierungsmethoden.

Scheffé, H. wurde 1944 Mitglied des Institute of Mathematical Statistics, 1952 der American Statistical Association und 1964 des International Statistical Institute

wurde zum Präsidenten des International Statistical Institute und zum Vizepräsidenten der American Statistical Association gewählt. Er starb in Kalifornien im Jahr 1979.

Ausgewählte Werke von Henry Scheffé:

1959 Die Varianzanalyse. Wiley, New York.

1952 Eine Varianzanalyse für paarweise Vergleiche. Marmelade. Stat. Assoc., 47, 381–400.

1953 Eine Methode zur Beurteilung aller Kontraste in der Varianzanalyse. Biometrie, 40, 87–104.

1956 Alternative Modelle zur Varianzanalyse. Ann. Mathematik. Stat., 27, 251–271.

1973 Eine statistische Theorie der Kalibrierung. Ann. Stat., 1:1–37.

WEITERLESEN

Varianzanalyse

Test auf den geringsten signifikanten Unterschied

Saisonindex

Betrachten Sie eine **ZeitreiheYt,** deren Komponenten Sind:

• **Säkularer Trend** *Tt* •

Zyklische Schwankungen *Ct* •

Saisonale Schwankungen *St* •

Unregelmäßige Schwankungen *It*

Der Einfluss saisonaler Schwankungen muss berücksichtigt werden über ein ganzes Jahr hinweg neutral sein, und saisonale Schwankungen *St* wiederholen sich theoretisch von Periode zu Periode auf identische Weise (was in einem realen Fall nicht überprüft wird).
Um die Anforderungen des theoretischen **Modells** zu erfüllen und die realen **Zeitreihen** untersuchen zu können , müssen periodische Variationen, die jedes Jahr identisch sind (Monat für Monat oder Trimester für Trimester), sogenannte saisonale Variationen, anstelle des beobachteten *St* geschätzt werden. Sie werden mit *Sj bezeichnet,* wobei *j* wie folgt variiert:

j = **1** bis 12 für Monate (über *n* Jahre) oder

j = **1** bis 4 für Trimester (über *n* Jahre).

Beachten Sie, dass es über *n* Jahre hinweg nur 12 saisonale Indexzahlen für die Monate und 4 für die Trimester gibt.

MATHEMATISCHE ASPEKTE

Es gibt viele Methoden zur Berechnung der Saisonindizes:

Machine Translated by Google

1. *Prozentsatz des Mittelwerts*

Die Daten, die jedem Monat entsprechen, sind ausgedrückt als **Prozentsätze** des **Mittelwerts.** Wir berechnen es mit dem Mittelwert (ohne). unter Berücksichtigung von Extremwerten) oderthe **Median** der Prozentsätze jedes Monats für die verschiedenen Jahre. Den Mittelwert nehmen Für jeden Monat erhalten wir 12 Prozentsätze gegebene Indizes. Wenn der Mittelwert nicht gleich ist 100 %, wir sollten eine Anpassung vornehmen Multiplikation der Indizes mit einem praktischen Faktor.

Betrachten Sie eine monatliche **Zeitreihe** *n* Jahre, bezeichnet mit *Yi,j,* wobei *j* = *1, 2,...,* 12 für die verschiedenen Monate und *i = 1,...,n,* angegeben nach Jahr. Zuerst gruppieren wir die Daten in einer Tabelle Typ:

	Monat		
Jahr 1	Januar ...	...	Dezember 12
1	*Y1,1*	...	*Y1,12*
...	...	...	...
N	*Yn,1*	...	*Yn,12*

Dann berechnen wir die Jahresmittelwerte für die *n* Jahre:

$$Mi = \frac{\bar{Y}i,.}{12}$$

für *i = 1,...,n.*

Somit berechnen wir für jedes Datum:

$$Qi,j = \frac{Yi,j}{Mi} \quad \text{für } i = 1,...,n$$

und *j = 1,...,* 12, worüber wir im berichten Folgende Tabelle:

	Monat		
Jahr 1	Januar ...	...	Dezember 12
1	*Q1,1*	...	*Q1,12*
...	...	...	...
N	*Qn,1*	...	*Qn,12*

In jede Spalte geben wir die Arithmetik ein **bedeuten,** bezeichnet mit $\bar{Q}\,i,.$ und wir berechnen die Summe dieser 12 Mittel; wenn es nicht gleich 1200, dann führen wir einen Korrekturfaktor ein:

$$k = \frac{1200}{\sum_{i=1}^{12} \bar{Q}\,i,.}.$$

Abschließend werden die Saisonindizes für angegeben jeden Monat durch den Mittelwert multipliziert mit k.

2. *Tendenzverhältnis*

Die Daten jedes Monats werden ausgedrückt als Prozentsätze des monatlichen Werts des **weltliche Tendenz.** Der Index wird erhalten aus einem geeigneten Mittelwert des prozentualen Alters der entsprechenden Monate. Wir sollte die Ergebnisse anpassen, wenn der Mittelwert dies tut nicht gleich 100 %, um die gewünschten Saisonindizes zu erhalten.

3. *Verhältnis des gleitenden Durchschnitts*

Wir berechnen einen **gleitenden Durchschnitt** auf 12 Monate. Wie die entsprechenden Ergebnisse sind über viele Monate ausgedehnt, statt in der Mitte des Monats sein wie in Aus den Originaldaten berechnen wir den gleitenden Durchschnitt über 2 Monate oder den gleitenden Durchschnitt von 12 Monaten. Das ist was Wir nennen einen zentrierten gleitenden Durchschnitt Over 12 Monate.

Dann drücken wir die Originaldaten aus jeden Monat als Prozentsatz des entsprechenden zentrierten gleitenden Durchschnitts 12 Monate. Wir nehmen den Mittelwert der entsprechenden Prozentsätze jedes Monats, was die Indizes ergibt. Das sollten wir dann tun Passen Sie den Wert auf 100 % an, um die gewünschten saisonalen Indizes zu erhalten.

4. *Relative Ketten*

Die Daten jedes Monats werden ausgedrückt als Prozentsätze der Daten des Vormonats (daher der Ausdruck relative Kette). Wir berechnen einen angemessenen

Machine Translated by Google

Mittelwert der Ketten für das entsprechende Monate. Wir können die relativen Prozentsätze jedes Monats im Verhältnis zu diesen ermitteln Januar (= 100 %). Wir finden dann dass der folgende Januar normalerweise einen Prozentsatz haben wird, der kleiner oder gleich ist. oder größer als 100 %. Die Tendenz ist also entweder steigend oder fallend. Der Endgültig erhaltene Indizes, angepasst an ihre Mit einem Mittelwert von 100 % ergeben sich die gewünschten Saisonindizes.

BEISPIELE

Wir betrachten die **Zeitreihe** jedes Phänomens; Wir haben Trimesterdaten 6 Jahre, zusammengefasst in der folgenden Tabelle:

Jahr	1 trimmen.	2 trimmen.	3 trimmen.	4 trimmen.
1	19.65	16.35	21.30	14,90
2	28.15	25.00	29,85	23.40
3	36,75	33,60	38,55	32.10
4	45.30	42,25	47.00	40,65
5	54.15	51,00	55,75	49,50
6	62,80	59,55	64,40	58.05

Wir ermitteln die Saisonindizes anhand der **Prozentsatz** des **Mittelwerts:** Zuerst berechnen wir das arithmetische Mittel für 6 Jahre:

Jahressummenmittel	
1	72,2 18.05
2	106,4 26,6
3	141,0 35,25
4	175,2 43,8
5	210,4 52,6
6	244,8 61,2

Wir dividieren die Zeilen der Anfangstabelle durch Trimestermittel des entsprechenden Jahres und mit 100 multiplizieren.

Dann, wenn die Ergebnisse starke Abweichungen von den erhaltenen Werten für das Gleiche zeigen Trimester ist es besser, den **Median** zu nehmen anstelle des arithmetischen Mittels, um das zu erhalten aussagekräftige Zahlen.

Jahr	1 trimmen.	2 trimmen.	3 trimmen.	4 trimmen.
1	108,86	90,58	118,01	82,55
2	105,83	93,98	112,22	87,97
3	104,26	95,32	109,36	91,06
4	103,42	96,46	107,31	92,81
5	102,95	96,96	105,99	94,11
6	102,61	97,30	105,23	94,85
Median	103,84	95,89	108,33	91,94

Der mittlere Prozentsatz für jedes Trimester beträgt in der letzten Zeile der vorherigen Tabelle angegeben Die Summe dieser Prozentsätze beträgt genau 400 %, Eine Anpassung ist nicht erforderlich, und die Zahlen, die wir in der letzten Zeile haben, repräsentieren die gewünschte Saisonindizes.

Wir haben gesehen, dass es noch drei weitere Meth gibt Odds zur Bestimmung der Saisonindizes und Sie geben dafür die folgenden Werte an Problem:

	Trimester			
Methode	**1**	**2**	**3**	**4**
Prozentsatz von gemein	103,84	95,89	108,33	91,94
Tendenz Verhältnis	112,62	98,08	104,85	84,45
Verhältnis von ziehen um Durchschnitt	111,56	98,87	106,14	83,44
Ketten	112,49	98,24	104,99	84,28

Wie wir sehen, stimmen die Ergebnisse miteinander überein, trotz der Vielfalt der verwendeten Methoden bestimmen sie.

Machine Translated by Google

WEITERLESEN

Zyklische Schwankung
Unregelmäßige Variation
Saisonale Unterschiede
Säkularen Trend
Zeitfolgen

Saisonale Unterschiede

Saisonale Schwankungen bilden eine der Grundlagen Komponenten der **Zeitreihe.** Sie sind Variationen periodischer Natur, die regelmäßig wiederkehren. Im wirtschaftlichen Kontext können saisonale Schwankungen durch folgende Ereignisse verursacht werden:

- Klimatische Bedingungen
- Für eine **Bevölkerung** angemessene Bräuche
- Religiöse Feiertage

Es handelt sich mehr oder weniger um periodische Schwankungen regelmäßig, die mit **außersaisonaler sonaler** Bewegung überlagert sind .

GESCHICHTE

Siehe **Zeitreihen.**

MATHEMATISCHE ASPEKTE

Die Auswertung saisonaler Schwankungen erfolgt durch die Bestimmung eines **Index** der saisonalen Variation, des so genannten **Saisonindex.**

BEISPIELE

Wir betrachten die **Zeitreihe** jedes Phänomens; Wir haben Trimesterdaten erfasst über einen Zeitraum von 6 Jahren, im Folgenden zusammengefasst Tisch:

Jahr 1		2	3	4
	trimmen.	trimmen.	trimmen.	trimmen.
1	19.65	16.35	21.30	14,90
2	28.15	25,00	29,85	23.40
3	36,75	33,60	38,55	32.10
4	45.30	42,25	47,00	40,65
5	54.15	51,00	55,75	49,50
6	62,80	59,55	64,40	58.05

Wir ermitteln die Saisonindizes anhand der **Prozentsatz** des **Mittelwerts:** Zuerst berechnen wir die Trimesterarithmetik bedeutet für die 6 Jahre:

Jahressummenmittel	
1	72,2 18.05
2	106,4 26,6
3	141,0 35,25
4	175,2 43,8
5	210,4 52,6
6	244,8 61,2

Wir dividieren die Zeilen der Anfangstabelle durch Trimestermittel des entsprechenden Jahres und mit 100 multiplizieren.
Dann, wenn die Ergebnisse große Abweichungen aufweisen für die erhaltenen Werte für dasselbe Trimester, es ist besser, stattdessen den **Median** zu nehmen aussagekräftigere Zahlen erhalten.

Jahr	1	2	3	4
	trimmen.	trimmen.	trimmen.	trimmen.
1	108,86	90,58	118,01	82,55
2	105,83	93,98	112,22	87,97
3	104,26	95,32	109,36	91,06
4	103,42	96,46	107,31	92,81
5	102,95	96,96	105,99	94,11
6	102,61	97,30	105,23	94,85
Median	103,84	95,89	108,33	91,94

Der mittlere Prozentsatz für jedes Trimester beträgt gegeben durch die letzte Zeile der vorherigen Tabelle. Da sich diese Prozentsätze genau addieren 400 %, eine Anpassung ist nicht erforderlich, und

Machine Translated by Google

Die Zahlen in der letzten Zeile stellen die gewünschten Saisonindizes dar.

Wir haben gesehen, dass es andere Methoden gibt die Saisonindizes zu bestimmen; Sie geben für dasselbe Problem die folgenden Werte an:

	Trimester			
Methode	**1**	**2**	**3**	**4**
Prozentsatz des Mittelwerts	103,84	95,89	108,33	91,94
Tendenzverhältnis	112,62	98,08	104,85	84,45
Verhältnis von gleitender Durchschn.	111,56	98,87	106,14	83,44
Ketten	112,49	98,24	104,99	84,28

Wir sehen, dass die Ergebnisse trotz der Vielfalt der Methoden, mit denen sie ermittelt wurden, übereinstimmen.

WEITERLESEN

- **Arithmetisches Mittel**
- **Zyklische Schwankung**
- **Prognose**
- **Indexnummer**
- **Gleitender Durchschnitt**
- **Saisonindex**
- **Säkularen Trend**
- **Zeitfolgen**

REFERENZEN

Fuller, AW: Einführung in statistische Zeitreihen. John Wiley'n'Sas (1976)

Säkularen Trend

Der säkulare Trend bildet einen der vier Grundkomponenten der **Zeitreihe.** Es beschreibt die langfristige Bewegung einer Zeitreihe, die global steigend, fallend oder stabil sein kann.

Der säkulare Trend kann linear sein oder nicht. Im Im letzten Fall sollten wir den Funktionstyp (exponentiell, quadratisch, logistisch usw.) wählen, der am besten an die beobachtete Verteilung in der **grafischen Darstellung** der Zeitreihe angepasst ist, bevor wir den säkularen Trend schätzen.

GESCHICHTE

Siehe **Zeitreihen.**

MATHEMATISCHE ASPEKTE

Sei Yt eine **Zeitreihe.** Sie kann mit Hilfe von vier Komponenten zerlegt werden: • Säkularer Trend *Tt* • **Zyklische Fluktuation** *Ct* • **Saisonale Schwankungen** *St* • **Unregelmäßige Schwankungen** *It* Während der Untersuchung der Zeitreihe beginnen wir mit der Analyse des säkularen Trends und entscheiden zunächst, ob dieser Trend sein kann angeblich linear ode Wenn wir annehmen können, dass es linear ist, kann die Bewertung des säkularen Trends auf viele Arten erfolgen: 1. *Grafische Methode oder die erhobene Hand*

Nachdem Sie eine **grafische Darstellung** der Zeitreihe erstellt haben, müssen Sie eine Trendlinie oder -kurve bei erhobener Hand im Diagramm anpassen. Dies ist die Annäherung an den säkularen Trend.

2. *Methode der kleinsten*

Quadrate Damit lässt sich mathematisch die Linie identifizieren, die die besten Daten darstellt, d. h.:

$$Yt = a + b \cdot t \quad ,$$

Dabei ist *Yt* der **Wert** der Zeitreihe zu einem bestimmten Zeitpunkt, a der Wert der Linie, wenn *t* am Ursprung liegt, *b* die Steigung der Linie und *t* einem anderen Wert im gesamten gewählten Zeitraum entspricht.

Die Linie der **kleinsten Quadrate** Y^t entspricht der Schätzung des säkularen Trends *Tt.*

Machine Translated by Google

Berechnung des säkularen Trends

Unter der Annahme, dass die Zeitreihe das Jahr ist, gehen wir von der Idee aus, den Ursprung auf das entsprechende Jahr in der Mitte der Jahre zu verschieben, in denen die Messungen durchgeführt wurden.

A. *Gerade Anzahl von Jahren n*

Sei *j* = *n* ÿ 1/2 und setzen wir *xi* = *ti* ÿ *tj* für *i* = *1, 2,...,n;* die Jahresmitte *tj* ist der neue Ursprung und die Einheit der neuen **Variablen** *x* ist das Jahr.

Daher

$$a\hat{} = \frac{\sum_{i=1}^{N} Yi}{N} \quad \text{und} \quad b\hat{} = \frac{\sum_{i=1}^{N} xiYi}{\sum x2}$$

gibt uns Yˆt = *ˆa+bxˆ* oder Yˆt = (aˆÿbtˆ j)+ *btˆ* , wobei *t* das Jahr ist, ohne den Ursprung zu ändern. B.

Ungerade Anzahl von Jahren n

Sei *j* = n/2 + 1, wobei *tj* der neue Ursprung ist, setze *xi* = 2 · *(ti* ÿ *tj)* + 1 für *i = 1, 2,...,n.*

Daher

$$a\hat{} = \frac{\sum_{i=1}^{N} Yi}{N} \quad \text{und} \quad b\hat{} = \frac{\sum_{i=1}^{N} xiYi}{\sum x2}$$

Gib uns Yˆt = *ˆa + bxˆ* . Die Einheit *x* ist das Halbjahr, gezählt ab dem 1. Januar des Jahres *tj.*

3. *Methode des gleitenden Durchschnitts* Sei Y1, Y2, *Y3,...* eine gegebene Zeitreihe. Konstruieren Sie eine neue Serie:

$$\frac{Y1 + ... + Yn\ ;}{N} \quad \frac{Y2 + ... + Yn+1\ ,...;}{N}$$

$$\frac{Yj + ... + Yn+jÿ1}{N},...$$

deren Einschätzung des säkularen Trends ist hergestellt nach einer der beiden vorherigen Methoden.

4. *Semimeans-Methoden*

Sei Y1, *Y2,...,Yn* eine gegebene Zeitreihe. Bezeichnen Sie mit *k* den ganzzahligen Teil $\frac{n}{2}$. von Wir berechnen:

$$y1 = \frac{Y1 + ... + Yk}{k} \quad \text{Und}$$

$$= n\ ÿ\ \frac{Yk+1 + ... + Yn\ y2}{kk + 1\ x1 =}$$

——— Und

$$x2 = \frac{2\ n + k + 1}{2}$$

und bestimmen Sie die Linie des weltlichen Trends durch zwei Punkte gehen:

(x1; y1) und (x2; y2).

Wenn der säkulare Trend nicht als linear angenommen werden kann, verfeinern wir die Methode der kleinsten Quadrate. Wir sollten es anwenden und *Yt = a + bt* durch eine andere Trendkurve ersetzen, zum Beispiel: • Eine Exponentialkurve: *Yt = a · bt* • Eine modifizierte Exponentialkurve: *Yt = c + a · bt* • Eine logistische Kurve: *Yt = (c + a · bt)ÿ1* • Eine Gompertz-Kurve: *Yt = c · abx*

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Analyse säkularer Trends hat folgende Ziele: • Erstellen eines beschreibenden Modells einer vergangenen Situation.

• Erstellen Sie Projektionen mit konstanter Struktur.

• Eliminieren Sie säkulare Trends, um andere Komponenten der **Zeitreihe zu untersuchen.**

Die vier Methoden zur **Schätzung** des Secu

Dieser Trend hat gewisse Nachteile. Lassen Sie uns jeweils die wichtigsten erwähnen: 1. *Grafische Methode* Zu subjektiv.

S

Machine Translated by Google

2. *Methode der kleinsten Quadrate*

Diese Methode kann nur für einen Zeitraum verwendet werden, in dem die Bewegung vollständig erfolgt
die gleiche Richtung. Als der erste Teil
entspricht einer aufsteigenden Bewegung
und ein anderer Teil zu einer absteigenden Bewegung,
müssen zwei lange Linien erstellt werden, um sich an
diese Daten anzupassen, die sich jeweils darauf beziehen
zu einer Zeit mit einer einzigartigen Richtung.
Diese Zeilen haben einige interessante Eigenschaften:

- Wir finden im selben Graphen die Menge von WerteYiderZeitreiheunddesTrends _ Linie. Die Summe der Abweichungen zwischen die beobachteten Werte *Yi* und die Schätzung des säkularen TrendsY^iwillimmergleich null:

$$\sum_{i=1}^{N} Yi \ddot{y} Y\hat{}i = 0 .$$

- Die Trendlinie minimiert die Summe Abweichungen:

$$\sum_{i=1}^{N} Yi \ddot{y} Y\hat{}i \quad ^2 = \text{minimaler Wert} .$$

3. *Methode des gleitenden Durchschnitts*

Die Daten vom Anfang und Ende des
Ursprüngliche Zeitreihen gehen „verloren“. Gleitende
Durchschnitte können Zyklen und andere Bewegungen
erzeugen, die in den Originaldaten nicht vorhanden sind.
Darüber hinaus sind sie stark davon betroffen
zufällige Ausreißer. Diese Methode ermöglicht es
Richtungsänderungen aufdecken.

4. *Semimeans-Methode*

Auch wenn diese Methode einfach anzuwenden ist, ist sie
kann Ergebnisse ohne Wert liefern. Gedanke
gültig, wenn Daten klassifiziert werden können
in zwei Gruppen einteilen, in denen die Tendenzen liegen
linear, die Methode ist nur anwendbar, wenn
die Tendenz ist global linear oder annähernd linear.
Wenn das Verfahren einfach ist
Die Anwendung ist nicht sehr streng.

BEISPIELE

Lassen Sie uns den säkularen Trend des Jahres abschätzen
Verkäufe der Surfin-Keksfabrik unter Verwendung des
folgende Daten:

Jahresumsatz der Keksfabrik Surfin (in Millionen).
in Euro)

Jahresumsatz (Mio. $)
1975 7.6
1976 6.8
1977 8.4
1978 9.3
1979 12.1
1980 11.9
1981 12.3

Lassen Sie uns die Methode der **kleinsten Quadrate** verwenden :
Verschieben wir den Ursprung in die Mitte der 7
Jahre, also 1978. Wir vervollständigen die folgende
Tabelle:

Jahresverkäufe (Mio. Euro)		Code Jahr		
Xt	Yt	T	tYt	T^2
1975 7.6		ÿ3 ÿ22,2		9
1976 6.8		ÿ2 ÿ13,6		4
1977 8.4		ÿ1	*ÿ8,4*	1
1978 9.3		0	0,0	0
1979 12.1		1	12.1	1
1980 11.9		2	23.8	4
1981 12.3		3	36.9	9
Gesamt 68,4		0	28,0 28	

Setze *ti = Xi* ÿ 1978 für *i = 1, 2,...,* 7. Die
Mitte 1978 ist der neue Ursprung und die Einheit
von *t* ist das Jahr.
Daher:

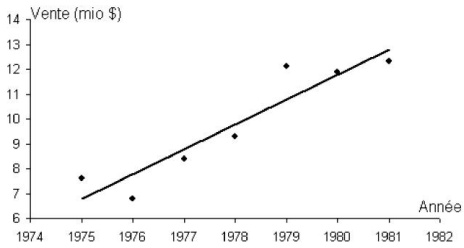
$$a\hat{} = \frac{\sum_{t=\ddot{y}3}^{3} Yt}{7} = \frac{68,4}{7}$$

= 9,7714 (Millionen Euro) und

Machine Translated by Google

$$b\hat{} = \frac{\sum_{t=\ddot{y}3}^{3} tYt}{\sum_{t=\ddot{y}3}^{3} 2t} = \frac{28}{} = 1{,}0 \text{ (Millionen Euro) , } 28$$

was uns $Tt = Y\hat{}t = 9{,}7714 + t$ für die Gleichung der säkularen Trendlinie ergibt .
Das ergibt grafisch:



WEITERLESEN

Zyklische Schwankung
Grafische Darstellung
Kleinsten Quadrate
Gleitender Durchschnitt
Zeitfolgen

VERWEISE

Chatfield, C.: Die Analyse von Zeitreihen.
Eine Einleitung. Chapman'n'Hall (2003)

Halblogarithmisch

Parzelle

Ein halblogarithmisches Diagramm ist eine **grafische Darstellung** einer **Zeitreihe.** Sie wird durch eine arithmetische Skala der Zeit *t* definiert , die auf der Abszissenachse aufgetragen ist, und eine logarithmische Skala, die auf der Ordinatenachse aufgetragen ist, was bedeutet, dass der Logarithmus des beobachteten **Werts** *Yt* in die Ordinate einer arithmetischen Skala übertragen wird.

Im Allgemeinen werden bedruckte halblogarithmische Papierbögen verwendet.

MATHEMATISCHE ASPEKTE

In einem Achsensystem werden ein Ursprung und eine Längeneinheit willkürlich gewählt; Die Abszisse ist rechnerisch in Millimetern unterteilt Papier.

Auf der Ordinatenachse kann die vom Hersteller des halblogarithmischen Papiers gewählte Einheitslänge geändert werden, die dem Abstand zwischen zwei aufeinanderfolgenden Zehnerpotenzen auf der logarithmischen Skala entspricht. Dieser Abstand, der die Längeneinheit der entsprechenden arithmetischen Skala darstellt, wird Modul genannt.

Die am Rand der logarithmischen Skala abgedruckten **Werte** können mit derselben Zahl multipliziert werden; Diese multiplikative Konstante wird so einfach wie möglich gewählt, sodass die erhaltenen Werte leicht für die Transkription der Punkte auf dem Diagramm verwendet werden können. Handelsübliche halblogarithmische Papierbögen haben üblicherweise ein, zwei, drei oder vier Module.

DOMÄNEN UND EINSCHRÄNKUNGEN

Das halblogarithmische Diagramm hat die folgenden Eigenschaften.

• Die Exponentialkurven *Y* werden durch eine Linie dargestellt. Wenn *y(x)* eine Exponentialfunktion ist:

$$y(x) = c \cdot ax \quad ,$$

durch Logarithmusbildung:

$$\log y(x) = \log c + x \cdot \log a$$

wobei $\log y(x) = m \cdot x + hm =$

Wo $\log a$ und $h = \log c$,

wir finden einen linearen Ausdruck in x. • Wenn *a = 1+i,* wobei *i* der jährliche Zinssatz von *Y* und einem anfänglichen Geldbetrag ist

S

Machine Translated by Google

von Y0, dann:

$$Y = Y0 \cdot (1 + i)^{x}$$

Gibt den Geldbetrag nach *x* Jahren an.

Es ist auch möglich, in der Napierian-Basis *(e)* zu arbeiten , da $a^x = e^{x \cdot \ln a}$ und daher $y(x) = c \cdot e^{x \cdot \ln a}$ oder $y(x) = c \cdot e^{r \cdot x}$, wobei $r = \ln a$ ist wird als sofortige Wachstumsrate von y bezeichnet. In diesem Fall ist es das Verhältnis der momentanen Variation $\frac{dy}{dx}$ zum **Wert** von y.

Die Verwendung des halblogarithmischen Diagramms ist in den folgenden Situationen sinnvoll:

- Wenn große Werteunterschiede in der **Variablen** bestehen, **um** ein Verlassen des Diagramms zu vermeiden.
- Wenn man relative Varia machen möchte erscheinen. •

Wenn Kumulationen von Wachstumsraten, die exponentiell erscheinen würden, linear dargestellt werden müssen.

WEITERLESEN

Grafische Darstellung
Zeitfolgen

Serielle Korrelation

Wir nennen serielle Korrelation oder **Autokorrelation** die Abhängigkeit zwischen den Beobachtungen von **Zeitreihen.**

GESCHICHTE

Siehe **Zeitreihen.**

MATHEMATISCHE ASPEKTE

Seien *x1,...,xn n* Beobachtungen einer ZufallsvariablenX , die von der Zeit abhängt (wir sprechen hier von Zeitreihen). Um die Abhängigkeit zwischen Beobachtungen und Zeit zu messen,

Wir verwenden den folgenden Koeffizienten:

$$d = \frac{\sum_{t=2}^{N} (xt - xt-1)^2}{\sum_{t=1}^{N} x_t^2},$$

was die Beziehung zwischen den beiden anzeigt aufeinanderfolgende Beobachtungen. Dieser Koeffizient nimmt Werte zwischen 0 und 4 an. Wenn es nahe bei liegt 0, so dass der Unterschied zwischen zwei aufeinanderfolgenden Beobachtungen minimal ist, spricht man von einer positiven Autokorrelation. Wenn er nahe bei 4 liegt, liegt eine negative Autokorrelation vor: Auf eine Beobachtung mit einem großen Wert folgt tendenziell ein niedriger Wert und umgekehrt. Für den Fall, dass Beobachtungen unabhängig sind

Abhängig von der Zeit hat der Koeffizient *d* einen Wert nahe 2.

Wir können auch die serielle Korrelation messen

der Beobachtungen *x1,...,xn* Berechnung des **Korrelationskoeffizienten** zwischen den

Reihe von *xt* und die von *xt* „auf k bewegt “:

$$rk = \frac{\sum_{t=k+1}^{N} (xt - \bar{x})(xt-k - \bar{x})}{\sum_{t=1}^{N} (xt - \bar{x})^2}$$

und messen Sie auf die gleiche Weise die serielle Korrelation für den Korrelationskoeffizienten

Verbindung zweier verschiedener Serien.

DOMÄNEN UND EINSCHRÄNKUNGEN

Autoregressive Modelle

Wenn eine serielle Korrelation beobachtet wird zwischen den Residuen eines Sim-Modells

B. lineare Regression, können praktische Modelle des Zeitreihenbereichs verwendet werden

Erfolg, zum Beispiel Modelle des Typs

$$Yt = \beta 0 + \beta 1 Xt + \varepsilon t ,$$

Machine Translated by Google

wobei die Fehlerterme *ÿt* nicht unabhängig voneinander sind, sondern durch eine Relation des Typs beschrieben werden

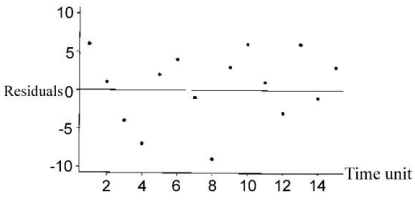
$$ÿt = ÿsÿtÿ1 + ÿt ,$$

wobei *ÿs* durch die serielle Korrelation von *ÿt* geschätzt werden können und wobei die *ÿt* unabhängige Fehler sind.

BEISPIELE

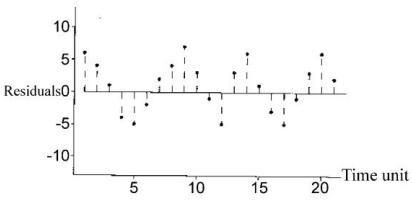
Diagramm der Residuen als Funktionen der Zeit

Diese Art von Diagramm wird insbesondere bei der Analyse von Zeitreihen verwendet, d. h. wenn die Zeit eine erklärende Variable des Modells ist.
Residuen relativ zur Zeit.



Eine grafische Darstellung von Residuen relativ zur Zeit bringt immer den Anschein einer seriellen Korrelation mit sich, positiv oder negativ, wie im Fall der obigen Abbildung. Eine positive oder negative serielle Korrelation impliziert, dass die Fehler nicht unabhängig sind.

Residuen, die eine serielle Korrelation darstellen.



WEITERLESEN

Autokorrelation

Hypothesentest

Zeitfolgen

VERWEISE

Bartlett, MS: Zur theoretischen Spezifikation der Stichprobeneigenschaften von Autocor-bezogenen Zeitreihen. J. Roy. Stat. Soc. Ser. B 8, 27–41 (1946)

Box, GEP, Jenkins, GM: Time Series Analysis: Forecasting and Control (Reihe in TimeSeriesAnalysis). HoldenDay, San Francisco (1970)

Durbin, J., Watson, GS: Testen auf serielle Korrelation in der Regression der kleinsten Quadrate, II. Biometrie **38,** 159–177 (1951)

Zeichentest

Der Vorzeichentest ist ein **nichtparametrischer Test** und kann verwendet werden, um die Hypothese zu testen, dass es keinen Unterschied zwischen der Verteilung von gibt zwei Zufallsvariablen. Wir verdanken ihm seinen Namen die Tatsache, dass anstelle quantitativer Werte „+“- und „ÿ“-Zeichen verwendet werden. Daher ist es auch dann anwendbar, wenn eine Reihe ein Ordnungsmaß hat.

GESCHICHTE

In der sicherlich ersten Veröffentlichung über nichtparametrische Tests untersuchte Arbuthnott, J. (1710) die Liste der in London in einem Zeitraum von 82 Jahren registrierten Geburten. Für jedes Jahr verglich er die Anzahl der geborenen Kinder beiderlei Geschlechts. Er bezeichnete mit „+“ das **Ereignis** „Es werden mehr Jungen als Mädchen geboren“ und mit „ÿ“ das gegenteilige Ereignis. (Es gab keine Gleichheit).
Zu seiner größten Überraschung kam er 82 Mal auf ein „+“-Zeichen und nie auf ein „ÿ“-Zeichen; Dies führte dazu, dass er die **Nullhypothese** einer Gleichheit der Geburten im Verhältnis zum Geschlecht des Kindes ablehnte.

MATHEMATISCHE ASPEKTE

Betrachten Sie *n* Beobachtungspaare (x1, y1), (x2, *y2),...,* *(xn,* yn). Für jedes Paar *(xi,* yi),

Machine Translated by Google

Wir machen den folgenden Vergleich entsprechend der Dimension von *xi* relativ zu *yi :*

„+ wenn *xi* < *yi*

„ÿ wenn *xi* > *yi*

" = wenn *xi* = *yi*

Wir subtrahieren von der Dimension des **Sam**

Geben Sie die Anzahl der erscheinenden „="-Zeichen an Das heißt, wir berücksichtigen nur die Paare, die einen positiven oder negativen Unterschied darstellen. Diese Anzahl Paare bezeichnen wir mit *m* .
Wir zählen dann die Anzahl der „+"-Zeichen, was wir mit T bezeichnen.

Hypothesen

Je nach Test einseitig oder zweiseitig, die Null- und Alternativhypothese
Dem Test entsprechend sind:
A: Zweiseitiger Fall:

H0 : *P* (*X* < Y) = *P* (*X* > Y)

H1 : *P* (*X* < Y) = *P* (*X* > Y)

B: Einseitiger Fall:

H0 : *P* (*X* < Y) ÿ *P* (*X* > Y)

H1 : *P* (*X* < Y) > *P* (*X* > Y)

C: Einseitiger Fall:

H0 : *P* (*X* < Y) ÿ *P* (*X* > Y)

H1 : *P* (*X* < Y) < *P* (*X* > Y)

Im Fall A stellen wir die Nullhypothese (H0) auf. dass die **Wahrscheinlichkeit ,** dass *X* kleiner als *Y* ist ist dasselbe wie die Wahrscheinlichkeit, dass X größer ist als Y.

Im Fall B nehmen wir a priori an, dass die Wahrscheinlichkeit, dass *X* kleiner als *Y* ist , kleiner ist als oder gleich der Wahrscheinlichkeit, dass *X* größer ist als Y.

Schließlich nehmen wir im Fall C a priori an, dass die Wahrscheinlichkeit, dass *X* kleiner als *Y* ist größer oder gleich der Wahrscheinlichkeit, dass *X* ist größer als Y.

Im zweiseitigen Fall ist die Wahrscheinlichkeit, dass *X* kleiner als *Y* ist (da *X* > Y gilt). $\frac{1}{2}$.

Der Vorzeichentest ist ein Sonderfall des **Bino** 1 **Mial-Test** mit $p = 2$, − das heißt, das unter dem **Nullhypothese** H0 wollen wir so viele bekommen „+" als „ÿ".

Entscheidungsregeln

Fall A

Wir verwenden die **Binomialtabelle** und suchen nach die Zahl *X* mit entsprechendem Binomial Wahrscheinlichkeit gleich $2\frac{1}{.}$

Wir sollten in der Tabelle den **Wert** finden, der am nächsten kommt Zu $\frac{ÿ}{2}$, wobei ÿ das **Signifikanzniveau ist.**

Wir lehnen H0 auf der Ebene ÿ ab, wenn:

$$T ÿ tÿ/2 \text{ oder } T ÿ m ÿ tÿ/2 ,$$

mit *T* = Anzahl der „+".

Wenn *m* > 20, können wir die **normale Tabelle** verwenden als Näherung der **Binomialverteilung.**

Wir transformieren **die Statistik** *T* in eine **Zufallsstatistik Variable** *Z,* die einer Standardnormalen folgt Verteilung:

$$Z = \frac{T ÿ ÿ}{ÿ} ,$$

Wo

$$ÿ = m \cdot p = m \quad \frac{1}{2}$$

$$\text{und } ÿ = ÿm \cdot p \cdot q = 2 \quad \frac{1}{-ÿm}$$

sind der **Mittelwert** und die **Standardabweichung** von die Binomialverteilung.
Somit erhalten wir:

$$Z = \frac{T ÿ \quad \frac{1}{2m}}{\frac{1}{2} ÿm} .$$

Machine Translated by Google

Die Näherung des Wertes der Binomialtabelle ist dann:

$$tÿ/2 = 2^{-1} \; m + zÿ/2 \; ÿm,$$

wobei zÿ/2 in der **Normaltabelle** auf der Ebene 2 zu finden ist . Dann ist die Entscheidungsregel ist dasselbe.

Fall B

Wir lehnen H0 auf der Ebene ÿ ab, wenn:

$$T ÿ m ÿ tÿ,$$

Dabei ist tÿ der Wert der entsprechenden Binomialtabelle und das **Signifikanzniveau** ÿ (oder der nächstliegende Wert).

Für $m > 20$ machen wir die Näherung

$$1 \; tÿ = 2^{-} \; m + zÿ \; ÿm,$$

wobei zÿ in der Normaltabelle auf der Ebene ÿ zu finden ist.

Fall C

Wir lehnen H0 auf der Ebene ÿ ab, wenn:

$$T ÿ tÿ,$$

wobei der Wert *von ÿ* derselbe ist wie für Fall B.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Voraussetzungen für die Verwendung des Vorzeichentests sind folgende: 1. Die Paare von Zufallsvariablen (Xi, Yi),i = *1, 2,...,nmüssen* voneinander unabhängig sein.

2. Der Maßstab muss mindestens ordinal sein, das heißt, jedem Paar kann ein „+“, „ÿ“ oder „=“-Zeichen zugeordnet werden.

3. Die Paare *(Xi,* Yi), *i = 1, 2,...,n* müssen eine Logik zwischen ihnen darstellen, das heißt, wenn P „+ > P „ÿ für ein Paar *(Xi,* Yi), dann P „+ > P „ÿ für alle anderen Paare.

Das gleiche Prinzip gilt, wenn P „+ < P „ÿ oder P „+ = P „ÿ .

BEISPIELE

Ein Schokoladenhersteller untersucht eine neue Verpackungsart für eines seiner Produkte. Dazu schlägt er 10 Verbrauchern vor, die alte und neue Verpackung auf einer Skala von 1 bis 6 zu bewerten (wobei 1 „mag ich überhaupt nicht“ und 6 „mag ich sehr“) bedeutet.

Wir führen **einen einseitigen** Test durch (entsprechend Fall C), bei dem die **Nullhypothese, die** wir testen möchten, geschrieben ist:

$$H0 : P „+ ÿ P „ÿ$$

relativ zur **Alternativhypothese:**

$$H1 : P „+ < P „ÿ .$$

Das „+“-Zeichen bedeutet, dass die neue Verpackung der alten vorgezogen wird.

Die erhaltenen Ergebnisse sind:

Konsumer	Rang alt Verpackung	Rang neu Verpackung	Zeichen der Differenz
14	5		+
23	5		+
31	1		=
45	4		
54	4		=
62	6		+
74	6		+
84	5		+
92	3		+
10	1	4	+

Die Tabelle zeigt, dass zwei Verbraucher sagten Es gab keinen Unterschied zwischen dem Neuen und dem alte Verpackung. Es gibt also acht Unterschiede *(m* = 8), bei denen in sieben Fällen die neue Verpackung gegenüber der alten bevorzugt wurde. Somit gilt T = 7 (Anzahl der „+“).

Also lehnen wir H0 auf der Ebene ÿ ab, wenn

$$T ÿ tÿ,$$

S

Machine Translated by Google

wobei tÿ der **Wert** der **Binomialtabelle** ist .

Wir wählen ÿ = 0,05; der Wert von tÿ ist 1 (für $m = 8$ und ÿ = 0,0352). Somit ist $T >$ tÿ , weil $T = 7$ und tÿ = 1. Wir lehnen die H0-Hypothese nicht ab : P „+ ÿ P „ÿ und der Schokoladenhersteller kann daraus schließen, dass Verbraucher im Allgemeinen die neue Verpackung bevorzugen.

Wir nehmen noch einmal das gleiche Beispiel, aber dieses Mal wird die Studie an 100 Verbrauchern durchgeführt. Die angenommenen Ergebnisse sind die folgenden:

83 Personen bevorzugen die neue Verpackung.

7 Personen bevorzugen die alte Verpackung.

10 Leute sind gleichgültig.

Somit haben wir:

$$m = 100 \text{ ÿ } 10 = 90 \quad ,$$
$$T = 83 \ .$$

Der Wert von tÿ kann durch die **Normalverteilung angenähert werden:**

$$1 \text{ tÿ } = 2 \quad m + z\ddot{y}\ \ddot{y}m \ ,$$

wobei zÿ der Wert der **Normaltabelle** für ÿ = 0,05 ist. Das gibt:

$$1 \text{ tÿ } = 2 \quad 90 \text{ ÿ } 1{,}64\ddot{y}\ 90$$
$$= 37{,}22 \ .$$

Ist T größer als tÿ (83 > 37,22), dann wird die Nullhypothese H0: P „+ ÿ P „ÿ nicht verworfen und der Schokoladenhersteller zieht die gleiche Schlussfolgerung wie zuvor.

WEITERLESEN

- **Binomialtabelle**
- **Binomialtest**
- **Hypothesentest**
- **Nichtparametrischer Test**
- **Normalverteilung**

VERWEISE

Arbuthnott, J.: Ein Argument für die göttliche Vorsehung, abgeleitet von der konstanten Regelmäßigkeit, die bei den Geburten beider Geschlechter beobachtet wird. Phi los. Trans. **27,** 186–190 (1710)(3.4).

Signifikanzniveau

Das Signifikanzniveau ist ein **Parameter** eines **Hypothesentests** und sein **Wert** wird vom Benutzer im Voraus festgelegt.

Das mit ÿ bezeichnete Signifikanzniveau eines Hypothesentests ist die **Wahrscheinlichkeit ,** die **Nullhypothese** H0 abzulehnen, wenn sie wahr ist:

$$\ddot{y} = P\{\text{reject H0}|\text{H0 true}\} \ .$$

Das Signifikanzniveau wird auch als Wahrscheinlichkeit eines **Fehlers vom Typ I bezeichnet.**

GESCHICHTE

Im Jahr 1928 diskutierten Jerzy Neyman und Egon Pearson die Probleme im Zusammenhang mit der Frage, ob eine Stichprobe als wahrscheinlich aus einer bestimmten Bevölkerungsgruppe stammend angesehen werden kann oder nicht. Sie identifizierten zwei Arten von Fehlern, die mit möglichen Entscheidungen verbunden sind. Die Wahrscheinlichkeit des Typ-I-Fehlers (Ablehnung der gegebenen Hypothese) entspricht dem Signifikanzniveau des Tests. Siehe **Hypothesentests.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Statistiker sind besorgt über die Möglichkeit eines **Fehlers ,** indem sie die **Nullhypothese** verwerfen, wenn sie wahr ist, und konstruieren einen Hypothesentest so, dass die **Wahrscheinlichkeit** eines Fehlers dieser Art einen im Voraus festgelegten festen Wert nicht übe
Dieser **Wert** entspricht dem Signifikanzniveau des Tests.

Machine Translated by Google

Beachten Sie, dass Statistiker, die einen **Hypothesentest** durchführen, sich mit der zweiten Art auseinandersetzen müssen **Fehler,** Typ-II-Fehler genannt . Es findet statt wenn die Nullhypothese nicht abgelehnt wird in Wirklichkeit ist es falsch.

Um das globale Risiko von Fehlern bei der Entscheidungsfindung zu verringern, reicht es nicht aus, das zu verringern Signifikanzniveau, aber ein Kompromiss sollte zwischen dem Signifikanzniveau und liegen die Wahrscheinlichkeit eines Fehlers zweiter Art.

Eine Methode basiert auf der Untersuchung der Testleistung.

BEISPIELE

Im folgenden Beispiel veranschaulichen wir das Einfluss der Wahl der Bedeutung Ebene bei der Durchführung eines **Hypothesentests.**

Betrachten Sie einen rechts **einseitigen** Test **Mittelwertÿ** einer **Grundgesamtheit ,** deren **Varianz** ÿ2 entspricht 36.

Stellen wir folgende Hypothese auf:

Nullhypothese H0 : ÿ = 30

Alternativhypothese H1 : ÿ > 30 .

Nehmen wir an, dass der Mittelwert x¯ einer **Stichprobe** ist der Dimension *n* = 100 aus dieser Grundgesamtheit ergibt 31.

Fall 1

Wählen Sie ein Signifikanzniveauÿ = 5%.Wir finden der **kritische Wert** von zÿ = 1,65 in der Normaltabelle. Die Obergrenze der **Akzeptanz Region** wird bestimmt durch:

$$ÿx¯ + zÿ \cdot ÿx¯,$$

Dabei ist ÿx¯ der Mittelwert der **Stichprobenverteilung** der Mittelwerte und *ÿx¯* der **Standard** Fehler (oder die **Standardabweichung** der Stichprobenverteilung der Mittelwerte):

$$ÿx¯ = \frac{ÿ}{\sqrt{ÿn}} \cdot$$

Diese Obergrenze beträgt:

$$30 + 1{,}65 \cdot \frac{6}{\sqrt{ÿ100}} = 30{,}99 .$$

Der **Ablehnungsbereich** der Nullhypothese entspricht dem **Intervall** *[30,99;* ÿ[und der Akzeptanzbereich zum Intervall] ÿ ÿ; 30,99[.

Fall 2

Wenn wir ein kleineres Signifikanzniveau festlegen, z Beispiel ÿ = 1 %, der Wert *von* zÿ in der **Normaltabelle** beträgt 2,33 und die Obergrenze von Der Akzeptanzbereich beträgt nun:

$$ÿx¯ + zÿ \cdot ÿx¯ = 30 + 2{,}33 \cdot \frac{6}{\sqrt{ÿ100}}$$

$$= 31{,}4 .$$

Der Ablehnungsbereich der Nullhypothese entspricht dem Intervall *[31,4;* ÿ[und der Akzeptanzbereich zum Intervall] ÿ ÿ; 31,4[.

Aus den beobachteten Mittelwerten x¯ = 31 können wir Sehen Sie, dass wir im ersten Fall (ÿ = 5%) müssen Lehnen Sie die Nullhypothese ab, da der Stichprobenmittelwert (31) im Ablehnungsbereich liegt. In Im zweiten Fall (ÿ = 1%) müssen wir das machen Umgekehrte Entscheidung: Die Nullhypothese kann aufgrund des Stichprobenmittelwerts nicht abgelehnt werden liegt im Akzeptanzbereich.

Wir sehen, dass die Wahl des Signifikanzniveaus einen direkten Einfluss auf die Entscheidung darüber hat die Nullhypothese und infolgedessen weiter der Entscheidungsprozess. Ein kleineres Signifikanzniveau (und damit eine kleinere **Wahrscheinlichkeit** eines **Fehlers erster Art)** schränkt die ein Ablehnungsbereich und erhöht die Akzeptanz Bereich der Nullhypothese.

Es besteht ein geringeres **Fehlerrisiko** eine kleinere Präzision.

S

Machine Translated by Google

WEITERLESEN

Hypothesentest

Fehler vom Typ I

VERWEISE

Neyman, J., Pearson, ES: Über die Verwendung und Interpretation bestimmter Testkriterien für Zwecke der statistischen Inferenz Teil I (1928). Nachdruck von Cambridge University Press im Jahr 1967

Einfache Indexnummer

Eine einfache **Indexzahl** ist das Verhältnis von zwei Werte, die dieselbe **Variable darstellen,** gemessen in zwei verschiedenen Situationen oder in zwei verschiedenen Situationen verschiedene Perioden. Zum Beispiel ein einfacher Index Die Preiszahl gibt die relative Schwankung des Preises zwischen der aktuellen Periode an und einen Bezugszeitraum. Die am häufigsten verwendeten einfachen Indexzahlen sind die von Preis, Menge und Wert.

GESCHICHTE

Siehe **Index.**

MATHEMATISCHE ASPEKTE

Die Indexzahl In/0, die in Bezug auf **eine Variable Gin** -Situation repräsentativ ist auf dieselbe Variable in Situation 0 (Referenz Situation) ist definiert durch:

$$In/0 = \frac{Gn}{G0},$$

Dabei ist *Gn* der **Wert** der **Variablen** *G* in Situation *n* und G0 der Wert der Variablen *G* in Situation 0.

Im Allgemeinen ist eine einfache Indexnummer ausgedrückt in Basis 100 in der Referenzsituation 0:

$$In/0 = \frac{Gn}{G0} \cdot 100 .$$

Eigenschaften einfacher Indexzahlen

- *Identität:* Wenn zwei verglichene Situationen (bzw zwei Perioden) identisch sind, ist der Wert der Indexnummer ist gleich 1 (oder 100):

$$In/n = I0/0 = 1 .$$

- Eine einfache Indexzahl ist *umkehrbar:*

$$In/0 = \frac{1}{I0/n} .$$

- Eine einfache Indexnummer ist *übertragbar.* Betrachten Sie zwei IndexzahlenI1/0 und I2/0 Darstellung zweier Situationen, 1 und 2, mit in Bezug auf die gleiche Grundsituation 0. Wenn wir Betrachten Sie die Indexzahl I2/1, wir können Sag das:

$$I2/0 = I2/1 \cdot I1/0 .$$

Beachten Sie, dass die Übertragbarkeit (auch genannt Zirkularität) impliziert Reversibilität:

Wenn $I0/1 \cdot I1/0 = I0/0 = 1$, dann $I0/1 = \frac{1}{I1/0}$.

DOMÄNEN UND EINSCHRÄNKUNGEN

In den Wirtschaftswissenschaften kommt dieser einfache **Index** sehr selten vor **Zahlen** sind an sich ausnutzbar; Die Die darin enthaltenen Informationen sind relativ begrenzt. Dennoch haben sie eine grundlegende Bedeutung bei der Konstruktion **zusammengesetzter Indexzahlen.**

BEISPIELE

Betrachten Sie die folgende Tabelle, die Folgendes darstellt die Preise eines bestimmten Artikels und die in verschiedenen Zeiträumen verkauften Mengen:

Preis		Menge	
Periode 1	Periode 2	Periode 1	Periode 2
P0 = 50	*Pn* = 70	Q0 = 25	*Qn* = 20

Machine Translated by Google

Wenn *Pn* der Preis der Zinsperiode ist und
P0 ist der Preis des Referenzzeitraums, wir erhalten:

$$In/0 = \frac{Pn}{P0} \cdot 100 = \frac{70}{50} \cdot 100 = 140 \quad ,$$

Das bedeutet, dass bei der Basis 100 in Periode 1 der Preisindex des Artikels in Periode 2 140 beträgt. Der Preis des Artikels ist zwischen dem Referenzzeitraum und dem tatsächlichen Zeitraum um 40 % (140 ÿ 100) gestiegen.
Wir können auch eine einfache Indexzahl der Menge berechnen:

$$In/0 = \frac{Qn}{Q0} \cdot 100 = \frac{20}{25} \cdot 100 = 80 \quad .$$

Die verkaufte Menge hat sich daher zwischen dem Referenzzeitraum und dem tatsächlichen Zeitraum um 20 % (100–80) verringert.
Mit diesen Zahlen können wir auch eine Wertindexzahl berechnen, wobei der Wert als Preis multipliziert mit der Menge definiert ist:

$$In/0 = \frac{Pn \cdot Qn}{P0 \cdot Q0} = \frac{70 \cdot 20}{50 \cdot 25} \cdot 100 = 112 \quad .$$

Dort liegt der Wert des betrachteten Gegenstandes
Daher stieg die Zahl zwischen den beiden betrachteten Zeiträumen um 12 % (112–100).

WEITERLESEN

Zusammengesetzte Indexnummer
Fisher-Index
Indexnummer
Laspeyres-Index
Paasche-Index

Einfach linear Rückschritt

Die einfache lineare Regression ist eine Analyse **der Regression ,** bei der die **abhängige Variable** vorliegt *Y* hängt linear von einer einzelnen **unabhängigen Variablen** X ab.

Die einfache lineare Regression zielt nicht nur darauf ab, die Regressionsfunktion relativ zum gewählten **Modell** zu schätzen , sondern auch die Zuverlässigkeit der erhaltenen Schätzungen zu testen.

GESCHICHTE

Siehe **Analyse der Regression.**

MATHEMATISCHE ASPEKTE

Das **Modell** der einfachen linearen Regression hat die Form:

$$Y = \ddot{y}0 + \ddot{y}1X + \ddot{y} \quad ,$$

Dabei ist Y **die abhängige Variable** (oder erklärte Variable), *X* die **unabhängige Variable** (oder erklärende Variable), ÿ der Term des zufälligen nicht beobachtbaren **Fehlers** und ÿ0 und ÿ1 die zu schätzenden Parameter.
Wenn wir eine Menge von *n* Beobachtungen (X1, Y1), ..., *(Xn,* Yn) haben, wobei *Yi* linear vom entsprechenden *Xi abhängig ist,* können wir schreiben

$$Yi = \ddot{y}0 + \ddot{y}1Xi + \ddot{y}i \, , \qquad i = 1,...,n.$$

Das Problem bei der **Regressionsanalyse** besteht darin, die Parameter ÿ0 und ÿ1 zu schätzen und die Werte ÿˆ 0 und ÿˆ 1 so zu wählen, dass der **Abstand** zwischen *Yi* und (ÿ0+ ÿ1 · *Xi)* minimal ist. Wir hätten sollen:

$$\ddot{y}i = Yi \ddot{y} \ddot{y}0 \ddot{y} \ddot{y}1 \cdot Xi$$

klein für alle *i = 1,...,n.* Um dies zu erreichen, können wir zwischen vielen Kriterien wählen: 1.
$\min_{\ddot{y}0,\ddot{y}1} \max_{i} |\ddot{y}i|$

2. $\min_{\ddot{y}0,\ddot{y}1} \sum_{i=1}^{N} |\ddot{y}i|$

3. $\min_{\ddot{y}0,\ddot{y}1} \sum_{i=1}^{n} \ddot{y}2$

S

Machine Translated by Google

Die am häufigsten verwendete Methode zur **Schätzung** von

Die Parameter ÿ0 und ÿ1 sind die dritten, die sogenannten **kleinsten Quadrate.** Ziel ist es, die Summe der Fehlerquadrate zu minimieren.

Die Schätzung der Parameter mit der Methode der kleinsten Quadrate ergibt die folgenden Schätzer ÿˆ 0 und ÿˆ 1:

$$\ddot{y}\hat{}\ 1 = \frac{\sum_{i=1}^{N} Xi \ddot{y} X^{-} Y\hat{}i \ddot{y} Y^{-}}{\sum_{i=1}^{N} Xi \ddot{y} X^{-}{}^{2}}$$

$$\ddot{y}\hat{}\ 0 = Y^{-} \ddot{y} \ddot{y}\hat{}\ 1X .$$

Wir können dann schreiben:

$$Y\hat{}i = \ddot{y}\hat{}\ 0 + \ddot{y}\hat{}\ 1Xi ,$$

wobei Y^i der geschätzte Wert *von Yi für* ein gegebenes *Xi* ist , wenn ÿˆ 0 und ÿˆ 1 bekannt sind.

Maß für die Zuverlässigkeit der Schätzung von

Wir haben eine Schätzung des Wertes berechnet von Y, oder Yˆ, mit Hilfe der Methode der kleinsten Quadrate, wobei unsere Berechnung auf einem linearen Modell basiert, um die Beziehung, die *Y* in Beziehung setzt, zu übersetzen zu X. Aber wie weit können wir diesem Modell vertrauen?

Um diese Frage zu beantworten, ist es sinnvoll, eine **Varianzanalyse** durchzuführen und die Hypothese anhand der Parameter ÿ0 und ÿ1 der Regressionsgeraden zu testen. Um diese Tests durchzuführen, müssen wir die folgenden Annahmen treffen: • Für jeden Wert von X ist *Y*

eine **Zufallsvariable,** die gemäß der **Normalverteilung verteilt ist.**

• Die **Varianz** von *Y* ist für alle X gleich ; es ist gleich ÿ2 (unbekannt). • Die unterschiedlichen Beobachtungen zu *Y* sind unabhängig voneinander abhängig, aber durch die Werte von X bedingt.

Varianzanalyse

Die Tabelle der Varianzanalyse, die wir erstellen müssen, ist die folgende:

Varianzanalyse

Quelle von variation	Grad freidom	Die Summe von Quadrate	Bedeutung von Quadrate
Regression	1	$\sum_{i=1}^{N} Y\hat{}\ i \ddot{y} Y^{-}{}^{2}$	$\sum_{i=1}^{n} Y\hat{}\ i \ddot{y} Y^{-}{}^{2}$
Resi Dual	$n \ddot{y} 2$	$\sum_{i=1}^{N} Y\hat{}\ i \ddot{y} Y^{-}{}^{2}$	$\frac{\sum_{i=1}^{N} Y\hat{}\ iÿY^{-}{}^{2}}{nÿ2}$
Insgesamt	$n \ddot{y} 1$	$\sum_{i=1}^{N} Yi \ddot{y} Y^{-}{}^{2}$	

Wenn das Modell korrekt ist, dann

$$S2 = \frac{\sum_{i=1}^{N} Y\hat{}i \ddot{y} Y^{-}{}^{2}}{n \ddot{y} 2}$$

ist ein erwartungstreuer Schätzer von ÿ2.

Die Varianzanalyse ermöglicht es uns, die **Nullhypothese zu testen:**

$$H0 : \ddot{y}1 = 0$$

gegen die **Alternativhypothese:**

$$H1 : \ddot{y}1 = 0$$

Berechnung der **Statistik:**

$$F = \frac{EMSE}{RMSE} = \frac{EMSE}{S2} .$$

Diese Statistik *F* muss mit dem Wert *Fÿ,1,nÿ2* der **Fisher-Tabelle verglichen werden,** wobei ÿ die **Signifikanz** des Tests ist.

ÿ

Wenn *F* ÿ *Fÿ,1,nÿ2* , dann akzeptieren wir H0.

Wenn *F* > *Fÿ,1,nÿ2* , dann lehnen wir H0 für H1 ab.

Das **Bestimmtheitsmaß** R2 wird wie folgt berechnet: ÿˆ **XY** ÿ **nY**ˆ 2

$$R2 = \frac{ESS}{TSS} = \frac{}{\mathbf{YY} \ddot{y} \mathbf{nY}\hat{}\ 2} ,$$

Dabei ist ESS die Summe der Quadrate der Regression und *TSS* die Gesamtsumme der Quadrate.

Machine Translated by Google

Hypothese über die Steigung

ÿ1 Im Fall einer einfachen linearen Regression ermöglicht die Statistik *F* das Testen einer **Hypothese** unter Verwendung der Parameter der Regressionsgleichung oder ÿ1, der Steigung der Linie. Eine andere Möglichkeit, denselben Test durchzuführen, ist wie folgt:

Wenn H0 wahr ist (ÿ1 = 0), dann ist die Statistik t:

$$t = \frac{\ddot{y}\hat{}}{1}$$

Sÿˆ 1 folgt einer **Student-Verteilung** mit (nÿ2) Freiheitsgraden. ist die

Sÿˆ **Standardabweichung von ÿˆ1,** Schätzung 1 ed aus der **Stichprobe:**

$$S_{\ddot{y}\hat{}\,1} = \frac{S2}{\sqrt{\sum_{i=1}^{N} Xi \ddot{y} X^{-}\,{}^{2}}}.$$

Die Statistik *t* muss mit dem Wert der **Student-**

ÿ2nÿ2 ist **Tabelle verglichen werden,** wobei ÿ das *t* Signifikanzniveau und *n* ÿ 2 die Anzahl der Freiheitsgrade. Die Entscheidungsregel lautet wie folgt:

/*t*/ ÿ *t* ÿ ,nÿ2
2

ÿ wir akzeptieren H0 (ÿ1 = 0).

/*t*/ > *t* ÿ ,nÿ2
2

ÿ wir lehnen H0 für H1 ab (ÿ1 = 0).

Es lässt sich zeigen, dass im Fall einer einfachen linearen Regression t^2 gleich F.
Die Statistik *t* ermöglicht die Berechnung einer **Konfidenz Intervall** für ÿ1.

Hypothese zur Ordinate im Ursprung ÿ0

Auf ähnliche Weise können wir ein Konfidenzintervall für ÿ0 konstruieren und die Hypothese testen:

H0 : ÿ0 = 0

H1 : ÿ0 = 0

Statistik t:

$$t = \frac{\ddot{y}\hat{}\,0}{S_{\ddot{y}\hat{}\,0}}$$

folgt ebenfalls einer Student-Verteilung mit *(nÿ* 2) Freiheitsgraden.
Die geschätzte Standardabweichung von ÿˆ 0, bezeichnet durch $S_{\ddot{y}\hat{}\,0}$ ist definiert durch:

$$S_{\ddot{y}\hat{}\,0} = \frac{S2\sqrt{\sum_{i=1}^{N} X2}}{\sqrt{N\sum_{i=1}^{N} Xi \ddot{y} X^{-}\,{}^{2}}}.$$

Die Entscheidungsregel lautet wie folgt:

/*t*/ ÿ *t* ÿ ,nÿ2
2

ÿ wir akzeptieren H0 (ÿ0 = 0).

/*t*/ > *t* ÿ ,nÿ2
2

ÿ wir lehnen H0 für H1 ab (ÿ0 = 0).

erhalten wird aus der Student-Tabelle für *t* ÿ ,nÿ2 2 (nÿ2) Freiheitsgrade und ein Signifikanzniveau ÿ.

Das **Bestimmtheitsmaß** R2 wird wie folgt berechnet:

$$R2 = \frac{ESS}{TSS} = \frac{\hat{y} XY \ddot{y} nY^{-}2}{YY \ddot{y} nY^{-}2},$$

Dabei ist ESS die Summe der Quadrate der Regression und *TSS* die Gesamtsumme der Quadrate.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die einfache lineare Regression ist ein Sonderfall der **multiplen linearen Regression.** Der in der multiplen linearen Regression offenbarte Matrixansatz **(Modell** mit vielen unabhängigen Variablen) gilt auch für den besonderen Fall, dass wir nur eine unabhängige Variable haben.

Siehe **Analyse der Regression.**

BEISPIELE

Die folgende Tabelle stellt das Bruttosozialprodukt (BSP) und die Nachfrage dar

S

Machine Translated by Google

Grundnahrungsmittel für den Zeitraum 1969 bis 1980 Länder.

Jahr	BSP	Nachfrage nach Grundnahrungsmitteln
	X	Y
1969	50	6
1970	52	8
1971	55	9
1972	59	10
1973	57	8
1974	58	10
1975	62	12
1976	65	9
1977	68	11
1978	69	10
1979	70	11
1980	72	14

Wir wollen den Bedarf an Grundnahrungsmitteln abschätzen
Abhängig vom GN Pac entsprechend dem Mod **el:**

$$Yi = ÿ0 + ÿ1Xi + ÿi, \qquad i = 1,...,n.$$

Die **Schätzung** der Parameter ÿ0 und ÿ1 durch Die Methode der **kleinsten Quadrate** liefert uns die folgenden Schätzer:

$$ÿ\hat{} 1 = \frac{\sum_{i=1}^{12} Xi ÿ X^{-} Yi ÿ Y^{-}}{\sum_{i=1}^{12} Xi ÿ X^{-}{}^{2}} = 0,226$$

$$ÿ\hat{} 0 = Y^{-} ÿ ÿ\hat{} 1X^{-} = ÿ4,04 .$$

Die geschätzte Linie kann wie folgt geschrieben werden:

$$Y\hat{} = ÿ4,04 + 0,226X .$$

Varianzanalyse

Wir berechnen die Freiheitsgrade als
sowie die Summe der Quadrate und der Mittelwert von Quadrate, um die Tabelle der zu erstellen

Varianzanalyse:

$$dlreg = 1 \quad ,$$

$$dlres = n ÿ 2 = 10 \quad ,$$

$$dltot = n ÿ 1 = 11 ;$$

$$ESS = \sum_{i=1}^{12} Y\hat{}i ÿ Y^{-}{}^{2}$$

$$= (7,20 ÿ 9,833)^{2} + \cdots + (12,23 ÿ 9,833)^{2}$$

$$= 30,457 ;$$

$$RSS = \sum_{i=1}^{12} Yi ÿ Y\hat{}i{}^{2}$$

$$= (6 ÿ 7,20)^{2} + \cdots + (14 ÿ 12,23)^{2}$$

$$= 17,21 ;$$

$$TSS = \sum_{i=1}^{12} Yi ÿ Y^{-}{}^{2}$$

$$= (6 ÿ 9,833)^{2} + \cdots + (14 ÿ 9,833)^{2}$$

$$= 47,667 ;$$

$$EMSE = = \frac{ESS}{dlreg} = \frac{30,457}{}$$

$$S2 = RMSE = = \frac{RSS}{dlres} = 1,721 .$$

Somit haben wir die folgende Tabelle:

Varianzanalyse

Quelle von var iation	Abschlüsse von kostenlos dom	Die Summe von Quadrate	Bedeutung von Quadrate	F
Betreff gres sion	1	30.457	30.457	17.687
Resi Dual	10	17.211	1.721	
Gesamt	11	47.667		

Machine Translated by Google

Bei einem **Signifikanzniveau** von ÿ = 5 % finden wir in der **Fisher-Tabelle** den Wert

$$F_{ÿ,1,nÿ2} = F_{0,05,1,10} = 4,96 .$$

Aufgrund von $F > F_{0,05,1,10}$ lehnen wir die **Nullhypothese** H0 ab: ÿ1 = 0, was bedeutet, dass ÿ1 signifikant von Null verschieden ist.
Zur Berechnung des **Determina-Koeffizienten**

Für dieses Beispiel genügt ein Blick auf die Tabelle der Varianzanalyse, da sie alle erforderlichen Elemente enthält:

$$R2 = \frac{ESS}{TSS} = \frac{30.457}{47.667}$$
$$= 0,6390 = 63,90\ \%$$

Wir können daraus schließen, dass nach dem gewählten Modell 63,90 % der Variation der Nachfrage nach Grundnahrungsmitteln durch die Variation des BSP erklärt werden können.
Es ist offensichtlich, dass der R2-Wert 100 nicht überschreiten darf; Der Wert 63,90 ist groß genug, liegt aber nicht nah genug an 100, um einen davon abzuhalten, das Modell zu verbessern.
Dies kann bedeuten:

- Dass außer dem BSP auch andere Variablen für eine Flosse berücksichtigt werden sollten er Bestimmung der Funktion des Nachfrage nach Grundnahrungsmitteln, daher erklärt das BSP nur einen Teil der Variation.
- Wir sollten ein anderes Modell testen (eines ohne konstanten Term, ein nichtlineares Modell usw.).

Durch Hypothesentests der Parameter können wir feststellen, ob sie sich signifikant von Null unterscheiden.

Hypothese zur Steigung ÿ1

In der Tabelle der Varianzanalyse haben wir:

$$S2 = RMSE = 1,721 ,$$

Dies ermöglicht die Berechnung **des Standarddevi**

1: **Kation** $S_{ÿ\hat{}\ 1}$ von ÿ̂

$$S_{ÿ\hat{}\ 1} = \frac{S2}{\sqrt{\sum_{i=1}^{12} Xi ÿ \bar{X}^2}}$$
$$= \frac{1.721}{\sqrt{(50 ÿ 61,416)^2 + \cdots + (72 ÿ 61,416)^2}}$$
$$= \frac{1.721}{\sqrt{596,92}} = 0,054 .$$

Wir können die Statistik berechnen:

$$t = \frac{ÿ\hat{}\ 1}{S_{ÿ\hat{}\ 1}} = \frac{0,226}{0,054} = 4,19 .$$

Wenn wir ein Signifikanzniveau von ÿ = 5 % wählen, erhalten wir:

$$ÿ_{2nÿ2} = t_{0,025,10} = 2,228 . t$$

Wenn wir diesen **Wert** von *t* mit dem Wert in der Tabelle vergleichen, erhalten wir:

$$|t| > t ÿ 2nÿ2 ,$$

was darauf hinweist, dass die Nullhypothese:

$$H0 : ÿ1 = 0$$

muss für die **Alternativhypothese verworfen werden:**

$$H1 : ÿ1 = 0 .$$

Wir kommen zu dem Schluss, dass eine lineare Beziehung besteht zwischen dem BSP und der Nachfrage nach Grundnahrungsmitteln. Wenn H0 akzeptiert würde, würde das bedeuten, dass es keinen linearen Zusammenhang zwischen diesen gäbe zwei Variablen.

Bemerkung: Das können wir sehen

$$F = t^2,$$
$$17,696 = 4,192$$

(ohne Berücksichtigung von Rundheitsfehlern).
Folglich können wir indifferently Test *t* oder Test *F* verwenden, um die Hypothese H0: ÿ1 = 0 zu testen .

S

Machine Translated by Google

Die Hypothese zur Ordinate im Ursprung ÿ0

Die **Standardabweichung** Sÿˆ 0 von ÿˆ 0 lautet:

$$S_{\hat{y}0} = \frac{\sqrt{\frac{12}{S2 \sum_{i=1}^{12} \bar{X}^2}}}{\sqrt{N \sum_{i=1}^{12} Xi \ddot{y} \bar{X}^2}} .$$

$$= \frac{\sqrt{1{,}721\ 502 + \cdots + 722\ (12 \cdot 596{,}92)}}{\sqrt{1{,}721 \cdot 45861}} = \frac{}{7163.04} = 3{,}31 .$$

Wir können die Statistik berechnen :

$$t = \frac{\ddot{y}\hat{}\ 0}{S\ddot{y}\hat{}\ 0} = \frac{\ddot{y}4{,}04}{3.31} = \ddot{y}1{,}22 .$$

Wenn wir diesen **Wert** von *t* mit dem Wert der **Student-Tabelle** t0,025,10 = 2,228 vergleichen, erhalten wir:

$$|t| < t\ \ddot{y}\ _{2}n\ddot{y}2 ,$$

was auf die Nullhypothese hinweist:

$$H0 : \ddot{y}0 = 0$$

muss auf dem Signifikanzniveau ÿ = 5 % akzeptiert werden. Wir kommen zu dem Schluss, dass sich der Wert von ÿ0 nicht wesentlich von Null unterscheidet und die Gerade durch den Ursprung verläuft.

Neues Modell: Regression Passing Through

Herkunft

Das neue Modell ist das Folgende:

$$Yi = \ddot{y}1Xi + \ddot{y}i , \qquad ich = 1,..., 12 \quad .$$

Wenn wir den Parameter ÿ1 nach der Methode der kleinsten Quadrate schätzen, erhalten wir:

$$\sum_{i=1}^{12} \ddot{y}2 = \sum_{i=1}^{12} (Yi \ddot{y}\ \ddot{y}1Xi)^2 .$$

Wenn wir die Ableitung nach ÿ1 auf Null setzen , erhalten wir:

$$\frac{\sum_{i=1}^{12} \ddot{y}\ \ddot{y}2}{\ddot{y}\ddot{y}1} = \ddot{y}2 \sum_{i=1}^{12} Xi\ (Yi \ddot{y}\ \ddot{y}1Xi) = 0$$

oder

$$\sum_{i=1}^{12} (XiYi) \ddot{y}\ \ddot{y}1 \sum_{i=1}^{12} X2 = 0 .$$

Der Wert von ÿ1 , der die Gleichung des Schätzers ÿˆ 1 von ÿ1 erfüllt , ist:

$$\ddot{y}\hat{}\ 1 = \frac{\sum_{i=1}^{12} (XiYi)}{\sum_{i=1}^{12} X2} = \frac{7382}{45861} = 0{,}161 .$$

Folglich wird die neue Regressionsgerade beschrieben durch:

$$Y\hat{}i = \ddot{y}\hat{}\ 1Xi$$

$$Y\hat{}i = 0{,}161Xi .$$

Die Tabelle der Varianzanalyse relativ zu diesem neuen Modell ist die folgende:

Varianzanalyse

Quelle von variation	Grade von Freiheit	Die Summe von Quadrate	Bedeutung von Quadrate	F
Betreff gression	1	1188,2	1188,2	660,11
Resi Dual	11	19.8	1.8	
Insgesamt	12	1208.0		

Beachten Sie, dass in einem Modell ohne Konstante die Anzahl der Freiheitsgrade für die Gesamtvariation gleich der Anzahl der Beobachtungen n ist. Die Anzahl der Freiheitsgrade für die Restvariation beträgt *n* ÿ 1.

Machine Translated by Google

Somit können wir R2 bestimmen:

$$R2 = \frac{ESS}{TSS} = \frac{1188.2}{1208.0} = 0{,}9836$$

$$= 98{,}36\ \%\ .$$

Wir haben beim ersten Modell den Wert gefunden

ue R2 = 63,90 %. Wir können sagen, dass dies neu ist Modell ohne KonstanteTermismusViel besser

als der vorherige.

WEITERLESEN

- **Analyse von Residuen**
- **Bestimmtheitsmaß**
- **Korrelationskoeffizient**
- **Kleinsten Quadrate**
- **Multiple lineare Regression**
- **Normale Gleichungen**
- **Regressionsanalyse**
- **Restwert**

VERWEISE

Seber, GAF (1977) Lineare Regression
Analyse. Wiley, New York

Einfache Zufallsstichprobe

Eine einfache Zufallsstichprobe ist eine **Stichprobe** Methode, bei der man *n* Einheiten wählt unter den *N* Einheiten einer Population in einer solchen eine Art und Weise, dass jeder der C_N^n mögliche **Proben** hat die gleiche **Wahrscheinlichkeit ,** ausgewählt zu werden.

GESCHICHTE

Siehe **Probenahme.**

MATHEMATISCHE ASPEKTE

Um eine einfache Zufallsstichprobe **der** Größe zu erhalten n, die Individuen einer **Population** stehen an erster Stelle nummeriert von 1 bis N; dann sind *n* Zahlen zwischen 1 und N gezeichnet. Die Zeichnungen sind erfolgt grundsätzlich ersatzlos.

Es können auch Tabellen mit **Zufallszahlen** verwendet werden um eine einfache Zufallsstichprobe zu erhalten. Das Ziel einer einfachen Zufallsstichprobe **soll** eine liefern **Schätzung** ohne **Verzerrung** des **Mittelwerts** und der **Varianz** der **Grundgesamtheit.** In der Tat, wenn wir mit y1, *y2,...,yN* die Merkmale einer Population bezeichnen und mit *y1,...,yn* die entsprechende Werte im einfachen Zufall Stichprobe erhalten wir folgende Ergebnisse:

	Bevölkerung:	Probe:
Gesamt	$T = \sum_{i=1}^{N} yi$	$t = \sum_{i=1}^{N} yi$
Mittelwert	$Y^- = \frac{T}{N}$	$y^- = \frac{T}{N}$
Varianz	$\sigma^2 = \frac{\sum_{i=1}^{N} (yi - Y^-)^2}{N - 1}$	$s2 = \frac{\sum_{i=1}^{N} (yi - y^-)^2}{n - 1}$

DOMÄNEN UND EINSCHRÄNKUNGEN

Eine einfache Zufallsstichprobe kann durchgeführt werden oder ersatzlos. Erfolgt eine Zeichnung mit Ersetzung, bleibt die **Grundgesamtheit** immer gleich. Die **Probe**

ist daher durch eine Reihe unabhängiger **Zufallsvariablen** gekennzeichnet , die identisch verteilt sind. Ob die **Grundgesamtheit** ausreichend groß ist und ob die Größe der **Stichprobe** relativ klein im Verhältnis zur Bevölkerung ist, kann man davon ausgehen, dass es sich um einen Zufall handelt Variablen sind **unabhängig,** auch wenn die Zeichnungen ohne Ersatz erstellt werden. Dann ist es so Es ist möglich, bestimmte Eigenschaften abzuschätzen die **Bevölkerung ,** indem man sie aus ermittelt **dieses Beispiel.** Eine einfache Zufallsstichprobe ist die Grundlage der Probenahmetheorie .

BEISPIELE

Betrachten wir eine **Population** von 658 Personen. Um eine einfache Zufallsstichprobe zu bilden , haben wir

S

Machine Translated by Google

Ordnen Sie jeder Person eine dreistellige Zahl zu:
001, *002, ...,* 658. Mit einer **Zufallszahlentabelle**
erhalten wir dann eine **Stichprobe** von
zehn Personen. Wir wählen zufällig einen ersten aus
Ziffer und eine Leserichtung. Wir lesen dann
die ersten drei Ziffern, bis wir zehn verschiedene
Zahlen zwischen 001 und 658 zur Definition erhalten
unsere probe.

WEITERLESEN

- **Cluster-Sampling**
- **Einschätzung**
- **Schätzer**
- **Probenahme**
- **Stichprobenverteilung**
- **Geschichtete Stichprobe**
- **Systematische Probenahme**

VERWEISE

Cochran, WG: Sampling Techniques, 2
Hrsg. Wiley, New York (1963)

Simulation

Simulation ist eine Methode zur Analyse,
Entwerfen und Betreiben komplexer Systeme.
Bei der Simulation wird ein Modell entworfen
ein System und die Durchführung von Experimenten damit
wie es voranschreitet.
Das grundlegende Problem der Simulation ist
beim Bau des künstlichen SAM
Werte relativ zu einer statistisch bekannten Verteilung.
Diese Verteilungen sind im Allgemeinen
empirisch bekannt: Sie führen zu einer statistischen
Studie, aus der wir die bestimmen können
Wahrscheinlichkeitsverteilungen der Zufallsvariablen,
die das Phänomen charakterisieren.
Sobald die Verteilungen bekannt sind, wird die Stichprobe erstellt
wird durch die Generierung von Zufallszahlen
konstruiert.

GESCHICHTE

Die Methode der Simulation hat eine lange Geschichte.
Es wurde erstmals von Student (1908) eingeführt.
wer entdeckte die Stichprobenverteilungen von
die *t*- **Statistik** und der **Korrelationskoeffizient**
lation.
In jüngerer Zeit dank Computer, Simulationsmethoden
und Monte-Carlo-Methoden
sind rasant vorangekommen.

BEISPIELE

Betrachten Sie das Phänomen des Wartens
die Kasse eines großen Ladens. Vermuten
dass wir das beste „System von“ kennen wollen
Kassierer“, das ist das, wo die Summe der
Kosten der Inaktivität von Kassierern und die Kosten
der wartenden Kunden sind am kleinsten.
Die Entwicklung des Phänomens hängt im
Wesentlichen von der Verteilung der Ankunft ab
der Kunden an der Kasse und im Distri
Abgrenzung der Dienstzeit. Die Simulation
besteht darin, eine **Stichprobe** der Kundenankünfte
und eine Stichprobe der Servicezeiten zu erstellen.
So können wir die Kosten eines Systems berechnen.
Das **Experiment** kann für verschiedene Systeme
wiederholt und das beste System ausgewählt werden.

WEITERLESEN

- **Bootstrap**
- **Generierung von Zufallszahlen**
- **Jackknife-Methode**
- **Monte-Carlo-Methode**
- **Zufallszahl**

VERWEISE

Gosset, SW „Student“: Das Wahrscheinliche
Fehler eines Mittelwerts. Biometrie 6, 1–25
(1908)S. 302–310.
Kleijnen, JPC: Statistische Techniken in
Simulation (in zwei Teilen). Bd. 9 der Statistik

Machine Translated by Google

Tics: Lehrbücher und Monographien. Marcel Dekker, New York (1974)

Kleijnen, JPC: Statistische Tools für Simulationspraktiker. Bd. 76 der Statistik

Tics: Lehrbücher und Monographien. Marcel Dekker, New York (1987)

Naylor, TH, Balintfy, JL, Burdick, DS, Chu, K.: Computersimulationstechniken. Wiley, New York (1967)

Snedecor, George Waddel

Snedecor, George Waddel wurde 1881 geboren in Memphis und starb 1974 in Amherst.
Im Jahr 1899 trat er in das Alabama Polytechnic Institute in Auburn ein und blieb dort mehrere Jahre 2 Jahre. Anschließend verbrachte er 2 Jahre damit, das Lehrerzertifikat zu erwerben. Als seine Familie umzog Tuscaloosa im Jahr 1903, Snedecor, George Wad del wurde an die Alabama University versetzt, wo er 1905 seinen B.S. in Mathematik und Physik erhielt. Seine erste akademische Stelle nahm er an der Militärakademie Selma an, wo er von 1905 bis 1907 lehrte.
Von 1907 bis 1910 unterrichtete er Mathematik und Griechisch am Austin College in Sher Mann.

Im Jahr 1910 zog George Waddel nach Snedecor Ann Arbor, MI, wo er 1913 seine Ausbildung an der Michigan State University abschloss, einen Masterabschluss erworben haben. Im gleichen Jahr wurde er an der University of Iowa angestellt in Ames, wo er bis 1958 blieb.
Im Jahr 1927 erschien die Mathematics Statistical Der Gottesdienst wurde in Ames mit eröffnet Snedecor, GW und Brandt, AE an der Spitze. Dieser Dienst war der Vorläufer des 1933 eingeweihten Statistischen Labors, in dem Snedecor, GW war Regisseur und whereCox, **Gertrude** hat gearbeitet.

Im Jahr 1931 wurde Snedecor, GW, zum Professor an der Fakultät für Mathematik ernannt Iowa State College. Während dieser Zeit, Snedecor, GW entwickelte eine Freundschaft mit **Fisher, Ronald Aylmer.** Im Jahr 1947 Snedecor, GW verließ das Statistiklabor. Er wurde Professor für Statistik am Iowa State College und blieb dort

Diese Position hatte er bis zu seiner Pensionierung im Jahr 1958 inne. Er starb im Februar 1974.

Hauptwerk von Snedecor, George Wad del:

1937 Statistische Methoden für Experimente in Landwirtschaft und Biologie. Collegiate, Ames, USA.

Geodaten

Geodaten sind eine besondere Art von Daten (häufig sogenannte Kartendaten, geografische Daten), die darauf verweisen auf eine räumliche Lokalisierung, wie zum Beispiel die Konzentration von Schadstoffen in einem bestimmten Gebiet, Wolken usw., und alle Informationen das in Form einer Karte dargestellt werden kann. Zu den räumlichen Daten gehören Daten mit eine abzählbare Sammlung räumlicher Standorte, z Beispiel: die Verteilung der Kindersterblichkeit in verschiedenen Städten und Ländern.

Wir können auch davon ausgehen, dass räumliche Daten die sind Realisierung einer regionalen Variablen. Die *Theorie der Regionalvariablen* setzt das voraus Jede Maßnahme kann als modelliert (visualisiert) werden Seindie Realisierung einer Zufallsfunktion (oder Zufallsprozess). Als Erkenntnisse können beispielsweise Bodenproben angesehen werden einer regionalen Variable (siehe **Geostatistik).**

GESCHICHTE

Mit der Entwicklung von Territorialinformationssystemen (TIS) bzw. geografischen Informationssystemen

Machine Translated by Google

tionssysteme (GIS), der Begriff der Geodaten wurde in der Statistik weit verbreitet.

DOMÄNEN UND EINSCHRÄNKUNGEN

Ein GIS ist eine Sammlung von Informationsprogrammen, die auf der Grundlage von Georeferenzen die die Integration räumlicher, nichträumlicher, qualitativer und quantitativer Daten in Datenbanken, die kann als einzigartiges System organisiert werden. Geodaten stammen aus ganz unterschiedlichen Bereichen, und erfordern oft eine Auswahl und Analyse nach spezifischen Kriterien für jede Art von Forschung. Satelliten sammeln enorme Datenmengen, nur ein winziger Bruchteil wird ausgewertet. Diese Fülle an Daten muss überprüft werden durch das Auswahlverfahren. Datenauswahl kann je nach Art der Probleme erfolgen aufzulösen und die zur Herstellung verwendeten Modelle Analyse.

BEISPIELE

Im Allgemeinen können räumliche Daten als behandelt werden Realisierungen von Zufallsvariablen wie in eine gewöhnliche statistische Analyse. Unten sind wir Präsentieren Sie als Beispiel für Geodaten Daten gesammelt im Schweizer Jura. Es ist ein Satz von 259 Daten zur Schwermetallbelastung in einer Region 14,5 km2, 10 davon sind hier vertreten:

X	Y	Ground	Rock	Cd	Co
2.386	3.077	3	3	1,740	9,32
2.544	1.972	2	2	1,335	10,00
2.807	3.347	2	3	1,610	10,60
4.308	1.933	3	2	2.150	11.92
4.383	1.081	3	5	1,565	16,32
3.244	4.519	3	5	1,145	3,50
3,925	3,785	3	5	0,894	15,08
2.116	3.498	3	1	0,525	4,20
1,842	0,989	3	1	0,240	4,52

Cr	Cu	Ni	Pb	Zn
38,32	25,72	21,32	77,36	92,56
40,20	24,76	29,72	77,88	73,56
47.00	8,88	21,40	30,80	64,80
43,52	22,70	29,72	56,40	90,00
38,52	34,32	26,20	66,40	88,40
40,40	31,28	22.04	72,40	75,20
30,52	27,44	21,76	60,00	72,40
25,40	66,12	9,72	141,00	72,08
27,96	22,32	11,32	52,40	56,40

„Boden“, was sich auf die Verwendung des bezieht Boden und „Gestein“, das den Typ beschreibt des zugrunde liegenden Gesteins am Probenort, sind die kategorialen Daten **(qualitative kategoriale Variablen).** Die verschiedenen Arten von Gestein sind:

1. Argovian,
2. Kimmeridgian,
3. Sequanisch,
4. Portlandianer,
5. Quartär.

Die Nutzung des Bodens entspricht:

1. Wald,
2. Weideland,
3. Prärie,
4. gepflügt.

WEITERLESEN

Einstufung
Geostatistik
Histogramm
Probenahme
Raumstatistik

VERWEISE

Atteia, O., Dubois, J.-P., Webster, R.: Geo Statistische Analyse der Bodenverunreinigung der Schweizer Jura. Umgebung. Umweltverschmutzung. **86,** 315–327 (1994)

Goovaerts, P.: Geostatistik für die Natur Ressourcenbewertung. Universität Oxford sity Press, Oxford (1997)

Machine Translated by Google

Matheron, G.: Die Theorie der Variablen
regionalisiert und ihre Anwendungen. Mas Sohn, Paris
(1965)

Raumstatistik

Die Raumstatistik befasst sich mit der statistischen Analyse **räumlicher Daten.** Die Raumstatistik umfasst alle Techniken zur Erforschung und beweisen Sie das Vorhandensein einer räumlichen Abhängigkeit zwischen im Raum verteilten Beobachtungen. Geodaten können aus der Geologie und der Erde stammen Naturwissenschaften, Bildbearbeitung, **Epidemiologie,** Landwirtschaft, Ökologie, Astronomie oder Wald Wissenschaften. Es wird davon ausgegangen, dass die Daten zufällig sind. und manchmal wird auch ihr Platz angenommen zufällig sein.

GESCHICHTE

Die statistische Behandlung räumlicher Datendaten zurück zu Halley, Edmond (1686). Halley, E. Überlagert auf einer Karte das Relief des Windes Richtungen und Monsun zwischen und um die Tropen und versuchte die Ursachen zu erklären. Das räumliche Statistikmodell erschien viel später. **Fisher, Ronald Aylmer** (1920–1930), während seiner Forschungen zum Experimentellen Station von Rothamsted in England, formulierte die Grundlage der Prinzipien des Zufalls Wahl, der Analyse nach Blöcken und der Replikation.

MATHEMATISCHE ASPEKTE

Die statistische Analyse räumlicher Datenformen das Hauptfach der räumlichen Statistik. Der Hauptwerkzeuge der statistischen Analyse sind das Histogramm, die Kurve der Häufigkeitskumulation, das **QQ-Diagramm,** der **Median,** der **Mittelwert** der Verteilung, die Variationskoeffizienten und das **Maß der Schiefe.** Im Bei vielen Variablen bestehen die am häufigsten verwendeten statistischen Ansätze darin, sie in der Abbildung darzustellen Form eines Streudiagramms zu erstellen und deren Korrelation zu untersuchen. Die räumliche Analyse hat die Unterscheidungskraft Merkmal, das räumliche Informationen und die zwischen ihnen bestehende Beziehung berücksichtigt zwei räumliche Daten. Zwei nahe beieinander liegende Daten werden einander mit größerer Wahrscheinlichkeit ähneln wenn sie weiter voneinander entfernt wären. Wenn wir also eine Karte analysieren, die die Schadstoffkonzentrationen am Boden zeigt, sehen wir keine zufälligen Werte, sondern im Gegenteil kleine Werte von Schadstoffen gruppiert sowie große Werte, und der Übergang zwischen den Klassen würde kontinuierlich sein.

DOMÄNEN UND EINSCHRÄNKUNGEN

Räumliche Darstellung in Form von Karten und Zeichnungen und räumliche Statistiken können sein vereint, um ein besseres visuelles Verständnis zu ermöglichen des Einflusses, den die benachbarten Werte haben aufeinander haben. Überlagerung statistisch Informationen auf einer Karte ermöglichen eine bessere Analyse die Informationen zu erfassen und das entsprechende zu zeichnen Schlussfolgerungen.

BEISPIELE

Siehe **räumliche Daten.**

WEITERLESEN

- **Autokorrelation**
- **Geostatistik**
- **Geodaten**

VERWEISE

Cressie, NAC: Statistik für räumliche Daten.
Wiley, New York (1991)

Goovaerts, P.: Geostatistik für die Natur
Ressourcenbewertung. Universität Oxford
sity Press, Oxford (1997)

Isaaks, EH, Srivastava, RM: Eine Einführung in die
angewandte Geostatistik. Oxford
University Press, Oxford (1989)

Machine Translated by Google

Journel, A., Huijbregts, CJ: Bergbaugeostatistik. Akademisch, New York (1981)

Matheron, G.: Die Theorie regionaler Variablen und ihre Anwendungen. Mas Sohn, Paris (1965)

Oliver, MA, Webster, R.: Kriging: eine Methode zur Interpolation für geografische Informationssysteme. Int. J. Geogr. Inf. Syst. 4(3), 313–332 (1990)

Speerführer-Rang Korrelationskoeffizient

Der Spearman-Rangkorrelationskoeffizient (Spearman ÿ) ist eine nichtparametrische Messkorrelation. Es wird zur Bestimmung verwendet **Beziehung** zwischen zwei Datensätzen .

GESCHICHTE

Spearman, Charles war Psychologe. 1904 führte er erstmals den Dienstgrad ein Korrelationskoeffizient. Es wird oft als ÿ von Spearman bezeichnet und ist eine der ältesten Rangstatistiken .

MATHEMATISCHE ASPEKTE

Seien (X1, *X2,...,Xn)* und (Y1, *Y2,...,Yn)* zwei **Stichproben** der Größe n. *RXi* bezeichnet den Rang von *Xi* im Vergleich zu den anderen **Werten** der *X*- Stichprobe, für *i = 1, 2,...,n. RXi* = 1, wenn *Xi* der kleinste Wert von X ist, *RXi* = 2, wenn *Xi* der zweitkleinste Wert ist usw., bis *RXi = n ,* wenn *Xi* der größte Wert von X ist. Ebenso bezeichnet *RYi* den Rang von *Yi ,* fori = *1, 2,...,n.*
Der Spearman-Rangkorrelationskoeffizient, im Allgemeinen mit ÿ bezeichnet, ist definiert durch:

$$\ddot{y} = 1 \ddot{y} \frac{6 \sum_{i=1}^{N} d2}{n(n2 \ddot{y} 1)},$$

wobei *di = RXi* ÿ *RYi .*

Wenn mehrere **Beobachtungen** genau den gleichen **Wert haben,** wird diesen Beobachtungen ein Durchschnittsrang zugewiesen. Wenn es viele Durchschnittsränge gibt, nehmen Sie am besten eine Korrektur vor und berechnen Sie:

$$\ddot{y} = \frac{Sx + Sy \ddot{y} \sum_{i=1}^{n} d2}{2\sqrt{Sx \cdot Sy}},$$

Wo

$$Sx = \frac{n\,n2 \ddot{y} 1 \ddot{y} 3 \sum_{i=1}^{G} t \ddot{y} ti\,i}{12},$$

wobei *g* die Anzahl der Gruppen mit durchschnittlichen Altersrängen und *ti* die Größe der Gruppe *i* für die ist **Xsample** und

$$Sy = \frac{n\,n2 \ddot{y} 1 \ddot{y} 3 \sum_{j=1}^{H} t_j \ddot{y} tj}{12},$$

wobei *h* die Anzahl der Gruppen mit durchschnittlichen Rängen und *tj* die Größe der Gruppe *j* für die *Y*-Stichprobe ist.
(Wenn es keine Durchschnittsränge gibt, werden die **Beobachtungen** als Gruppen der Größe 1 angesehen, was bedeutet, dass *g = h = n* und *ti = tj* = 1 für *i, j = 1, 2,...,n* und *Sx = Sy = n(n2* ÿ 1)/12.)

Hypothesentest

Der Spearman-Rangkorrelationskoeffizient wird oft als statistischer Test verwendet, um festzustellen, ob Es besteht eine Beziehung zwischen zwei **Zufällen Variablen.** Der Test kann ein **bilateraler Test** sein oder ein **einseitiger Test.** Die Hypothesen sind:

A: *Bilateraler Fall*

H0: *X* und *Y* sind voneinander unabhängig.

H1: Es besteht entweder eine positive oder eine negative Korrelation zwischen *X* und Y.

Machine Translated by Google

Es besteht eine positive Korrelation, wenn die große **Werte** von *X* neigen dazu, mit großen Werten von *Y* und kleinen Werten von verbunden zu sein *X* mit kleinen Werten von Y. Es gibt ein Negativ Korrelation, wenn große Werte von X häufig mit kleinen Werten von verknüpft werden müssen *Y* und umgekehrt.

B: *Einseitiger Fall*

H0: *X* und *Y* sind voneinander unabhängig.

H1: Es besteht eine positive Korrelation zwischen *X* und Y.

C: *Einseitiger Fall*

H0: *X* und *Y* sind voneinander unabhängig.

H1: Es besteht eine negative Korrelation zwischen *X* und Y.

Entscheidungsregeln

Die Entscheidungsregeln sind je nach unterschiedlich die **Hypothesen.** Deshalb gibt es die Entscheidungsregeln A, B und C relativ zu den vorherigen Fälle.

Entscheidungsregel A

Lehnen Sie H0 auf dem **Signifikanzniveau** ÿ ab, wenn

$$ÿ > tn,1ÿÿ \quad _2 \quad \text{oder } ÿ < tn,ÿ \quad _2 \,,$$

wobei *t* der **kritische Wert** des gegebenen Tests ist am Spearman-Tisch.

Entscheidungsregel B

Lehnen Sie H0 auf dem **Signifikanzniveau** ÿ ab, wenn

$$ÿ > tn,1ÿÿ .$$

Entscheidungsregel C

Lehnen Sie H0 auf dem **Signifikanzniveau** ÿ ab, wenn

$$ÿ < tn,ÿ .$$

Anmerkung: Die Notation *t* bedeutet das nicht

Die Spearman-Koeffizienten hängen damit zusammen des Studenten.

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Rangkorrelationskoeffizient nach Spearman beträgt Wird als **Hypothesentest** verwendet , um die Abhängigkeit zwischen zwei **Zufallsvariablen zu untersuchen.** Es kann als ein **Test der Unabhängigkeit angesehen werden.**

Als nichtparametrisches Korrelationsmaß kann es auch mit Nominalwerten verwendet werden oder ordinale **Daten.**

Eine Korrelationsmessung zwischen zwei **Variablen** müssen die folgenden Punkte erfüllen:

1. Es nimmt **Werte** zwischen ÿ1 und +1 an.
2. Es besteht eine positive Korrelation zwischen *X* und *Y* , wenn der Wert des Korrelationskoeffizienten ist effizient ist positiv; Eine perfekte positive Korrelation entspricht einem Wert von +1.
3. Es besteht eine negative Korrelation zwischen *X* und *Y* , wenn der Wert des Korrelationskoeffizienten ist effizient ist negativ; Eine perfekte negative Korrelation entspricht einem Wert von ÿ1.
4. Es besteht keine Korrelation zwischen *X* und *Y* wenn der Korrelationskoeffizient nahe bei liegt null; man kann auch sagen, dass X und *Y* es nicht sind korreliert.

Der Rangkorrelationskoeffizient nach Spearman bietet folgende Vorteile:

- Bei den **Daten** kann es sich um nichtnumerische **Beobachtungen** handeln **solange** sie klassifiziert werden können nach bestimmten Kriterien.
- Es ist leicht zu berechnen.
- Der zugehörige Statistiktest funktioniert nicht eine **Grundhypothese** basierend auf dem formulieren Form der Verteilung der **Bevölkerung,** aus der die **Proben** entnommen werden de.

Die Spearman-Tabelle gibt die theoretischen **Werte** des Spearman-Rangkorrelationskoeffizienten unter der **Hypothese** der **Unabhängigkeit** zweier **Zufallsvariablen an.**

Hier ist ein Beispiel der Spearman-Tabelle für *n* = 6, 7 und 8 und ÿ = 0,05 und 0,025:

Machine Translated by Google

N	ÿ = 005.	ÿ = 0025
6	0,7714	0,8286
7	0,6786	0,7450
8	0,6190	0,7143

Eine vollständige Spearman-Tabelle finden Sie in Glasser, GJ und Winter, RF (1961).

BEISPIELE

In diesem Beispiel nehmen acht Paare echte Zwillinge auf Intelligenztests. Das Ziel besteht darin, herauszufinden, ob dies der Fall ist **Unabhängigkeit** zwischen dem Test des erstgeborenen Zwillings und dem des zweitgeborenen Zwillings.

Die Daten sind in der folgenden Tabelle aufgeführt Die höchsten Punktzahlen entsprechen den Besten Ergebnisse.

Zwillingspaar am 1. geboren	Xi	Geboren am 2	Yi
1	90		88
2	75		79
3	99		98
4	60		66
5	72		64
6	83		83
7	86		86
8	92		95

DieX werden untereinander klassifiziert und das *Y* untereinander und *di* wird berechnet. Das gibt:

Ein Paar Zwillinge	Geboren 1 Xi	RXi	Geboren 2 Yi	RYi	di
1	90	6	88	6	0
2	75	3	79	3	0
3	99	8	98	8	0
4	60	1	66	2	ÿ1
5	72	2	64	1	1
6	83	4	83	4	0
7	86	5	86	5	0
8	92	7	95	7	0

Der Rangkorrelationskoeffizient nach Spearman beträgt dann berechnet:

$$\ddot{y} = 1 \ddot{y} \frac{6 \cdot \sum_{i=1}^{8} d2_{_}}{n(n2 \ddot{y} 1)} = 1 \ddot{y} \frac{6 \cdot 2}{8(82 \ddot{y} 1)}$$

$$= 0{,}9762 .$$

Dies zeigt, dass es eine (nahezu perfekte) positive Korrelation zwischen der Intelligenz Tests.

Angenommen, die Ergebnisse für 7 und 8 lauten geändert. Wir erhalten dann die folgende Tabelle:

Ein Paar Zwillinge	Geboren 1 Xi	RXi	Geboren 2 Yi	RYi	di
1	90	6,5	88	5.5	1
2	75	3	79	3	0
3	99	8	98	7,5	ÿ0,5
4	60	1	66	2	ÿ1
5	72	2	64	1	1
6	83	4,5	83	4	0,5
7	83	4,5	88	5.5	ÿ1
8	90	6,5	98	7.5	ÿ1

Da es durchschnittliche Ränge gibt, verwenden wir die Formel

$$\ddot{y} = \frac{Sx + Sy \ddot{y} \sum_{i=1}^{8} d2_{_}}{2\sqrt{Sx \cdot Sy}}$$

um die Spearman-Rangkorrelation zu berechnen Koeffizient. Wir berechnen zunächst *Sx:*

$$Sx = \frac{n \, n2 \ddot{y} 1 \ddot{y} \sum_{i=1}^{G} T^3_{_} \ddot{y} ti}{12}$$

$$= \frac{8\,82 \ddot{y} 1 \ddot{y} 23 \ddot{y} 2 + 23 \ddot{y} 2}{12}$$

$$= 41 .$$

Machine Translated by Google

Dann *Sy:*

$$Sy = \frac{n\,n2 \ddot{y}\ 1 \ddot{y} \quad \sum_{j=1}^{H} 3\,t_j \ddot{y}\ tj}{12} = \frac{8\ 82 \ddot{y}\ 1 \ddot{y}\ 23 \ddot{y}\ 2 + 23 \ddot{y}\ 2}{12} = 41 \ .$$

Der Wert von ÿ wird zu Folgendem:

$$\ddot{y}41 \cdot 41 \frac{41 + 41 \ddot{y}\ (1 + 0 + \ldots + 1) \ddot{y} = 2}{} = \frac{76{,}5}{82} = 0{,}9329 \ .$$

Wir sehen, dass es in diesem Fall auch ein Fast gibt perfekte positive Korrelation.
Wir führen nun den **Hypothesentest durch:**

H0: Es besteht Unabhängigkeit zwischen den Intelligenztests eines Zwillingspaares.

H1: Es besteht ein positiver Zusammenhang zwischen den Intelligenztests.

Wir wählen ein **signifikantes Niveau** von ÿ = 0,05.
Da wir uns im Fall B befinden, wird H0 abgelehnt, wenn

$$\ddot{y} > t8{,}0{,}95 \ ,$$

wobei t8,0,95 der **Wert** der Spearman-Tabelle ist , also wenn

$$\ddot{y} > 0{,}6190 \ .$$

In beiden Fällen (ÿ = 0,9762 und ÿ = 0,9329) wird H0 verworfen. Wir können daraus schließen, dass zwischen den Ergebnissen der Intelligenztests eines Zwillingspaares eine positive Korrelation besteht.

WEITERLESEN

- **Hypothesentest**
- **Nichtparametrischer Test**
- **Test der Unabhängigkeit**

VERWEISE

Glasser, GJ, Winter, RF: Kritische Werte des Rangkorrelationskoeffizienten für Prüfung der Unabhängigkeitshypothese. Biometrie **48,** 444–448 (1961)

Spearman, C.: Der Beweis und die Messung der Assoziation zwischen zwei Dingen. Bin. J. Psychol. **15,** 72–101 (1904)

Spearman-Tisch

Die Spearmant-Tabelle gibt die theoretischen Werte des Spearman-Rangkorrelationskoeffizienten unter **der Hypothese der Unabhängigkeit** zweier Zufallsvariablen an.

GESCHICHTE

Siehe **Rangkorrelationskoeffizient nach Spearman.**

WEITERLESEN

- **Rangkorrelationskoeffizient nach Spearman**

Standardabweichung

Die Standardabweichung ist ein **Maß für dis Person.** Sie entspricht der positiven Quadratwurzel der **Varianz,** wobei die Varianz der Mittelwert der quadratischen **Abweichungen** jeder **Beobachtung** in Bezug auf den **Mittelwert** der Beobachtungsmenge ist.

Es wird normalerweise mit ÿ bezeichnet, wenn es relativ zu einer **Grundgesamtheit** ist , und mit *S* , wenn es relativ zu einer **Stichprobe ist.**
In der Praxis wird die Standardabweichung ÿ **einer Population** anhand der Standardabweichung *S* einer **Stichprobe** dieser Population geschätzt.

S

Machine Translated by Google

GESCHICHTE

Der Begriff Standardabweichung steht in engem Zusammenhang mit den Arbeiten der beiden englischen Mathematiker **Pearson, Karl** und **Gosset, WS .** Tatsächlich verwendete Pearson, K. den Begriff während einer Konferenz, die er 1893 vor der London Royal Society hielt, zum ersten Mal.

Er verwendete es erneut in seinem Artikel „On the Dissection of Asymmetrical Frequency Curves“ im Jahr 1894, und es war auch Pearson, K., der das Symbol ÿ zur Bezeichnung der Standardabweichung einführte. Gosset, WS (Student), arbeitete ebenfalls an diesen Problemen. Er erklärte, warum es wichtig ist, *S* (Standardabweichung relativ zu einer **Stichprobe)** von ÿ (Standardabweichung relativ zu einer **Population) zu unterscheiden.**

In seinem Artikel vom März 1908, **Gosset,** definierte WS die Standardabweichung einer **Stichprobe** durch:

$$S = \sqrt{\frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}{N}}.$$

Es stellte sich die Frage, ob dieser Ausdruck durch *n* oder durch *n* ÿ 1 geteilt werden sollte. Pearson, K. fragte Gosset, WS nach seiner Meinung, der in einem Brief vom 13. März 1927 antwortete, dass beide Formeln möglich seien
in der Literatur gefunden. Der mit einem nÿ1
Der Nenner ergibt einen Mittelwert , der unabhängig von der Größe der **Stichprobe** und damit der **Grundgesamtheit ist.** Bei großen Stichproben hingegen ist der Unterschied zwischen beiden Formeln vernachlässigbar und die Berechnungen werden einfacher, wenn die Formel mit n *Nenner* verwendet wird.

Insgesamt deutet Gosset, WS darauf hin, dass die Verwendung von *n* ÿ 1 wahrscheinlich angemessener ist für kleine **Stichproben,** wohingegen die Verwendung von *n* für große Stichproben vorzuziehen ist.

Beachten Sie, dass die Entdeckung der Standardabweichung in den Kontext der Theorie der **Schätzung** und der **Hypothesenprüfung gestellt werden muss.** Auch die Untersuchung der Variabilität erregte zunächst die Aufmerksamkeit der Astronomen, weil sie an Entdeckungen im Zusammenhang mit der Variabilität interessiert waren Verteilung von **Fehlern.**

MATHEMATISCHE ASPEKTE

Betrachten Sie eine **Zufallsvariable** *X* mit einem **Erwartungswert** *E[X].* Die **Varianz** ÿ2 der **Grundgesamtheit** ist definiert durch:

$$\ddot{y}2 = E\,(X \ddot{y} E\,[X])^2$$

oder

$$\ddot{y}2 = E\,X2 \ddot{y} E\,[X]^2 .$$

Die Standardabweichung ÿ ist die positive Quadratwurzel von ÿ2:

$$\ddot{y} = \sqrt{E\,(X \ddot{y} E\,[X])^2}.$$

Um die Standardabweichung ÿ von abzuschätzen

Bei einer **Grundgesamtheit** sollte die Verteilung der Grundgesamtheit bekannt sein, in der Praxis ist es jedoch zweckmäßig, ÿ anhand eines Zufalls zu schätzen **Stichprobe** *x1,...,xn* mit der Formel:

$$S = \sqrt{\frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}{n \ddot{y} 1}}$$

wobei *S* der **Schätzer** von ÿ und x¯ der ist **Arithmetisches Mittel** der **Stichprobe:**

$$\bar{x} = \frac{\sum_{i=1}^{N} xi}{N}.$$

Machine Translated by Google

Die Standardabweichung kann ebenfalls berechnet werden unter Verwendung der folgenden alternativen Formel:

$$S = \sqrt{\frac{\sum_{i=1}^{N} x_2 \ddot{y} n\bar{x}^2}{n \ddot{y} 1}}$$

$$= \sqrt{\frac{N \sum_{i=1}^{N} x_2 \ddot{y} \left(\sum_{i=1}^{N} x_i\right)^2}{n(n \ddot{y} 1)}}.$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Als Maßstab wird die Standardabweichung verwendet Messung in **Hypothesentests** und **Vertrauensintervalle.** Andere **Maßnahmen von Streuung** wie der **Bereich** oder der **Mittelwert Abweichungen werden** auch **in beschreibenden Statistiken verwendet,** sie spielen jedoch bei Tests eine untergeordnete Rolle von Hypothesen. Es kann Folgendes überprüft werden:

$$\text{mittlere Abweichung} \ddot{y} S \ddot{y} 2 \frac{\text{Bereich}}{\sqrt{\frac{N}{n \ddot{y} 1}}}.$$

Wenn wir eine **Zufallsvariable** betrachten , folgt: ergibt eine **Normalverteilung** des **Mittelwerts** ÿ und Standardabweichung ÿ, das können wir überprüfen 68,26 % der **Beobachtungen** befinden sich in das **Intervall** ÿ ± ÿ. Außerdem 95,44 % der Beobachtungen liegen im Intervall ÿ ± 2ÿ, und 99,74 % der Beobachtungen liegen im Intervall ÿ ± 3ÿ.

BEISPIELE

Fünf Studenten haben nacheinander zwei bestanden Prüfungen, bei denen sie Folgendes erreichten Noten:
Prüfung 1:

$$\frac{3{,}5 + 4{,}5 + \bar{x} = = 4}{\frac{20}{5}}$$

Prüfung 2:

$$\frac{2{,}5 + 5{,}5 + 3{,}5 + 4{,}5 + 4 \bar{x} = = 4}{\frac{20}{5}}$$

Das **arithmetische Mittel** x¯ dieser beiden Mengen der **Beobachtungen** ist identisch. Nichtsdestotrotz, die Streuung der Beobachtungen rundherum Mittelwert ist nicht dasselbe.

Um die Standardabweichung zu berechnen, berechnen wir zunächst müssen jeweils die **Abweichungen** berechnen **Beobachtung** in Bezug auf die **Arithmetik Mittelwert** und quadrieren Sie dann diese Abweichungen:

Prüfung 1:

Grad	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3.5	ÿ0,5	0,25
4	0,0	0,00
4.5	0,5	0,25
3.5	ÿ0,5	0,25
4.5	0,5	0,25
$\sum_{i=1}^{5}(x_i \ddot{y} \bar{x})^2$		1,00

Prüfung 2:

Grad	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2.5	ÿ1,5	2,25
5.5	1.5	2,25
3.5	ÿ0,5	0,25
4.5	0,5	0,25
4	0,0	0,00
$\sum_{i=1}^{5}(x_i \ddot{y} \bar{x})^2$		5.00

Die Standardabweichung für jede Prüfung beträgt gleich:
Prüfung 1:

$$S = \sqrt{\frac{\sum_{i=1}^{5} (x_i \ddot{y} \bar{x})^2}{n \ddot{y} 1}} = \sqrt{\frac{1}{4}} = 0{,}5$$

Prüfung 2:

$$S = \sqrt{\frac{\sum_{i=1}^{5} (x_i \ddot{y} \bar{x})^2}{n \ddot{y} 1}} = \sqrt{\frac{5}{4}} = 1{,}118$$

Da die Standardabweichung der Noten in Die zweite Prüfung ist größer, die Noten sind höher

S

Machine Translated by Google

um das **arithmetische Mittel** verteilt als
für die erste Prüfung. Die Variabilität der zweiten
Prüfung ist daher größer als die der
Erste.

WEITERLESEN

Variationskoeffizient
Mittlere absolute Abweichung
Maß der Streuung
Varianz

VERWEISE

Barnard, GA, Plackett, RL, Pearson, S.:
„Student“: eine statistische Biographie von
William Sealy Gosset: basierend auf Schriften
von ES Pearson. Universität Oxford
Press, Oxford (1990)

Pearson, K.: Beiträge zur mathematischen
Evolutionstheorie. I. In: Karl
Pearsons frühe statistische Arbeiten. Cambridge
University Press, Cambridge,
S. 1–40 (1948). Erstveröffentlichung 1894
as: Zur Zerlegung asymmetrischer
Frequenzkurven. Philos. Trans. Roy. Soc.
London. Ser. A **185,** 71–110

Gosset, SW „Student“: Der wahrscheinliche Fehler
eines Mittelwerts. Biometrie 6, 1–25 (1908)

Standart Fehler

Der Standardfehler **ist** die Quadratwurzel von
geschätzte **Varianz** einer **Statistik,** Bedeutung
die **Standardabweichung** der **Stichprobe**
Verteilung dieser Statistik.

GESCHICHTE

Der Standardfehlerbegriff wird zugeschrieben
Gauss, CF (1816), der dies offenbar nicht tat
kennen das Konzept **der Stichprobenverteilung**
einer **Statistik ,** die zur Schätzung des **Werts** von verwendet wird
ein **Parameter** einer **Population.**

MATHEMATISCHE ASPEKTE

Der **Standardfehler des Mittelwerts oder Standards**
Abweichung der **Stichprobenverteilung** von
Der mit ÿx¯ bezeichnete Mittelwert wird wie folgt berechnet
eine Funktion der Standardabweichung der
Bevölkerung und der jeweiligen Größe der
endliche Grundgesamtheit der **Stichprobe:**

$$ÿx^{-} = \frac{ÿ}{\sqrt{ÿn}} \cdot \sqrt{\frac{N ÿ n}{N}},$$

wobei ÿ die Standardabweichung der Bevölkerung
ist, *N* die Größe der Bevölkerung ist und
n ist die Stichprobengröße.
Wenn die Größe *N* der Population unendlich ist, ist die
Korrekturfaktor:

$$\sqrt{\frac{N ÿ n}{N}}$$

kann weggelassen werden, da es gegen 1 tendiert
wenn *N* gegen Unendlich tendiert.
Wenn die **Grundgesamtheit** unendlich ist oder
wenn die **Stichprobe** nicht erschöpfend ist (mit Ersetzung),
dann ist der **Standardfehler** des **Mittelwerts** gleich
zur Standardabweichung der **Grundgesamtheit**
dividiert durch die Quadratwurzel der Größe des
Probe:

$$ÿx^{-} = \frac{ÿ}{\sqrt{ÿn}}.$$

Wenn die **Standardabweichung** ÿ der
Grundgesamtheit unbekannt ist, kann sie durch geschätzt werde
Standardabweichung *S* der **Stichprobe** gegeben durch
die Quadratwurzel der **Varianz** S2:

$$S2 = \frac{\sum_{j=1}^{N} xj ÿ \bar{x}^{2}}{n ÿ 1}.$$

Für den Standard gelten die gleichen Formeln
Fehler eines **Anteils** oder **Standardabweichung**
der Stichprobenverteilung **der Anteile ,** bezeichnet
mit *ÿp:*

$$ÿp = \frac{ÿ}{\sqrt{ÿn}} \cdot \sqrt{\frac{N ÿ n}{N}}$$

Machine Translated by Google

im Fall einer endlichen **Population** und

$$\ddot{y}p = \frac{\ddot{y}}{\sqrt{\ddot{y}n}}$$

im Fall einer unendlichen Population, wobei ÿ die **Standardabweichung** der Population ist:

$$\ddot{y} = p \cdot (1 \ddot{y} p) = \ddot{y}p \cdot q ,$$

Dabei ist *p* die **Wahrscheinlichkeit ,** dass ein Element das untersuchte Merkmal besitzt, und *q* = 1 ÿ *p* die Wahrscheinlichkeit, dass dies nicht der Fall ist.

Der **Standardfehler** der Differenz zwischen zwei unabhängigen Größen ist die Quadratwurzel der Summe der quadrierten Standardfehler beider Größen. Zum Beispiel der Standardfehler der Differenz zwischen zwei **Mittelwerten** ist gleich:

$$\ddot{y}x^{-}1\ddot{y}^{-}x2 = \ddot{y}2_{x^{-}1} + \ddot{y}2_{x^{-}2} = \frac{\ddot{y}2_1}{n1}\ddot{y}2 + \frac{2}{n2} ,$$

Dabei sind ÿ2 und 1ÿ2 die jeweiligen Varianzen der beiden zu vergleichenden unendlichen **Populationen** und n1 und n2 die jeweiligen Größen der beiden **Stichproben.**

Wenn die **Varianzen** der beiden **Grundgesamtheiten** unbekannt sind, können wir sie anhand der für die beiden **Stichproben berechneten Varianzen schätzen,** was Folgendes ergibt:

$$\ddot{y}x^{-}1\ddot{y}^{-}x2 = \sqrt{\frac{S2_1}{n1} + \frac{S2_2}{n2}} ,$$

wobei S2 1und S2 2die jeweiligen **Varianzen** sind der beiden **Proben.**

Wenn wir bedenken, dass die **Varianzen** der beiden

Wenn **die Populationen** gleich sind, können wir den Wert dieser Varianzen mit einer **gepoolten Varianz** S2 schätzen *p*, die als Funktion von S2 und S2 2 berechnet wird:

$$S2_p = \frac{(n1 \ddot{y} 1) \cdot S2 + (n2 \ddot{y} 1) \cdot S2_2}{n1 + n2 \ddot{y} 2} ,$$

was ergibt:

$$\ddot{y}x^{-}1\ddot{y}^{-}x2 = Sp \cdot \sqrt{\frac{1}{n1} + \frac{1}{n2}} .$$

Der **Standardfehler** der Differenz zwischen zwei **Proportionen** wird auf die gleiche Weise berechnet wie der Standardfehler der Differenz zwischen zwei **Mittelwerten.** Deshalb:

$$\ddot{y}p1\ddot{y}p2 = \ddot{y}2_{p1} + \ddot{y}2_{p2}$$

$$= \frac{p1 \cdot (1 \ddot{y} p1)}{n1} + \frac{p2 \cdot (1 \ddot{y} p2)}{n2} ,$$

Dabei sind p1 und p2 die **Proportionen für** unendliche **Populationen ,** die anhand der beiden **Stichproben** der Größe n1 und n2 berechnet wurden.

BEISPIELE

(1) *Mittlerer Standardfehler*

Ein Hersteller möchte die Präzision einer neuen Maschine zur Herstellung von Schrauben mit einem Durchmesser von 8 mm testen

eter. An einer Charge von N = 10000 Stück wird eine **Stichprobe** von *n* = 100 Stück untersucht; die **Standardabweichung** *S* beträgt 1,2 mm.

Der Standardfehler des **Mittelwerts** ist gleich:

$$\ddot{y}x^{-} = \frac{\ddot{y}}{\sqrt{\ddot{y}n}} \cdot \sqrt{\frac{N \ddot{y} n}{N}} .$$

Da die **Standardabweichung** ÿ der **Grundgesamtheit** unbekannt ist, können wir sie mithilfe der Standardabweichung **der Stichprobe S schätzen, was Folgendes** ergibt:

$$\ddot{y}x^{-} = \frac{S}{\sqrt{\ddot{y}n}} \cdot \sqrt{\frac{N \ddot{y} n}{N}}$$

$$= \frac{1.2}{10} \cdot \sqrt{\frac{10000 \ddot{y} 100}{10000}}$$

$$= 0{,}12 \cdot 0{,}995 = 0{,}119 .$$

Beachten Sie, dass der Faktor

$$\sqrt{\frac{N \ddot{y} n}{N}} = 0{,}995$$

S

Machine Translated by Google

hat ebenso wenig Einfluss auf das Ergebnis wie die Größe der **Bevölkerung** ist groß genug.

(2) *Standardfehler eines Anteils*

Ein Kandidat führt eine Umfrage **zu einer Stichprobe durch** 200 Leute wollen wissen, ob er mehr als haben wird 50 % (= ÿ0) der Stimmen, wobei ÿ0 der angenommene Wert des **Parameters** ÿ **(Anteil)** ist der **Bevölkerung).**

Der Standardfehler des Anteils *ÿp* beträgt gleich:

$$ÿp = \sqrt{\frac{ÿ0 \cdot (1 \text{ ÿ } ÿ0)}{N}},$$

Dabei ist *n* die **Stichprobengröße** (in unserem Beispiel 200). Wir können davon ausgehen, dass die **Bevölkerung** ist unendlich, weshalb wir nicht darauf eingehen Berücksichtigen Sie den Korrekturfaktor:

$$ÿp = \sqrt{\frac{0{,}5 \cdot (1 \text{ ÿ } 0{,}5)}{200}} = 0{,}0354 .$$

(3) *Standardfehler der Differenz zwischen zwei Mengen*

Eine Versicherungsgesellschaft entscheidet sich für die Ausstattung Büros mit Mikrocomputern. Es will kaufen Diese Mikrocomputer bestehen aus zwei verschiedenen Lieferanten, sofern zwischen den beiden Marken kein wesentlicher Unterschied in der Haltbarkeit besteht. Es testet eine Stichprobe von 35 Mikrocomputern Marke 1 und 32 von Marke 2, dabei die Uhrzeit notieren das verging vor der ersten Panne. Der Die beobachteten Daten sind die folgenden:

Marke 1: Standardabweichung S1 = 57,916

Marke 2: Standardabweichung S2 = 57,247

Fall 1:

Wir nehmen an, dass die **Varianzen** der beiden **Populationen** sind gleich. Der gepoolte Standard Abweichung ist gleich:

$$Sp = \sqrt{\frac{(n1 \text{ ÿ } 1) \cdot S2_{\,1} + (n2 \text{ ÿ } 1) \cdot S2_{\,2}}{n1 + n2 \text{ ÿ } 2}}$$

$$= \sqrt{\frac{34 \cdot 57{,}9162 + 31 \cdot 57{,}2472}{35 + 32 \text{ ÿ } 2}}$$

$$= 57{,}5979 .$$

Wenn wir die gewichtete **Standardabweichung kennen,** können wir den **Wert** der Standardabweichung der **Stichprobenverteilung** berechnen mit der folgenden Formel:

$$ÿx^{-}1ÿ^{-}x2 = Sp \cdot \sqrt{\frac{1}{n1} + \frac{1}{n2}}$$

$$= 57{,}5979 \cdot \sqrt{\frac{1}{35} + \frac{1}{32}} = 14{,}0875 .$$

Fall 2:

Wenn wir davon ausgehen, dass die **Varianzen** der beiden **Populationen** ungleich sind, müssen wir zunächst den Standardfehler des **Mittelwerts** berechnen jede **Probe:**

$$ÿx^{-}1 = \frac{S1}{\sqrt{ÿn1}} = \frac{57.916}{\sqrt{ÿ35}} = 9{,}79 ,$$

$$ÿx^{-}2 = \frac{S2}{\sqrt{ÿn2}} = \frac{57{,}247}{\sqrt{ÿ32}} = 10{,}12 ,$$

$$ÿx^{-}1ÿ^{-}x2 = \sqrt{ÿ2_{\,x^{-}1} + ÿ2_{\,x^{-}2}}$$

$$= 9{,}792 + 10{,}122$$

$$= 14{,}08 .$$

Wir stellen in diesem Beispiel fest, dass die beiden Fälle Werte des Standardfehlers ergeben fast gleich. Dies wird durch die erklärt Sehr geringfügiger Unterschied zum **Standard Abweichungen** der beiden **Proben** S1 und S2.

WEITERLESEN

Hypothesentest

Stichprobenverteilung

Standardabweichung

Machine Translated by Google

VERWEISE

Box, GEP, Hunter, WG, Hunter, JS:
Statistiken für Experimentatoren. Eine Einführung
in Design, Datenanalyse und
Modellbau. Wiley, New York (1978)

Cochran, WG: Sampling Techniques, 2
Hrsg. Wiley, New York (1963)

Gauss, CF: Bestimmung der Genauigkeit
der Beobachtungen, vol. 4, S. 109–117
(1816). In: Gauss, CF Werke (veröffentlicht
im Jahr 1880). Dieterichsche Universitätsdruckerei,
Göttingen

Gauss, CF: Gauss' Werk (1803–1826) über
die Theorie der kleinsten Quadrate. Trotter, HF,
trans. Statistische Technikforschung
Group, Technical Report, Nr. 5, Princeton University
Press, Princeton, NJ (1957)

Standardisierte Daten

Standardisierte Daten sind Daten, aus denen wir
Subtrahieren Sie den Mittelwert der entsprechenden Beobachtungen
edzur **Zufallsvariablen** (zugeordnet zur
Daten) und dann durch die Standardabweichung der
Beobachtungen dividieren. Somit standardisiert
Die Daten haben einen Mittelwert von 0 und eine Varianz von 1.

MATHEMATISCHE ASPEKTE

Die Standardisierung einer Datenreihe X mit

Die Terme $xi,\ i = 1,...n$ sind dasselbe wie das Ersetzen des Termes xi der Reihe durch xs :

$$xs = \frac{xi - \overline{xi}}{s},$$

wobei $\overline{xi}$ und s jeweils der Mittelwert und sind die Standardabweichung von xi.

DOMÄNEN UND EINSCHRÄNKUNGEN

Bestimmte Statistiker haben die Angewohnheit, Stan zu sagen
Darstellen erklärender Variablen in einer linearen Form
Regressionsmodell zur Vereinfachung der numerischen
Schwierigkeiten, die sich aus Matrixberechnungen
ergeben, und zur Erleichterung der Interpretation
Vergleich der Regressionskoeffizienten.
Die Anwendung der Methode der kleinsten Quadrate
Die Verwendung standardisierter Daten führt zu
Ergebnissen, die denen der Originaldaten entsprechen.
Eine Regression wird im Allgemeinen durch modelliert

$$Yi = \sum_{j=1}^{P} \ddot{y}jXij, \qquad i = 1,...,n,$$

wobei Xj die nichtstandardisierten Daten sind. Wenn Xij ist standardisiert (Mittelwert 0 und Varianz 1), $\ddot{y}j$ wird als standardisierter Regressionskoeffizient bezeichnet und entspricht der Korrelation, wenn Y ist ebenfalls standardisiert. Die gleiche Terminologie kann verwendet werden, wenn Xj auf eine Varianz von 1 reduziert wird, aber nicht unbedingt auf einen Mittelwert von Null. Der Koeffizient ÿ0 absorbiert die Differenz.

BEISPIELE

Im folgenden Beispiel die Beobachtungen
beziehen sich auf 13Zementmischungen.Jede Mischung
besteht aus den vier in der Tabelle angegebenen Zutaten
Tisch. Das Ziel des Experiments besteht darin, zu
bestimmen, wie die Größen X1, X2, X3 und X4 funktionieren
Diese vier Inhaltsstoffe beeinflussen die Menge an
Wärme Y , die bei der Aushärtung abgegeben wird
Zement.

Tabelle: Von Zement abgegebene Wärme

Mischen-tur	Zutat X_1	X_2	X_3	X_4	Hitze Y
1	7	26	6	60	78,5
2	1	29	15	52	74,3
3	11	56	8	20	104,3
4	11	31	8	47	87,6
5	7	52	6	33	95,9
6	11	55	9	22	109.2
7	3	71	17	6	102,7
8	1	31	22	44	72,5

Machine Translated by Google

Mischen- tur	Zutat X_1	X_2	X_3	X_4	Hitze Y
9	2	54	18	22	93.1
10	21	47	4	26	115,9
11	1	40	23	34	83,9
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Quelle: Birkes und Dodge (1993)

yi Wärmemengen, die durch die Härtung entstehen i-te Mischung (in Joule);

xi1 Menge der Zutat 1 (Aluminat von Tricalcium) in dieser Mischung;

xi2 Menge der Zutat 2 (Tricalciumsilikat) in der Mischung;

xi3 Menge des Inhaltsstoffs 3 (Aluminoferrit). Tetracalcium) in der Mischung;

xi4 Menge der Zutat 4 (Silikat von Dicalcium) in dieser Mischung.

Das erhaltene Modell einer einfachen linearen Regression ist:

$$Y = 62{,}4 + 1{,}55X1 + 0{,}51X2 + 0{,}102X3 - 0{,}144X4 + \varepsilon .$$

Die folgende Tabelle enthält Daten dazu Zement mit Daten relativ zum Standard erklärende Variablen, die wir mit bezeichnen Xs_1, Xs_2, Xs_3, Xs_4 und Xs

Daten zu Zement mit standardisierten erklärenden Variablen

Mischen- tur	Standardisierte Zutat Xs_1	Xs_2	Xs_3	Xs_4	Hitze Y
1	−0,079	−1,424	−0,901	1,792	78,5
2	−1,099	−1,231	0,504	1,314	74,3
3	0,602	0,504	−0,589	−0,597	104,3
4	0,602	−1,102	−0,589	1,016	87,6
5	−0,079	0,247	−0,901	0,179	95,9
6	0,602	0,440	−0,432	−0,478	109,2
7	−0,759	1,468	0,817	−1,434	102,7
8	−1,099	−1,102	1,597	0,836	72,5

Mischen- tur	Standardisierte Zutat Xs_1	Xs_2	Xs_3	Xs_4	Hitze Y
9	−0,929	0,376	0,973	−0,478	93,1
10	2,302	−0,074	−1,213	−0,239	115,9
11	−1,099	−0,524	1,753	0,239	83,9
12	0,602	1,147	−0,432	−1,075	113,3
13	0,432	1,275	−0,589	−1,075	109,4

Die Schätzung nach der Methode der kleinsten Quadrate des kompletten Modells unter Verwendung der standardisierten erklärende Variablen ergibt:

$$Y = 95{,}4 + 9{,}12Xs_1 + 7{,}94Xs_2 + 0{,}65Xs_3 - 2{,}41Xs_4 .$$

Beachten Sie, dass der Koeffizient $\beta_1 = 9{,}12$ von Xs_1 dürfen durch Multiplikation des Koeffizienten berechnet werden $\beta_1 = 1{,}55$ von X1 im Ausgangsmodell durch die Standardabweichung $s1 = 5{,}882$ von X1. Auch Beachten Sie, dass der Mittelwert $\bar{X}1 = 7{,}46$ von X1 nicht in der Beziehung zwischen γ_s berücksichtigt wird und β_1 aber in der Beziehung zwischen β_0^S und β_0. Nehmen nur die ersten beiden standardisierten Erläuterungen Variablen erhalten wir die Lösung

$$Y = 95{,}4 + 8{,}64Xs_1 + 10{,}3Xs_2 .$$

WEITERLESEN

- Daten
- Ridge-Regression

VERWEISE

Birkes, D., Dodge, Y.: Alternative Methoden der Regression. Wiley, New York (1993)

Statistiksoftware

Bei statistischer Software handelt es sich um eine Reihe von Computerprogrammen und -verfahren zur Behandlung von statistische **Daten.**

Machine Translated by Google

Bibliotheken

Eine Bibliothek ist eine Reihe von Programmen oder Unterroutinen, die unter derselben installiert werden Betriebssystem (DOS, Windows oder Unix) und kann durch Eingabe individueller Befehle genutzt werden.

IMSL (Internationale Mathematik und Statistikbibliothek): eine Sammlung von ca 540 in FORTRAN geschriebene Unterprogramme speziell in Bezug auf Mathematik und Statistiken.

NAG (Gruppe Numerische Algorithmen): eine Bibliothek von Algorithmen, die mit dem geschrieben wurden Hilfe von drei Sprachen (ALGOL 60, ALGOL 68 und FORTRAN ANSI). Wir finden hier mehr als 600 Fortgeschrittene Unterprogramme zur **Simulation,** Regressionsanalyse, Optimierungstechniken , GLIMs (allgemeine lineare Modelle), GenStat, Zeitreihen und grafische Darstellungen Darstellungen. Darüber hinaus können wir behandeln auf einfache Weise Antworten erhalten durch Volkszählung.

Software

Eine Software ist eine Reihe komplexer Algorithmen stellt eine gemeinsame Datenstruktur dar und erfordert ein Minimum an vorheriger Programmierung seitens des Benutzers.

BMDP: eine Reihe von Unterprogrammen, die in FOR geschrieben sind TRAN, die es ermöglichen, Anfragen zu bearbeiten durch **Umfrage erstellt,** um **grafische Darstellungen zu erstellen, um** multivariate Techniken oder **nichtparametrische Tests anzuwenden,** oder eine lineare oder nichtlineare durchzuführen **Regressionsanalyse** ohne Vergessen das Studium von **Zeitreihen.**

IMSL (Internationale Mathematik und Statistikbibliothek): Satz von Unterprogrammen geschrieben in FORTRAN speziell für die Bereiche Mathematik und **Statistiken.**

MINITAB: mit Abstand das am einfachsten zu bedienende System. Da es interaktiv genutzt werden kann Im Modus können Analysen durchgeführt werden verschiedener Art: **deskriptive Statistik, Varianzanalyse, Regressionsanalyse , nichtparametrische Tests, Zufallszahlengenerierung** oder die Untersuchung von **Zeitreihen,** unter anderem.

NAG: eine Bibliothek von **Algorithmen ,** mit denen geschrieben wurde die Hilfe von drei Sprachen (ALGOL 60, ALGOL 68, FORTRAN ANSI). Es stehen erweiterte Unterprogramme zur Verfügung zum Thema **Simulation, Regression** Analyse, **Optimierungstechniken , Zeitreihen** und **grafische Darstellungen.** Darüber hinaus wurden die Antworten erhalten durch **Umfrage** kann leicht behandelt werden, und Sowohl multivariate Analysen als auch **nichtparametrische Tests** können angewendet werden ihnen.

P-STAT: ein interaktives System, das behandelt **Versuchspläne ,** grafische **Darstellungen, nichtparametrische Tests,** und **Datenanalyse.**

SAS (Statistisches Analysesystem): eine Reihe von Software einschließlich einer grundlegenden SAS-Sprache sowie spezifischer Programme. Sicherlich das umfassendste Paket ermöglicht **die Analyse von Daten aus** sehr unterschiedlichen Bereichen Felder.

SPSS (Statistical Package for the Social Naturwissenschaften): eine der häufigsten gebrauchte Systeme. Es bietet interaktive Möglichkeiten der Behandlung großer **Datenbanken .** Auch einige **Algorithmen** zur **grafischen Darstellung** sind im Programm enthalten.

Machine Translated by Google

BMDP: Grundsätzlich ist BMDP eine komplette Software für biomedizinische Anwendungen Kit bestehend aus eingeschriebenen Unterprogrammen FORTRAN, das die Bearbeitung von Anfragen ermöglicht durch Volkszählung erstellt , um grafische Darstellungen zu generieren Darstellungen, um multivariate anzuwenden Techniken und nichtparametrische Tests, oder eine lineare oder nichtlineare **Regressionsanalyse** durchzuführen , ohne die Untersuchung von Zeitreihen zu vernachlässigen. Jedes Unterprogramm ist basierend auf äußerst wettbewerbsfähigen Zahlen Algorithmen.

DataDesk: eine interaktive Software für Ana Analyse statistischer Datenverarbeitungen. Es stellt Werkzeuge grafisch dar und ermöglicht eine um Zusammenhänge zwischen Daten, Tendenzen, Untergruppen und Ausreißer.

EViews: Software im Wesentlichen für Wirtschaftswissenschaftler, EViews ist das Werkzeug der Wahl für Analyse, Prognose und ökonomische Modellierung. EViews können verwendet werden für allgemeine statistische Analysen, die Schätzung von Zeitreihen, groß angelegte Simulationen, Grafiken oder einfaches Datenmanagement.

Excel: ermöglicht dem Benutzer die Darstellung und Analyse von Daten mithilfe grundlegender statistischer Funktionen Methoden: Deskriptive Statistik, Analyse von Varianz, Regression und Korrelation, und Hypothesentests.

Gauss: ein statistisches Programm mit orientierter Matrizen, Gauss ist ein leistungsstarkes ökonomisches und grafisches Werkzeug.

JMP: Statistiksoftware, insbesondere für experimentelle Designs, auch JMP enthält eine umfassende Sammlung von Statistiken kalische Werkzeuge.

Lisrel: Wird in den Sozialwissenschaften und mehr verwendet speziell für die faktorielle Analyse und Modellierung ist Lisrel besonders nützlich zur Modellierung der Darstellung von Latent Variablen.

Maple: eine vollständige mathematische Umgebung, die für die Manipulation von entwickelt wurde algebraische Ausdrücke und die Auflösung von Gleichungen und Integralen; es auch verfügt über zwei sehr leistungsstarke Grafiktools und drei Dimensionen. Der Benutzer kann auch Programm in Maple.

Mathematica: Statistiksuite, die es ermöglicht deskriptive univariate durchzuführen und Multivariate statistische Analyse, Daten Glättung, klassische Hypothesentests, Schätzung des Konfidenzintervalls und lineare und nichtlineare Regression.

Matlab: eine interaktive Umgebung, die kann zur Analyse wissenschaftlicher und verwendet werden statistische Daten. Die grundlegenden Objekte von Matlab sind Matrizen. Der Benutzer kann numerische Analysen und Behandlungen durchführen Zeichen und Bildbearbeitung. Der Benutzer kann die Funktionen mit C und programmieren FORTRAN.

MIM: ein Programm zur grafischen Modellierung von diskrete und kontinuierliche Daten. Unter Als Familien verfügbarer Grafikprogramme in dieser Software-Suite erwähnen wir logarithmische lineare Modelle, Gaußsche Grafikmodelle, Manova-Standardmodelle, und viele andere Modelle, die für die multivariate Analyse nützlich sind.

P-STAT: 1960 für die Behandlung entwickelt der pharmakologischen Daten ist P-STAT jetzt verfügbar von Forschern in den Sozialwissenschaften verwendet. Es ermöglicht die Durchführung einer Analyse von Volkszählungen. Die Leistung von P-STAT liegt in seinen Makros und ist für die damit verbundenen statistischen Anwendungen sehr nützlich zum Geschäft. Es ermöglicht die Verarbeitung großer Datenmengen legt experimentelle Designs fest und behandelt sie,

Machine Translated by Google

grafische Darstellungen, nichtparametrische Tests und **Datenanalyse.**

R- Programmiersprache ähnlich S, Ris auch in der Wissenschaft weit verbreitete Freeware. Es stellt eine andere Implementierung als S dar, aber viele S-Befehle funktionieren in R. R enthält eine Vielzahl statistischer Befehle (linear und nichtlinear). Modellierung, Durchführung klassischer Tests, Analyse von Zeitreihen, Klassifizierung, Segmentierung, . . .) sowie grafisch Techniken.

SAS (Statistical Analysis System): eine Suite von Softwareprogrammen inklusive einer Grundversion SAS-Sprache sowie spezifische Programme. Zweifellos das vollständigste Paket, das die Analyse von Daten ermöglicht verschiedene Ursprünge: Forschung, Technik, Medizin und Geschäftsanwendungen. Es umfasst eine beträchtliche Anzahl statistischer Werkzeuge: Varianzanalyse, Regressionsanalyse, kategoriale Datenanalyse, multivariate Analyse, Überlebensanalyse, Entscheidungsbäume, nichtparametrische Analyse, und „Data-Mining“.

S-plus: Software in einem leistungsstarken Format geschrieben und flexible objektorientierte Programmiersprache namens S ermöglicht Matrix-, mathematische und statistische Berechnungen. Es enthält ein Werkzeug für explorative visuelle Analyse, die auch ermöglicht eine erweiterte Datenanalyse sowie deren Modellierung. S-plus ermöglicht Geostatistiker, um einen Block zusammenzustellen „S+SpatialStats“ enthält verschiedene Werkzeuge der **räumlichen Statistik.**

Spad-T: basiert ausschließlich auf statistischen Techniken der multivariaten Analyse. Kontingenztabellen der behandelten Variablen sind mit Hilfe multifaktorieller Methoden erstellt Methoden weichen grafischen Darstellungen.

SPSS (Statistical Package for the Social Wissenschaften): erstmals 1968 veröffentlicht, SPSS ist tatsächlich eines der am weitesten verbreiteten Programme von Statistikern in verschiedenen Domänen. Es bietet interaktive Möglichkeiten zur Bearbeitung großer Datenbanken. Es enthält Algorithmen für grafische Darstellungen.

Stata: eine Statistiksoftware, die es einem ermöglicht Führen Sie statistische Analysen linear durch Modellierung auf einem verallgemeinerten linearen Modell, Behandlung binärer Daten, Verwendung nichtparametrischer Daten Methoden, führen multivariate Analysen durch und Analyse von Zeitreihen und nutzen Methoden der grafischen Datenexploration. Darüber hinaus ermöglicht die Software die Programmierung der benutzerspezifischen Befehle. Stata enthält unter anderem statistische Angaben Werkzeuge für Epidemiologen.

Statistica: erstmals 1993 veröffentlicht, Statisti ca ist eine äußerst komplexe Software Suite, die leistungsstarke Erkundungen enthält Tools für große Datenbanken. Das Grundlegende Die Software umfasst grundlegende Statistiken, Varianzanalyse, nichtparametrische Statistiken, lineare und nichtlineare Modelle, Techniken der explorativen multivariaten Analyse, neuronale Netze und die Berechnung von Stichproben Abmessungen, Werkzeuge zur experimentellen Entwurfsanalyse, Werkzeuge zur Qualitätskontrolle usw.

StatView: Statistiksoftware für medi technische Wissenschaftler und Forscher. StatView ermöglicht die Erstellung deskriptiver Statistiken, parametrisch und nichtparametrisch durchführen Hypothesentests, Korrelationen und Kovarianzen berechnen, lineare, nichtlineare und logistische Regression, erstellen Kontingenztabellen und führen sie durch Faktorielle Analyse, Überlebensanalyse,

Machine Translated by Google

Qualitätskontrolle und Pareto-Analyse.
Eine gute grafische Oberfläche ist verfügbar.

Systat: leistungsstarke Grafiksoftware. Ermöglicht für die Visualisierung von Daten sowie die Verwendung grundlegender Statistiken, um eine angemessene Lösung zu finden Modell.

Solas: Software zur Behandlung fehlender Daten. Solas integriert Algorithmen der multiplen Imputation.

Vista: Entwickelt für Einsteiger Statistik sowie für Professoren. Vista Kann als Tutorial-Software verwendet werden der Bereich der Univariaten, Multivariaten, grafische und rechnerische Statistiken.

WinBugs: ein Programm mit Bayes'schen Daten Modellierung basierend auf Monte-Carlo-Simulationstechniken. Es beinhaltet die Darstellung von Bayes'schen Modellen mit Befehle, mit denen Simulationen durchgeführt werden können. Darüber hinaus steht eine Schnittstelle zur Verfügung, über die die Analyse gesteuert werden kann jeder Zeit. Die grafischen Werkzeuge ermöglichen eine, um die Konvergenz der zu steuern Simulation.

XploRe: eine umfassende statistische Umgebung für die explorative Datenanalyse. Es ist in der Matrix-orientierten Sprache programmiert und enthält als solche eine große Gruppe statistischer Operationen und verfügt über eine leistungsstarke interaktive Funktion Schnittstelle. Das Programm unterstützt benutzerdefinierte Makros.

GESCHICHTE

Der erste Versuch von Statistikern, ihre statistischen Berechnungen zu vereinfachen war die Erstellung statistischer Tabellen. Mit die Ankunft von Computern Mitte der 1940er Jahre, Die Kodierung statistischer Unterprogramme dauerte Ort. Die Programmiersprachen dabei Zeit waren den Forschern nicht allgemein bekannt, und sie mussten bis zum Anfang warten 1960, als die Programmiersprache FORTRAN wurde für die Entwicklung statistischer Softwareprogramme freigegeben. Zu den ersten von IBM eingeführten Routinen gehörte Am bekanntesten ist die Pseudozufallszahl Generator RANDU. Die systematische Entwicklung statistischer Software begann im Jahr 1960 mit der Entstehung der Programme BDM und SPSS, gefolgt von SAS.

VERWEISE

Asimov, D.: Die große Tour: ein Werkzeug zum Anzeigen mehrdimensionaler Daten. SIAM J. Sci. Stat. Berechnen. 6(1), 128–143 (1985)

Becker, RA, Chambers, JM, Wilks, AR: Die neue S-Sprache. Waldsworth und Brooks/Cole, Pacific Grove, CA (1988)

Dixon, WJ Chefredakteur: BMDP Statistical Softwarekommunikation. The University of California Press. Berkeley, Kalifornien (1981)

Chambers, JM und Hastie, TK (Hrsg.): Sta tistische Modelle in S. Waldsworth und Brooks/Cole, Pacific Grove, CA (1992)

Dodge, Y., Hand, D.: What Should Future Wie sieht eine Statistiksoftware aus? Berechnen. Stat. Daten Anal. (1991)

Dodge, Y., Whittaker, J. (Hrsg.): Computational Statistics, Bde. 1 und 2. COMP STAT, Neuchâtel, Schweiz, 24–28 August 1992 (Englisch). Physica-Verlag, Heidelberg (1992)

Edwards, D.: Einführung in die Grafik Modellieren. Springer, Berlin Heidelberg New York (1995)

Francis, I.: Statistische Software: eine vergleichende Überprüfung. Nordholland, New York (1981)

Machine Translated by Google

Hayes, AR: Statistische Software: Eine Übersicht und Kritik ihrer Entwicklung. Büro für Marineforschung, Arlington, VA (1982)

Internationale Mathematik und Statistik Bibliotheken.: IMSL-Bibliotheksinformationen. IMSLK, Houston (1981)

Ripley, BD, Venables, WN (2000) S Programming. Springer, Berlin Heidelberg New York

Statistische Wissenschaften: Statistische Analyse in S-Plus, Version 3.1. Seattle: StatSci, eine Abteilung von MathSoft (1993)

Venables, WN, Ripley, BD: Moderne angewandte Statistik mit S-PLUS. Springer, Berlin Heidelberg New York (1994)

Wegman, EJ, Hayes, AR: Statistische Software. Enzykl. Stat. Wissenschaft. 8, 667–674 (1988)

Statistische Tabelle

Eine statistische Tabelle gibt Werte der **Verteilungsfunktion** oder der individuellen **Wahrscheinlichkeit** einer **Zufallsvariablen** an , die einer bestimmten **Wahrscheinlichkeitsverteilung folgt.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Notwendigkeit einer statistischen Tabelle ergibt sich aus der Tatsache, dass bestimmte Verteilungsfunktionen tionen können nicht in einer einfach zu verwendenden mathematischen Form ausgedrückt werden. In anderen Fällen ist es unmöglich (oder zu schwierig), ein Primitiv der **Dichtefunktion** für die direkte Berechnung der entsprechenden Verteilungsfunktion zu finden.

Die Werte der Verteilungsfunktion sind vorab berechnet und in einer statistischen Tabelle dargestellt, um die Verwendung zu vereinfachen.

BEISPIELE

Beispiele für statistische Tabellen sind die **Normaltabelle,** die **Student-Tabelle,** Fisher-Tabellen und das Chi-Quadrat.

WEITERLESEN

- **Binomialtabelle**
- **Chi-Quadrat-Tisch Verteilungsfunktion**
- **Fisher-Tisch**
- **Kruskal-Wallis-Tisch**
- **Normaler Tisch**
- **Studententisch**
- **Von Wilcoxon signierter Tisch Wilcoxon-Tisch**

VERWEISE

Kokoska, S., Zwillinger, D.: CRC Standard Probability and Statistics Tables und für Mulae. CRC Press (2000)

Statistik

Statistik ist das Ergebnis der Anwendung einer Funktion auf eine Datenstichprobe, wobei die Funktion selbst unabhängig von der Stichprobenverteilung ist.

MATHEMATISCHE ASPEKTE

Eine Statistik ist eine beobachtbare Zufallsvariable fähig, dies unterscheidet ihn von einem Parameter. Vollständigkeit, Suffizienz und Unvoreingenommenheit sind wichtige wünschenswerte Eigenschaften der Statistik.

BEISPIELE

Arithmetisches Mittel, Varianz, Standardabweichung.

WEITERLESEN

- **Chi-Quadrat-Test der Unabhängigkeit**
- **Perzentil**

Machine Translated by Google

p-Wert
Quantil
Schülertest

Statistiken

Das aus dem Lateinischen abgeleitete Wort Statistik bezieht sich auf den Begriff des Staates (Status): „Was ist relativ zum Staat“. Regierungen haben ein großes Bedürfnis, zahlreiche Ereignisse und Aktivitäten, wie z. B. Veränderungen, zu zählen und zu messen Demografie, Geburten, Einwanderungs- und Auswanderungstrends, Veränderungen der Beschäftigungsquoten, Unternehmen usw.

In dieser Perspektive ist der Begriff „Statistik“. Wird verwendet, um einen Satz verfügbarer Daten über anzuzeigen ein bestimmtes Phänomen (z. B. Arbeitslosigkeit). Statistiken).

Im moderneren und genaueren Sinne In diesem Sinne gilt „Statistik“ als eine Disziplin, die sich mit numerischen Daten befasst. Es besteht aus einer Reihe von Techniken für Erkenntnisse aus unvollständigen Daten gewinnen, von einem strengen wissenschaftlichen System zur Verwaltung der Datenerhebung, ihrer Organisation, Analyse und Interpretation, sofern möglich um sie in numerischer Form darzustellen.

In einer Population von Individuen möchten wir im Sinne der statistischen Theorie zu wissen, ob eine bestimmte Person ein Auto hat oder ob sie raucht. An Andererseits würden wir gerne wissen, wie Viele Menschen haben ein Auto und sind Raucher. und ob es eine **Beziehung** zwischen Besitzen gibt ein Auto und Rauchgewohnheiten in der untersuchten Bevölkerung.

Wir möchten die Merkmale der Bevölkerung weltweit kennen lernen, ohne uns um jede Person oder jedes Objekt in der Welt zu kümmern Bevölkerung.

Wir unterscheiden zwei Untergruppen von Techniken: (1) diejenigen, die **deskriptive Statistiken** beinhalten und (2) solche, die **inferenzielle Statistiken** beinhalten . Das wesentliche Ziel der deskriptiven Statistik besteht darin, Informationen in einem verständlichen und nutzbaren Format darzustellen. Inferenz Statistiken hingegen zielen darauf ab, die Verallgemeinerung dieser Informationen zu erleichtern bzw. genauer gesagt, um Schlussfolgerungen (in Bezug auf Populationen) auf der Grundlage von Stichproben zu ziehen diese Populationen.

GESCHICHTE

Der Begriff „Statistik“ stammt aus dem Lateinischen „status“ (Zustand), wurde zum ersten Mal verwendet, nach Kendall, MG (1960), von einem Historiker namens Ghilini, Girolamo im Jahr 1589.

Laut Kendall, MG (1960) das Wahre Der Ursprung der modernen Statistik geht auf zurück 1660.

Statistische Methoden und statistische Verteilung Butionen entwickelten sich hauptsächlich im 19. Jahrhundert Jahrhundert dank der wichtigen Rolle von Statistik in der experimentellen und menschlichen Wissenschaft zen.

Im 20. Jahrhundert wurde die Statistik aufgrund des Reichtums und der Vielfalt zu einer eigenständigen Disziplin Vielfalt der verwendeten Methoden. Das auch erklärt die Geburt spezialisierter statistischer Enzyklopädien wie der *International Encyclopedia of Statistics* (1978) veröffentlicht in zwei Bänden und *die Encyclopedia of Statistical Sciences* (1982–1988) in neun Bänden umes.

DOMÄNEN UND EINSCHRÄNKUNGEN

Statistiken werden in vielen Bereichen verwendet sie unterscheiden sich sehr voneinander. Daher Wir finden, dass es in der industriellen Produktion eingesetzt wird in der Umwelt- und Medizinwissenschaft, as sowie in offiziellen Stellen.

Schauen wir uns einige der Bereiche an, in denen dies der Fall sein könnte Interesse eines Statistikers:

Machine Translated by Google

- *Grundlagenforschung in der Statistik:* Forschung in den Bereichen Wahrscheinlichkeiten, statistische Methoden und Theorie.
- *Biologie:* Grundlagenforschung und Experimente im Zusammenhang mit den Hauptphänomenen lebender Organismen und Biometrie.
- *Handel:* Datenmanagement, Verkaufsvolumen, Management, Inventarmethoden, Industrieplanung, Kommunikation und theoretische Kontroll- und Buchhaltungsverfahren.
- **Demografie:** Untersuchung der Zunahme menschliche Bevölkerung (Geburten- und Sterberate, Migrationsbewegung), Untersuchung der Struktur der Bevölkerung (persönliche, soziale und wirtschaftliche Merkmale).
- *Wirtschaft:* Maß für das Produktionsvolumen, den Handel, die Ressourcen, die Beschäftigung und den Lebensstandard; Analyse von das Verbraucher- und Herstellerverhalten, Reaktionen des Marktes auf Preisänderungen, Auswirkungen von Werbung und Regierung Richtlinien.
- *Bildung:* Maßnahmen und Tests im Zusammenhang mit der Prozess des Lernens.
- *Ingenieurwesen:* Forschung und Experimente einschließlich Design- und Effizienztests, Verbesserungen der Testmethoden, Fragen im Zusammenhang mit **Inferenz-** und Präzisionstests, Verbesserungen der Qualitätskontrollmethoden.
- *Gesundheit:* Kosten für Unfälle, medizinische Versorgung, Krankenhausaufenthalt.
- *Versicherung:* Bestimmung der Sterblichkeit und Unfallraten der Versicherten und von die allgemeine Bevölkerung; Bestimmung der Leitzins für jede **Risikokategorie ;** Anwendung statistischer Techniken auf Verwaltung von Versicherungspolicen.
- *Marketing und Konsumforschung:* Probleme im Zusammenhang mit Märkten, der Vertriebssystem, Erkennung von Trends; Untersuchung der Verbraucherpräferenzen und des Verbraucherverhaltens.
- *Medizin:* Epidemiologie, Grundlagen Forschung und Experimente zu den Ursachen, Diagnose, Behandlung und Prävention von Krankheiten.
- *Management und Verwaltung:* Probleme im Zusammenhang mit der Personalführung, Ausrüstung, Produktionsmethoden und Arbeitsbedingungen.
- *Psychologie und Psychometrie:* Probleme im Zusammenhang mit der Auswertung der Studie Fähigkeiten, Intelligenz, persönliche Eigenschaften sowie normales und abnormales Verhalten einer Person sowie der Einrichtung ment von Maßstäben und Instrumenten Maßangaben.
- *Sozialwissenschaften:* Etablierung von Tech Techniken der **Probenahme** und theoretische Tests in Bezug auf soziale Systeme und gesellschaftliche Wohlbefinden. Analyse der Kosten der Sozialwissenschaften; Analyse kultureller Unterschiede auf Werteebene und sozialer Ebene Verhalten.
- *Weltraumforschung:* Interpretation von Experimenten und Analyse der bei Weltraummissionen gesammelten Daten.
- *Naturwissenschaften (allgemein):* grundlegend Forschung in den Natur- und Sozialwissenschaften zen.

WEITERLESEN

Bayesianische Statistik
Beschreibende Statistik
Genetische Statistiken
Inferenzstatistik
Offizielle Statistiken
Raumstatistik

Machine Translated by Google

VERWEISE

Kendall, MG: Wo soll die Geschichte von Statistiken beginnen? Biometrie **47,** 447–449 (1960)

Kotz, S., Johnson, NL: Enzyklopädie von StatisticalSciences, Bd. 1–9. Wiley, Neu York (1982–1988)

Kruskal, WH, Tanur, JM: International Encyclopedia of Statistics, Bde. 1–2. Der Freie Presse, New York (1978)

Stigler, S.: Die Geschichte der Statistik, die Messung der Unsicherheit vorher 1900. Belknap, London (1986)

Stiel und Blatt

Siehe **Stamm-Blatt-Diagramm.**

Stamm-Blatt-Diagramm

Das Stamm-Blatt-Diagramm ist ein grafisches Diagramm Darstellung von **Frequenzen.** Das Grundlegende Die Idee besteht darin, Informationen über die **Häufigkeitsverteilung** bereitzustellen und die **Werte** beizubehalten der Daten gleichzeitig. Tatsächlich, der Stamm entspricht den Klassenintervallen **und** dem Blatt zur Anzahl der **Beobachtungen** im Klasse, die durch die verschiedenen **Daten dargestellt wird.** Es ist Dann ist es möglich, die **Werte** von direkt abzulesen die Daten.

GESCHICHTE

Der Ursprung des Stamm-Blatt-Diagramms ist oft mit Tukey, JW (1977) in Verbindung gebracht. Es ist Das Konzept basiert auf dem **Histogramm,** das stammt aus dem 18. Jahrhundert.

MATHEMATISCHE ASPEKTE

Um jeweils ein Stamm-Blatt-Diagramm zu erstellen Zahl muss in einen Hauptteil (Stamm) geschnitten werden und einen sekundären Teil (Blatt). Zum Beispiel 4.2, 42 oder 420 können für den Stiel in 4 Teile geteilt werden und 2 für das Blatt. Die Stämme werden dann aufgelistet vertikal (eine Zeile pro Stamm) in der Reihenfolge angezeigt Größe und die Blätter stehen daneben sie, auch in der Größenordnung. Wenn ein Satz von **Die Daten** umfassen die **Werte** 4,2, 4,4, 4,8, 4,4, und 4.1, dann sind Stamm und Blatt wie folgt:

4|12448 .

DOMÄNEN UND EINSCHRÄNKUNGEN

Das Stamm-Blatt-Diagramm liefert Informationen über Verteilung, Symmetrie, Konzentration, leere Mengen und **Ausreißer.**

Manchmal gibt es zu viele Zahlen der gleiche Stamm. In diesem Fall ist es nicht möglich um sie alle auf eine Linie zu bringen. Am meisten Die übliche Lösung besteht darin, die zu trennen zehn Blätter in zwei Gruppen von fünf Blättern, z. B. 4|0122444 und 4|567889.

Es ist möglich, zwei **Datenreihen** zu vergleichen indem man die beiden Blätter nach links und rechts legt des gleichen Stammes:

012244|4|567889 .

BEISPIELE

Pro-Kopf-BSP – 1970 (in Dollar)

AFRIKA			
Algerien	324	Mauritius	233
Angola	294	Marokko	224
Benin	81	Mosambik	228
Botswana	143	Niger	90
Burundi	67	Nigeria	145
Zentral	127	Wiedervereinigung	769
Afrikanischer Repräsentant.			
Tschad	74	Ruanda	60
Komoren	102	Senegal	236

Machine Translated by Google

AFRIKA			
Kongo	238	Sierra Leone	167
Ägypten	217	Somalia	89
Gleich. Guinea	267	Südafrika	758
Äthiopien	72	Rhodesien (Süd)	283
Gabun	670	Sudan	117
Gambia	101	Swasiland	270
Ghana	257	Togo	135
Guinea	82	Tunesien	281
Guinea Bissau	259	Uganda	135
Elfenbeinküste	347	Vereinigte Rep. von Kamerun	185
Kenia	143		
Lesotho	74	Vereinigte Rep. von Tansania	100
Liberia	268		
Madagaskar	133	Hochvolt	59
Malawi	74	Zaire	88
Mali	54	Sambia	421
Mauretanien	165		

Quelle: Jahresstatistik der UN, 1976

Die obige Tabelle gibt das Pro-Kopf-BSP in an
Dollar, für die meisten afrikanischen Länder im Jahr 1970.
Erstellen eines Stamm-Blatt-Diagramms aus
Diese **Daten** werden zunächst der Reihe nach klassifiziert
der Größenordnung. Der Stamm wird dann ausgewählt: in
In diesem Beispiel werden wir Hunderter verwenden. Die Zehner
Die Nummer wird dann rechts daneben platziert
entsprechende Hundert, was das Blatt ergibt
Teil des Diagramms. Die Einheiten werden nicht übernommen
berücksichtigen.
Der Stiel und das Blatt sind wie folgt:

Einheit 1	1 = 110
0	5566777788889
1	00012333444668
2	12233355667889

Einheit 1	1 = 110
3	24
4	2
5	
6	7
7	56

WEITERLESEN

- Häufigkeitsverteilung
- Grafische Darstellung
- Histogramm

VERWEISE

Tukey, JW: Erklärende Datenanalyse,
limitierte Vorabauflage. Addison Wesley,
Reading, MA (1970–1971)

Stochastischer Prozess

Ein stochastischer Prozess ist eine Familie von Zufallsvariablen. In der Praxis dient es der Modellierung einer Vielzahl zeitlicher Phänomene wo der Zufall ins Spiel kommt.
Wir unterscheiden viele Arten stochastischer Prozesse anhand bestimmter mathematischer Eigenschaften. Die bekanntesten sind:

- Markov-Prozess (oder Markov-Kette)
- Martingal
- Stationärer Prozess
- Prozess unabhängiger Inkremente

S

GESCHICHTE

Ursprünglich kam es zu stochastischen Prozessen von Fortschritten im frühen 20. Jahrhundert in bestimmte angewandte Zweige der Statistik wie z Statistische Mechanik (von Gibbs, JW, Boltzmann, L., Poincaré, H., Smoluchowski, M., und Langevin, P.). Die theoretischen Grundlagen wurden später von Doob, JL, Kolmogorov, AN und andere (1930–1940).
In dieser Zeit wurde das Wort „stochastisch“

Machine Translated by Google

vom griechischen Wort für „raten, glauben,
Stellen Sie sich vor“, begann zu verwenden. Andere Fortschritte wurden ausgehend von einer Brownschen Bewegung in der Physik gemacht (von Einstein, A., Levy, P., und Wiener, N.).

Der Markov-Prozess wurde erfunden von **Markow, Andrej Andrejewitsch** (1856–1922), der den Grundstein für die Theorie legte von Markov-Prozessen in endlicher Zeit, verallgemeinert vom russischen Mathematiker Kol mogorov, A. (1936).

MATHEMATISCHE ASPEKTE

Ein stochastischer Prozess ist eine Familie (Xt)tÿI (wobei *I* eine diskrete oder kontinuierliche Menge ist) von Zufallsvariablen, die auf demselben Wahrscheinlichkeitsraum definiert sind. Die Prozesse *in diskreter Zeit* werden durch die folgenden Familien vertreten (Xt)tÿ (das istmitI =) und die Prozesse *in kontinuierlicher Zeit* durch die Familien (Xt)tÿ+ (also mit *I* = = [0, ÿ)). +

Für Satz *I* wird der Begriff „stochastischer Prozess“ verwendet ist unendlich. In der Physik repräsentieren die Elemente von *I* die Zeit.

Für ein **Ereignis** *w, das* zur **Grundmenge gehört,** definieren wir *die Trajektorie* des untersuchten Prozesses als die folgende Familie:

$$(Xt\,(w))tÿI\ ,$$

so dass die „Reihe“ (für einen Prozess in diskreter Form) Zeit) der Werte der Zufallsvariablen
wird für ein ausgewähltes und festgelegtes Element angenommen.

Die Hauptziele der Zufallsforschung Prozess sind die Eigenschaften von Zufallsvariablen und Trajektorien.

Wir charakterisieren die verschiedenen Arten von Prozessen anhand der Wahrscheinlichkeitsbedingungen mit denen sie in Zusammenhang stehen.

1. *Markov-Prozess*

   Ein Markov-Prozess ist ein Zufallsprozess (Xt)tÿ in diskreter Zeit, für die die Wahrscheinlichkeit eines Ereignisses (in der Zeit *t* + 1) hängt nur von der unmittelbaren Vergangenheit ab (Zeit t).

   Dies wird mathematisch ausgedrückt durch folgende Bedingung:

   Für jedes *t* ÿ und für jedes i0, *i1,...,* it+1 ÿ + wir haben:

$$P(Xt+1 = it+1|Xt = it, Xtÿ1 = itÿ1,...,X0 = i0) = P(Xt+1 = it+1|Xt = it)\ .$$

2. *Martingal*

   Ein Martingal ist ein Zufallsprozess (Xt)tÿ in diskreter Zeit, die Folgendes bestätigt Zustand:

   Für jedes *t* ÿ :

$$E(Xt+1\,|Xt, Xtÿ1,...,X0) = Xt\ .$$

3. *Stationärer Prozess*

   Ein stationärer Prozess (im engeren Sinne). der Term) ist ein Zufallsprozess(Xt)tÿ für wobei die gemeinsamen Verteilungen des Prozesses nur von der Zeitdifferenz abhängen (zwischen den betrachteten Variablen in der gemeinsame Verteilung) und nicht auf den „absoluten“ Zeitpunkt. Mathematisch:

   Für jedes t1, *t2,...,tk* ÿ I(k ÿ 2) und für jedes *a* ÿ I, *Xt1 , Xt2 ,...,Xtk* haben die gleiche gemeinsame Verteilung wie Xt1+a, *Xt2+a,...,Xtk+a.*

4. *Prozess mit unabhängigen Inkrementen*

   Ein Prozess mit unabhängigen Inkrementen ist ein Zufallsprozess (Xt)tÿI für den:

   Für jedes t1 < t2 < < *tk* ÿ I÷*Xt2* ÿ *Xt1 , Xt3 ÿXt2 ,...,Xtk ÿXtkÿ1* der Zufall Variablen sind wechselseitig unabhängig.

DOMÄNEN UND EINSCHRÄNKUNGEN

Theoretisch gibt es keinen Unterschied zwischen das Studium stochastischer Prozesse und der Statistik selbst. In der Praxis konzentriert sich die Untersuchung stochastischer Prozesse mehr auf die Struktur

Machine Translated by Google

und Eigenschaften von Modellen als auf Schlussfolgerungen aus realen Daten, was eigentlich die Domäne der Statistik ist.

BEISPIELE

1. Hier geben wir ein Beispiel, das die engen mathematischen Beziehungen zwischen zeigt Statistik und stochastische Prozesse. Wir betrachten die Zufallsvariable Xi = Alter des Individuums i einer Population (potenziell unendlich). Ein typisches Problem besteht darin, den Mittelwert $E(X)$ mit der folgenden Hypothese zu schätzen : • Die Zufallsvariable Xi ist unabhängig. • Alle Xi sind identisch verteilt.

2. *Zufälliger Spaziergang.* Seien Zn, n ÿ 2, unabhängige und identisch verteilte Zufallsvariablen, die den Wert +1 mit der Wahrscheinlichkeit p und den Wert ÿ1 mit der Wahrscheinlichkeit q = 1 ÿ p annehmen.

   Wir definieren den stochastischen Prozess des Random Walk wie folgt:

$$X0 = 0,\ Xt = \sum_{n=1}^{T} Zn\ ,\ \text{wenn } t > 1 \quad \cdot$$

   Eine Stichprobe eines Random Walk wird typischerweise wie folgt dargestellt:

   (0, *1, 2, 1, 0,* ÿ1, *0,* ÿ1, ÿ2, ÿ3, *ÿ2,...).*

   Es ist leicht zu erkennen, dass der definierte stochastische Prozess ein Markov-Prozess ist (diese Tatsache ist eine direkte Folge der Unabhängigkeit von Xt) mit den folgenden Wahrscheinlichkeiten:

$$P\ (Xt+1 = i + 1\ |Xt = \text{i}) = p\ ,\ P\ (Xt+1 = i$$

$$\text{ÿ}\ 1\ |Xt = \text{i}) = q\ .$$

   Wir haben das:

$$E(Xt+1\ |Xt) = Xt + (+1) \cdot p + (\text{ÿ}1) \cdot q$$

$$= Xt + (p\ \text{ÿ}\ q).$$

   Die Irrfahrt ist genau dann ein Martingal, wenn p = q = 0,5.

WEITERLESEN

Konvergenz
Zufällige Variable
Statistiken
Zeitfolgen

VERWEISE

Barlett, MS: Eine Einführung in stochastische Prozesse. Cambridge University Press, Cambridge (1960)

Billingsley, P.: Statistische Inferenz für Markov-Prozesse. University of Chicago Press, Chicago (1961)

Cox, DR, Miller, HD: Theorie stochastischer Prozesse. Wiley, New York; Methuen, London (1965)

Doob, JL: Was ist ein stochastischer Prozess? Bin. Mathematik. Monatlich **49,** 648–653 (1942)

Doob, JL: Stochastische Prozesse. Wiley, New York (1953)

Kolmogorov, AN: Mathe. Sbornik. NS 1, 607–610 (1936)

Geschichtete Stichprobe

Bei der geschichteten **Stichprobe** wird die **Bevölkerung** zunächst in Teilpopulationen, sogenannte Strata, aufgeteilt. Diese dürfen sich nicht gegenseitig durchdringen, und die Gesamtheit dieser Schichten muss die bilden ganze Bevölkerung.

Sobald die Schichten bestimmt sind, wird aus jeder Schicht eine Zufallsprobe **(nicht** unbedingt von gleicher Größe) entnommen, wobei diese Probenahme in den verschiedenen Schichten unabhängig voneinander durchgeführt wird Schichten.

GESCHICHTE

Siehe **Probenahme.**

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Bezüglich der Bestimmung der Zahl der aus jeder Schicht zu entnehmenden Personen, dort Es gibt hauptsächlich zwei Ansätze: die **repräsentative geschichtete Stichprobe** und die optimal geschichtete Stichprobe.

Repräsentative geschichtete Stichprobe

In jeder Schicht gibt es eine Anzahl von Individuen, die proportional zur Größe der Schicht ist vergriffen.

Sei *Ni* die Größe der Schicht i, *N* die Anzahl der Individuen der Bevölkerung und *n* die Größe der globalen **Stichprobe.** Die Größe von Eine Teilstichprobe *ni* ist definiert durch:

$$ni = \frac{Ni}{N} \cdot n,\ i = 1, 2,...,k,$$

Dabei ist *k* die Anzahl der Bevölkerungsschichten .

Optimale geschichtete Probe

Ziel ist es, die Größe der zu entnehmenden Teilstichproben zu bestimmen, um die bestmögliche **Schätzung zu erhalten.** Dies geschieht durch Minimieren die **Varianz** des **Schätzers** *m* unter dem Zwang

$$\sum_{i=1}^{k} ni = n \quad ,$$

wobei *m* der **Mittelwert** der definierten **Stichprobe** ist von:

$$m = \frac{1}{N} \sum_{i=1}^{k} Nix\bar{}i \ ,$$

wobei x¯i der **Mittelwert** der Teilstichprobe i ist . Das ist ein klassisches **Optimierungsproblem** und gibt folgendes Ergebnis:

$$ni = \frac{N}{\sum_{i=1}^{k} Ni \cdot ÿi} Ni \cdot ÿi \ ,$$

wobei *ÿi* die **Standardabweichung** der Schicht i ist.

Bei einem konstanten Faktor n/ $\sum_{i=1}^{k} Ni \cdot ÿi$, das Die Größe der Teilmenge ist proportional zur Größe der Schicht und zum **Standard Abweichung.**

Je größer die Streuung der untersuchten Variablen in der Schicht und desto größer die Größe *Ni,* desto größer ist die Größe der Teilstichprobe *ni.*

DOMÄNEN UND EINSCHRÄNKUNGEN

Die geschichtete **Probenahme** basiert auf der Idee Erhalten, als Ergebnis der Kontrolle bestimmter Variablen, ein **Beispiel ,** das ein Bild neu erstellt Das ist so nah wie möglich an der **Bevölkerung.** Die Streuung in der **Bevölkerung** kann sein stark; in diesem Fall, um **die Probenahme** aufrechtzuerhalten **Fehler** auf einem akzeptablen Niveau, sehr groß Es muss **eine Stichprobe** erstellt werden, die zunehmen kann die Kosten erheblich. Um dieses Problem zu vermeiden, Eine geschichtete Probenahme kann eine gute Alternative sein. Die Bevölkerung kann so in Schichten unterteilt werden, dass die Variation innerhalb der Schichten ist relativ niedrig, was die reduzieren wird erforderliche **Stichprobengröße** für eine bestimmte Stichprobe Fehler.

Eine geschichtete **Probenahme** kann hierfür gerechtfertigt sein folgende Gründe:

1. Wenn die **Bevölkerung** zu heterogen ist, dann ist vorzuziehen, homogener zu sein Gruppen, um eine **Stichprobe** zu erhalten , die repräsentativer für die Bevölkerung ist.
2. Wenn man Informationen über bestimmte Aspekte einer **Bevölkerung** erhalten möchte (z. B. mit (bzgl. unterschiedlicher Bundesländer) ist hierfür die geschichtete Stichprobe besser geeignet.

Ein Steuerzeichen, das zur Definition der Schichten verwendet wird muss folgende Regeln beachten:

- Es muss in engem Zusammenhang mit dem stehen untersuchte Variablen.

Machine Translated by Google

• Es muss für jede Einheit einen bekannten **Wert** haben der **Bevölkerung.**

Einige der am häufigsten verwendeten Steuerelemente Variablen sind Geschlecht, Alter, Region, sozioprofessionelle Schicht usw.

BEISPIELE

Stellen Sie sich **eine Bevölkerung** von 600 Studenten vor, die in vier Schichten aufgeteilt ist. Uns interessiert das Ergebnisse eines Tests mit Werten im Bereich von 0 bis 100.

Angenommen, wir haben die folgenden Informationen:

Größe Ni	Standardabweichung ÿ
Schicht 1 285	8.6
Schicht 2 150	5.2
Schicht 3 90	2.2
Schicht 4 75	1.4

Wenn wir eine geschichtete **Stichprobe** von 30 nehmen wollen Individuen erhalten wir folgende Teilstichproben:

1. *Repräsentative geschichtete Stichprobe*

$$ni = \frac{Ni}{N} \cdot n$$

Wir erhalten dann die folgenden Werte: auf die nächste ganze Zahl gerundet:

$$n1 = \frac{N1}{N} \cdot n$$

$$= \frac{285}{600} \cdot 30$$

$$= 14{,}25$$

ÿ 14

n2 = *7,5* ÿ 7

n3 = *4,5* ÿ 5

n4 = 3,75 ÿ 4

2. *Optimal geschichtete Stichprobe*

$$ni = \frac{N}{\sum_{i=1}^{k} Ni \cdot \ddot{y}i} \cdot Ni \cdot \ddot{y}i$$

Wir haben:

N1 · ÿ1 = 2451

N2 · ÿ2 = 780

N3 · ÿ3 = 198

N4 · ÿ4 = 105

Der konstante Faktor $n/\sum_{i=1}^{4} Ni \cdot \ddot{y}i$ ist gleich Zu:

$$\frac{N}{\sum_{i=1}^{k} Ni \cdot \ddot{y}i} = \frac{30}{3534} = 0{,}0085 .$$

Wir erhalten folgende gerundete Werte auf die nächste ganze Zahl:

n1 = 0,0085 · 2451 = *20,81* ÿ 21

n2 = 0,0085 · 780 = 6,62 ÿ 6

n3 = 0,0085 · 198 = 1,68 ÿ 2

n4 = 0,0085 · 105 = 0,89 ÿ 1 .

Wenn wir beide Methoden vergleichen, können wir Folgendes erkennen: Bei einer homogeneren Schicht nimmt die zweite Methode eine größere Anzahl von Individuen in der ersten Schicht an (die am weitesten verstreut ist). und eine kleinere Anzahl von Personen in der zweite Schicht.

WEITERLESEN

Einschätzung

Schätzer

Probenahme

Einfache Zufallsstichprobe

VERWEISE

Yates, F.: Stichprobenmethoden für Volkszählungen und Umfragen. Griffin, London (1949)

Student

Siehe **Sealy, Gosset William.**

S

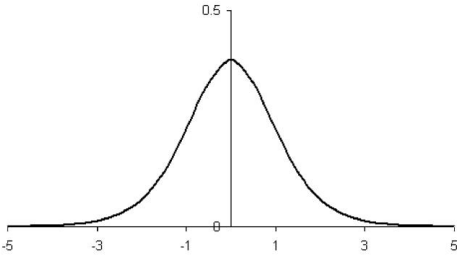
Machine Translated by Google

Studentenverteilung

Die Zufallsvariable *T* folgt einem Schüler Verteilung, wenn ihre **Dichtefunktion** die Form hat:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{2t}{\nu}\right)^{\frac{\nu+1}{2}},$$

Wo ist die Gammafunktion (siehe **Gammaverteilung)** *undvdie* Anzahl **der Freiheitsgrade?**



Studentenverteilung, ν = 7

Die Student-Verteilung ist **kontinuierlich Wahrscheinlichkeitsverteilung.**

GESCHICHTE Die Student-Verteilung wurde **von Gos Set, WS,** bekannt als „Student“, erstellt, der 1908 einen Artikel veröffentlichte, in dem er die **Dichtefunktion** der Differenz zwischen dem **Mittelwert** einer **Stichprobe** und dem Mittelwert der **Grundgesamtheit ,** aus der die Stichprobe entnommen wurde, dividiert durch den **Standard ,** beschrieb **Abweichung** der Probe. Im selben Artikel lieferte er auch die erste Tabelle der entsprechenden **Verteilungsfunktion.**

Der Student setzte seine Forschungen fort und im Jahr 1925 veröffentlichte einen Artikel, in dem er neue, umfangreichere Tabellen vorschlug etwas genauer.

Fisher, RA interessierte sich für Gossets Werke. Er schrieb ihm 1912 einen Antrag eine geometrische Demonstration der Student-Verteilung und Einführung des Begriffs **Freiheitsgrade.** 1925 veröffentlichte er einen Artikel, in dem er das Verhältnis t definiert:

$$t = \frac{Z}{\sqrt{\frac{X}{\nu}}},$$

Dabei sind *X* und *Z* zwei unabhängige **Zufallsvariablen, X** wird gemäß einer **Chi-Quadrat-Verteilung** mit *v* **Freiheitsgraden** verteilt und *Z* wird gemäß der zentrierten und reduzierten **Normalverteilung verteilt.**

Dieses Verhältnis wird als *t* von Student bezeichnet .

MATHEMATISCHE ASPEKTE

Wenn *X* eine **Zufallsvariable** ist , folgt Folgendes eine **Chi-Quadrat-Verteilung** mit *v* **Freiheitsgraden** und *Z* eine gemäß der **Standardnormalverteilung** verteilte Zufallsvariable ist , dann die Zufallsvariable

$$T = \frac{Z}{\sqrt{\frac{X}{\nu}}}$$

folgt einer Student-Verteilung mit *v* Freiheitsgraden, wenn *X* und *Z* unabhängig sind.
Betrachten Sie **die Zufallsvariable** T:

$$T = \frac{\sqrt{n}\,(\bar{x} - \mu)}{S},$$

Dabei ist $\bar{x}$ der **Mittelwert** einer **Stichprobe,** ÿ der Mittelwert **der Grundgesamtheit,** aus der die Stichprobe entnommen wurde, und *S* die **Standardabweichung** der Stichprobe, gegeben durch:

$$S = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

Dann folgt *T* einer Student-Verteilung mit *n* ÿ 1 **Freiheitsgrade.**

Machine Translated by Google

Der **erwartete Wert** der **Zufallsvariablen** *T*

Im Anschluss ist eine Studentenverteilung gegeben durch:

$$E[T] = 0 \quad , \quad v > 1 \quad ,$$

und die **Varianz** ist gleich:

$$\mathrm{Var}(T) = v \ddot{y} \frac{v}{2} \quad , \quad v > 2 \quad ,$$

wobei *v* die Anzahl der **Freiheitsgrade** ist **dom.**

Die Studentenverteilung hängt mit anderen zusammen

kontinuierliche Wahrscheinlichkeitsverteilungen:

- Wenn die Anzahl **der Freiheitsgrade** ausreichend groß ist, wird der Student
  Die Verteilung nähert sich einer **Normalverteilung an.**
- Wenn die Anzahl der **Freiheitsgrade** *v* ist gleich 1, die Student-Verteilung ist identisch mit einer **Cauchy-Verteilung** mit ÿ = 0 und ÿ = 1.
- Wenn **die Zufallsvariable** *X* einem Student folgt Verteilung mit *v* **Freiheitsgraden,** dann folgt die Zufallsvariable X2 einer **Fisher-Verteilung** mit 1 und *v* Grad von Freiheit.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Student-Verteilung ist symmetrisch Verteilung mit einem Mittelwert gleich 0. Je größer die Anzahl der **Freiheitsgrade,** desto größer kleiner wird seine **Standardabweichung .**

Die Student-Verteilung wird in **der Inferenzstatistik** in Bezug auf das *t*- Verhältnis von Stu verwendet

Delle beim **Hypothesentest** und bei der Konstruktion von **Konfidenzintervallen** für die

Mittelwert einer **Bevölkerung.**

In Tests zur **Varianzanalyse** kann die Student-Verteilung als Summe verwendet werden

der Quadrate zwischen den Gruppen (oder die Summe von Quadrate der Faktoren) nur verglichen werden hat einen **Freiheitsgrad.**

Die **Normalverteilung** kann als verwendet werden eine Annäherung an die Studentenverteilung wenn die Anzahl der **Beobachtungen** *n* groß ist *(n* > 30 laut den meisten Autoren).

WEITERLESEN

- **Cauchy-Verteilung**
- **Kontinuierliche Wahrscheinlichkeitsverteilung**
- **Fisher-Verteilung**
- **Normalverteilung**
- **Studententisch**
- **Schülertest**

VERWEISE

Fisher, RA: Anwendungen von „Student's“ Verteilung. Metron 5, 90–104 (1925)

Gosset, SW „Student“: Der wahrscheinliche Fehler eines Mittelwerts. Biometrie 6, 1–25 (1908)

Gosset, SW „Student“: Neue Tabellen für Prüfung der Aussagekraft von Beobachtungen. Metron 5, 105–120 (1925)

Studententisch

Die Student-Tabelle gibt die Werte der an **Verteilungsfunktion** einer **Zufallsvariable fähig** nach einer **Studentenverteilung.**

GESCHICHTE

Gosset, William Sealy, bekannt als „Student“, entwickelte 1908 den **Student-Test** und der Tisch, der seinen Namen trägt.

MATHEMATISCHE ASPEKTE

Die **Zufallsvariable** *T* folge einem **Student Verteilung** mit *v* Freiheitsgraden. Der **Dichtefunktion** der Zufallsvariablen *T* ist gegeben von:

$$f(s) = \frac{\frac{v+1}{2}}{\ddot{y}v\ddot{y} \cdot \frac{v}{2}} \left(1 + \frac{s2}{v}\right)^{\ddot{y}(v+1)/2} \quad ,$$

S

Machine Translated by Google

wo stellt die Gammafunktion dar (siehe **Gammaverteilung).**
Die **Verteilungsfunktion** der Zufallsvariable

fähiges *T* ist definiert durch:

$$F(t) = P(T \ddot{y} t) = \int_{\ddot{y}\ddot{y}}^{T} f(s)\, ds .$$

Für jeden **Wert** von v gibt die Student-Tabelle den Wert der Verteilungsfunktion *F* (t) an.

Da die Student-Verteilung relativ zum Ursprung symmetrisch ist, gilt:

$$F(t) = 1 \ddot{y} F(\ddot{y}t) ,$$

wodurch sich F (t) für jeden negativen Wert von t bestimmen lässt.

Die Studenttabelle wird im umgekehrten Sinne verwendet:

um die Werte zu finden, die einer gegebenen **Wahrscheinlichkeit entsprechen.** In diesem Fall gibt die Studenttabelle den **kritischen Wert** für die Anzahl der **Freiheitsgrade** *v* und das **Signifikanzniveau** ÿ an. Wir bezeichnen im Allgemeinen mit *tv,ÿ* den Wert der Zufallsvariablen *T* , für die:

$$PT \ddot{y} tv,\ddot{y} = 1 \ddot{y} \ddot{y} .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Siehe **Schülertest.**

BEISPIELE

Siehe Anhang E.
Wir betrachten einen einseitigen **Student-Test** eine **Stichprobe** der Dimension n. Für *v* = *n* ÿ 1 Freiheitsgrade ermöglicht die Student-Tabelle die Bestimmung des **kritischen Wertes** *tv,ÿ* des Tests für das gegebene **Signifikanzniveau** ÿ.

Ein Beispiel für die Verwendung der Student-Tabelle finden Sie unter **Student-Test.**

WEITERLESEN

Statistische Tabelle
Studentenverteilung
Schülertest

VERWEISE

Fisher, RA: Anwendungen der „Student“-Verteilung. Metron 5, 90–104 (1925)

Gosset, SW „Student“: Der wahrscheinliche Fehler eines Mittelwerts. Biometrie 6, 1–25 (1908)

Gosset, SW „Student“: Neue Tabellen zur Prüfung der Aussagekraft von Beobachtungen. Metron 5, 105–120 (1925)

Schülertest

Der Student-Test ist ein parametrischer **Hypothesentest** zum **Stichprobenmittelwert** oder zum Vergleich der Mittelwerte zweier Stichproben.

GESCHICHTE

Siehe unter Geschichte der **Student-Tabelle.**

MATHEMATISCHE ASPEKTE

Sei *X* eine verteilte **Zufallsvariable**

gemäß einer **Normalverteilung.** Die oben definierte Zufallsvariable *T* folgt einem **Stu**

Beulenverteilung mit nÿ1 Freiheitsgraden:

$$T = \frac{\bar{X} \ddot{y} \ddot{y}}{\frac{S}{\ddot{y}\sqrt{N}}} ,$$

Dabei ist $\bar{X}$ das **arithmetische Mittel** der **Stichprobe,** ÿ der Mittelwert der **Grundgesamtheit,** *S* die **Standardabweichung** der Stichprobe und *n* die **Stichprobengröße.**

Der Student-Test besteht darin, dieses Verhältnis für die Beispieldaten zu berechnen und es mit dem theoretischen **Wert** des **Student** zu vergleichen **Tisch.**

A. *Studententest für eine oder einzelne Stichprobe*

1. Die Hypothese, die wir testen möchten, ist:
   – **Nullhypothese:**

$$H0 : \ddot{y} = \ddot{y}0 .$$

Machine Translated by Google

Wir sollten testen, ob die Stichprobe aus einer Grundgesamtheit mit einem bestimmten **Mittelwert** ÿ0 stammt oder ob ein statistisch signifikanter Unterschied zwischen dem beobachteten Mittelwert in der Stichprobe und dem hypothetischen Mittelwert der Grundgesamtheit (ÿ0) unter H0 besteht.

– **Alternativhypothese:** Die Alternativhypothese kann drei Formen annehmen:

H1 : ÿ>ÿ0 **(einseitiger Test** rechts)

H1 : ÿ<ÿ0 **(einseitiger Test** links)

H1 : ÿ = ÿ0 **(zweiseitiger Test) .**

2. Berechnung der t-Statistik als:

$$= \frac{x^- \ddot{y} \ddot{y}0\ t}{\frac{S}{\ddot{y}\ N}},$$

Dabei ist ÿ0 der Mittelwert der durch H0 angegebenen Grundgesamtheit, x^- der Stichprobenmittelwert, *S* die Stichprobenstandardabweichung und *n* die Stichprobengröße.

3. Wählen Sie das **Signifikanzniveau** ÿ des Tests. 4. Vergleichen Sie den berechneten Wert *t* mit dem entsprechenden **kritischen Wert** tnÿ1 (dem Wert aus der Student-Tabelle mit nÿ1 Freiheitsgraden). Lehnen Sie H0 ab , wenn der Absolutwert von *t* größer als der kritische Wert ist tnÿ1.

Die kritischen Werte für verschiedene Freiheitsgrade und unterschiedliche Signifikanzniveaus sind in der Student-Tabelle angegeben. Für einen einseitigen Test nehmen wir den Wert *tnÿ1,1ÿÿ* in der Tabelle, und für einen zweiseitigen Test nehmen wir den Wert $tn\ddot{y}1,1\ddot{y}\ddot{y}_{\bar{2}}$.

B. *Student-Test für zwei Stichproben*

1. Die Hypothesen, die wir testen möchten, sind:

– **Nullhypothese:**

H0 : ÿ1 = ÿ2 .

Bei einem Test zum Vergleich zweier Stichproben möchten wir wissen, ob ein signifikanter Unterschied zwischen dem Mittelwert zweier Grundgesamtheiten besteht, aus denen die Stichproben gezogen wurden.

– **Alternativhypothese:** Es kann die folgenden drei Formen annehmen:

H1 : ÿ1 > ÿ2 (einseitiger Test rechts)

H1 : ÿ1 < ÿ2 (einseitiger Test links)

H1 : ÿ1 = ÿ2 (zweiseitiger Test) .

2. Berechnen Sie die t-Statistik mit:

$$t = \frac{x^-1 \ddot{y}\ \bar{x}2}{Sp \cdot \sqrt{\frac{1}{n1} + \frac{1}{n2}}},$$

wobei n1 und n2 die jeweiligen Stichprobengrößen sind und *Sp* die gepoolte Standardabweichung ist:

$$Sp = \sqrt{\frac{(n1 \ddot{y} 1)\ S2_1 + (n2 \ddot{y} 1)\ S2_2}{n1 + n2 \ddot{y} 2}}$$

$$= \sqrt{\frac{\sum_{i=1}^{2} \sum_{j=1}^{ni} (xij \ddot{y} \bar{x}i)^2}{n1 + n2 \ddot{y} 2}},$$

wobei $S2_1$ die Stichprobenstandardabweichung 1, S2 die von Stichprobe 2 und *xij* **Beobachtung** *j* von Stichprobe i darstellt.

3. Wählen Sie das Signifikanzniveau ÿ von Test

4. Vergleichen Sie den berechneten Wert von *t* mit dem kritischen Wert $tn1+n2\ddot{y}2,1\ddot{y}\ddot{y}_{\bar{2}}$ (dem Wert der Student-Tabelle mit *(n1* + n2 ÿ 2) Freiheitsgraden) . Lehnen Sie H0 ab , wenn der Absolutwert von *t* größer als der kritische Wert ist.

Wenn der Test einseitig ist, nehmen Sie den Wert tn1+n2ÿ2,1ÿÿ aus der Student-Tabelle, und wenn er zweiseitig ist, nehmen Sie den Wert $tn1+n2\ddot{y}2,1\ddot{y}\ddot{y}_{\bar{2}}$

Machine Translated by Google

Signifikanztest des

Regressionskoeffizienten Der Student-Test kann auch auf Regressionsprobleme angewendet werden. Betrachten Sie ein einfaches lineares Regressionsmodell, **dessen** Daten aus *n* Beobachtungspaaren *(xi,* yi), *i = 1,...,n,* modelliert von bekommen

$$yi = \ddot{y}0 + \ddot{y}1 \cdot xi + \ddot{y}i ,$$

wobei *ÿi* die zufälligen unabhängigen und normalverteilten Fehler mit Mittelwert 0 und **Varianz** ÿ2 sind.

Die Schätzer von ÿ0 und ÿ1 sind gegeben durch:

$$\ddot{y}\hat{}\,0 = \bar{y} \ddot{y} \ddot{y}\hat{}\,1 \cdot \bar{x} ,$$

$$\ddot{y}\hat{}\,1 = \frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})(yi \ddot{y} \bar{y})}{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2} .$$

Der Schätzerÿˆ 1 ist normalverteilt mit Mittelwert ÿ1 und Varianz

$$\mathrm{Var}(\ddot{y}\hat{}\,1) = \frac{\ddot{y}2}{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2} .$$

Wenn ÿ2 durch den Mittelwert der Quadrate der Residuen S2 geschätzt und im Ausdruck durch die Varianz von ÿˆ 1 ersetzt wird, dann ist das

$$T = \frac{\text{Verhältnis } \ddot{y}\hat{}\,1 \ddot{y} \ddot{y}1}{\mathrm{Var}\ \ddot{y}\hat{}\,1}$$

ist nach der Student-Verteilung mit *n* ÿ 2 Freiheitsgraden verteilt.

Bei einem bestimmten Signifikanzniveau ÿ können wir einen Student-Test des Koeffizienten ÿ1 durchführen. Die Hypothesen lauten wie folgt:

Nullhypothese H0 : ÿ1 = 0

Alternativhypothese H1 : ÿ1 = 0 .

Wenn der absolute Wert des berechneten Verhältnisses *T* ist größer als der Wert von *t* in der Student-Tabelle *tnÿ2,1ÿÿ* , sodass der $\frac{}{2}$, dann können wir schließen Koeffizient ÿ1 signifikant von Null verschieden ist. Im gegenteiligen Fall können wir die Nullhypothese (ÿ1 = 0) nicht ablehnen.

Wir kommen in diesem Fall zu dem Schluss, dass die Steigung der Regressionsgeraden nicht signifikant von Null abweicht.

Ein ähnlicher Test kann für die Konstante durchgeführt werden ÿ0:

Der Schätzerÿˆ 0 ist normalverteilt mit Mittelwert ÿ0 und Varianz

$$\mathrm{Var}(\ddot{y}\hat{}\,0) = \ddot{y}2 \cdot \left(\frac{\ddot{y}}{\sum_{i=1}^{N} x2} \right) \ddot{y} .$$

Wenn ÿ2 durch den Mittelwert der Quadrate der Residuen S2 geschätzt wird, wird er durch den Ausdruck ersetzt Var ÿˆ 0 , dann das Verhältnis

$$T = \frac{\ddot{y}\hat{}\,0 \ddot{y} \ddot{y}0}{\mathrm{Var}\ \ddot{y}\hat{}\,0}$$

ist nach der Student-Verteilung mit *n* ÿ 2 Freiheitsgraden verteilt.

Der Student-Test betrifft die folgenden Hypothesen:

Nullhypothese H0 : ÿ0 = 0

Alternativhypothese H1 : ÿ0 = 0 .

Wenn der absolute Wert des berechneten Verhältnisses *T* ist größer als der Wert von *t* in der Student-Tabelle *tnÿ2,1ÿÿ* , sodass der $\frac{}{2}$, dann können wir schließen Koeffizient ÿ0 signifikant von Null verschieden ist. Im gegenteiligen Fall können wir die Nullhypothese (ÿ0 = 0) nicht ablehnen.

Wir kommen zu dem Schluss, dass die Regression nicht signifikant von Null abweicht.

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Im Allgemeinen sollten wir sehr aufmerksam sein
die Wahl des durchzuführenden Tests. Der Student-
Test wird oft falsch angewendet. Tatsächlich ist es
bedeutet nur etwas, wenn die Beobachtungen
Sie stammen aus einer normalverteilten **Population**
oder sind einer normalverteilten Population nahe
eins. Ist diese Voraussetzung nicht erfüllt, erfolgt die
Der kritische Wert des Student-Tests ist nicht gegeben
bieten eine absolute Garantie.
Wenn die Normalverteilung anwendbar ist, ist die
Ein Schülertest an einer **Probe** wird verwendet, wenn die
Standardabweichung der Grundgesamtheit ist
unbekannt und wann die Dimension der
Die Stichprobe ist nicht groß *(n* < 30).
Ebenso gilt, wenn zwei Proben vorhanden sind
Der Studententest wird verwendet, wenn das Standardgerät verwendet wird
ationen von zwei Populationen sind unbekannt und
angenommen wird, dass sie gleich ist und wenn die
Summe der Dimensionen der Stichproben relativ klein ist *(n1* +
n2 < 30).

BEISPIELE

Schülertest an einer Probe

Betrachten Sie eine **Stichprobe** mit der Dimension *n* = 10
wobei die Beobachtungen normalerweise verteilt
sind und wie folgt lauten:

47 51 48 49 48 52 47 49 46 47 .

Wir wollen folgende Hypothesen testen:

Nullhypothese H0 : ÿ = 50

Alternativhypothese H1 : ÿ = 50 .

Das **arithmetische Mittel** der Beobachtungen
entspricht:

$$x^{-} = \frac{\sum_{i=1}^{10} xi}{N} = \frac{484}{10} = 48{,}4 .$$

Wir fassen die Berechnung im Folgenden zusammen
Tabelle:

xi	$xi - x^{-2}$
47	1,96
51	6,76
48	0,16
49	0,36
48	0,16
52	12.96
47	1,96
49	0,36
46	5,76
47	1,96
Gesamt 32,4	

Die **Standardabweichung** beträgt:

$$S = \sqrt{\frac{\sum_{i=1}^{10} (xi - x^{-})^2}{n - 1}} = \sqrt{\frac{32.4}{9}} = 1{,}90 .$$

Wir berechnen das Verhältnis:

$$t = \frac{x^{-} - \ddot{y}}{\frac{S}{\sqrt{N}}} = \frac{48{,}4 - 50}{\frac{1{,}90}{\sqrt{10}}} = -2{,}67 .$$

Die Anzahl der mit dem Test verbundenen
Freiheitsgrade beträgt *n* ÿ 1 = 9.
Wenn wir ein **Signifikanzniveau** ÿ von 5 % wählen,
dann der **Wert** *tnÿ1,1ÿÿ* $_{\overline{2}}$ des **Studententisches**
entspricht:

$$t_{9;0{,}975} = 2{,}26 .$$

Als |*t*| = 2,67 > 2,26, wir lehnen die Null ab
Hypothese und kommen zu dem Schluss, dass die
Stichprobe kein Signifikanzniveau von 5 % erreicht
aus einer Population mit einem Mittelwert ÿ = 50.

Schülertest an zwei Proben

Es wird eine Studie durchgeführt, um den Mittelwert zu vergleichen
Lebensdauer von zwei Reifenmarken. Eine Probe war
für jede Marke genommen. Die Ergebnisse sind:

- Für Marke 1:

  Mittlere Lebensdauer: $x^{-}1 = 72000$ km

  Standardabweichung: S1 = 3200 km

  Probendimension: n1 = 50

S

Machine Translated by Google

• Für Marke 2:

Mittlere Lebensdauer: $\bar{x}_2 = 74400$ km

Standardabweichung: $S2 = 2400$ km

Stichprobengröße: $n2 = 40$

Wir möchten wissen, ob es einen signifikanten Unterschied in der Lebensdauer der beiden Marken gibt von Reifen, bei einem Signifikanzniveau ÿ von 1 %. Dies impliziert die folgenden Hypothesen:

Nullhypothese H0 : $\ddot{y}1 =$ $\ddot{y}2$ oder $\ddot{y}1 \ddot{y} \ddot{y}2 = 0$

AlternativhypotheseH1 :

$\ddot{y}1 = \ddot{y}2$ oder $\ddot{y}1 \ddot{y} \ddot{y}1 = 0$.

Die gewichtete Standardabweichung *Sp* beträgt:

$$Sp = \frac{\sqrt{(n1 \ddot{y} 1) \cdot S2 +_1 (n2 \ddot{y} 1) \cdot S2^{\,2}}}{n1 + n2 \ddot{y} 2}$$

$$= \frac{\sqrt{49 \cdot 32002 + 39 \cdot 24002\ 50 +}}{40 \ddot{y} 2}$$

$$= 2873{,}07 .$$

Wir können das Verhältnis t berechnen:

$$t = \frac{\bar{x}_1 \ddot{y} \bar{x}_2}{Sp \cdot \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

$$= \frac{72000 \ddot{y} 74400}{2873.07 \cdot \sqrt{\frac{1}{40} + \frac{1}{50}}}$$

$$= \ddot{y}3{,}938 .$$

Die Anzahl der mit dem Test verbundenen Freiheitsgrade beträgt:

$$v = n1 + n2 \ddot{y} 2 = 88 .$$

Der Wert von $t_{v,1\ddot{y}\frac{\ddot{y}}{2}}$ gefunden im Student table für ÿ = 1 % ist:

$$t_{88,0{,}995} = 2{,}634 .$$

Als absoluter Wert des berechneten Verhältnisses ist größer als dieser Wert, $|t| = |\ddot{y}3{,}938| >$ 2.634 lehnen wir die Nullhypothese für die Alternativhypothese ab und kommen zu dem Schluss, dass die beiden Reifenmarken beim Signifikanzniveau 1 % nicht die gleiche mittlere Lebensdauer haben.

WEITERLESEN

Hypothesentest
Parametrischer Test
Studentenverteilung
Studententisch

VERWEISE

Fisher, RA: Anwendungen der „Student“-Verteilung. Metron 5, 90–104 (1925)

Gosset, SW „Student“: Der wahrscheinliche Fehler eines Mittelwerts. Biometrie 6, 1–25 (1908)

Gosset, SW „Student“: Neue Tabellen zur Prüfung der Aussagekraft von Beobachtungen. Metron 5, 105–120 (1925)

Sicheres Ereignis

Siehe **Beispielraum.**

Umfrage

Eine Untersuchung durch Umfrage oder einfach Umfrage ist eine Untersuchung, die an einem begrenzten Teil einer **Bevölkerung durchgeführt wird.** Dieser Teil der Grundgesamtheit stellt die **Stichprobe dar,** und die Methoden, mit denen man diese Stichprobe erstellen kann, werden als **Stichprobenmethoden bezeichnet.** Im Rahmen der Befragung erhalten wir Informationen über die untersuchten Einheiten. Mithilfe der **Inferenzstatistik** verallgemeinern wir diese Informationen dann auf die gesamte Bevölkerung. Diese Verallgemeinerung führt zu einem gewissen **Fehler.** Die Bedeutung dieses Fehlers hängt von der **Stichprobengröße und davon ab,** wie sie im Stichprobenschema ausgewählt wurden

Machine Translated by Google

der zur Entnahme der Stichprobe verwendet wurde,
und der verwendete **Schätzer .** Je repräsentativer
Je größer die Stichprobe, desto geringer ist der Fehler. Probenahme
Methoden sind zum Zeichnen sehr nützlich
Proben, die zu einem minimalen Fehler führen.
Eine andere Technik besteht darin, alles zu beobachten
Individuen einer Population. Wir verweisen hier
zu einer **Volkszählung.** Umfragen haben Kosten und Geschwindigkeit
Vorteile gegenüber Volkszählungen, die
Stichprobenfehler weitgehend ausgleichen. Darüber hinaus ist es so
manchmal ist es unmöglich, eine Volkszählung durchzuführen,
zum Beispiel, wenn die Tatsache der Beobachtung einer Einheit
einer Bevölkerung bedeutet Zerstörung oder wann
die Bevölkerung ist unendlich.

WEITERLESEN

Volkszählung

Datensammlung

Inferenzstatistik

Panel

Bevölkerung

Probe

Probenahme

VERWEISE

Gourieroux, C.: Théorie des sondages. Öko
Nomica, Paris (1981)

Grosbras, JM: Statistische Methoden des
Sondagen. Economica, Paris (1987)

Systematische Probenahme

Bei der systematischen Stichprobe handelt es sich um eine Zufallsmethode
Probenahme. In festgelegten Zeitabständen werden
Individuen aus einer **Population** entnommen ,
Raum oder Reihenfolge des Auftretens, wobei das
erste Individuum zufällig gezogen wird.

GESCHICHTE

Siehe **Probenahme.**

MATHEMATISCHE ASPEKTE

Betrachten Sie eine **Population** bestehend aus N
Individuen. Es wird eine systematische **Stichprobe** erstellt
von Individuen, deren Anzahl eine darstellt
arithmetische Folge.
Eine erste Zahlb wird zufällig ausgewählt
1 und r, wobei $r = n \frac{N}{N}$ die Größe der **Stichprobe** und
eine gemeinsame Differenz zwischen
aufeinanderfolgenden Formen der arithmetischen Folge ist.
Einzelpersonen, die auf diese Weise Proben nehmen, werden dies haben
folgende Nummern:

$$b, b + r, b + 2r, \ldots, b + (n \ddot{y} 1)r .$$

DOMÄNEN UND EINSCHRÄNKUNGEN

Eine systematische **Probenahme** ist nur möglich, wenn die
Einzelpersonen werden in einer bestimmten Reihenfolge klassifiziert.
Der Hauptvorteil dieser Methode liegt darin
Einfachheit und daher niedrige Kosten.
Andererseits hat es den Nachteil
ohne Berücksichtigung einer eventuellen Periodizität des
untersuchten Merkmals. Schwerwiegende **Fehler** könnten sein
Dies kann insbesondere dann der Fall sein, wenn der Zeitraum überschritten wird
ein Teiler des Verhältnisses der Arithmetik
Fortschreiten.

BEISPIELE

Stellen Sie sich eine **Population** von 600 Studenten vor
in alphabetischer Reihenfolge gruppiert
Wir wollen eine **Stichprobe** von 30 Personen nehmen
nach systematischer **Stichprobe.**
Das Verhältnis der arithmetischen Folge ist
gleich:

$$r = \frac{N}{N} = \frac{600}{30} = 20 .$$

Eine erste Zahl wird zufällig ausgewählt
1 und 20, zum Beispiel $b = 17$.
Die **Stichprobe** besteht aus den Studierenden
mit den Zahlen:

$$17, 37, 57, \ldots, 597 .$$

S

Machine Translated by Google

WEITERLESEN

Cluster-Sampling

Probenahme

VERWEISE

Deming, WE: Beispieldesign in der Wirtschaftsforschung. Wiley, New York (1960)

Machine Translated by Google

Zielbevölkerung

Das Konzept der Zielgruppe ist definiert als eine **Bevölkerung ,** zu der wir würden die Ergebnisse einer Untersuchung gerne anwenden.

BEISPIELE

Wir wollen eine Ausländeruntersuchung durchführen Leben in StadtX. **Die Bevölkerung** wird durch die Liste der im Land gemeldeten Ausländer repräsentiert Stadt.
Die Zielbevölkerung wird durch die Gesamtheit der in der Stadt

WEITERLESEN

Bevölkerung

Test der Unabhängigkeit

Ein Unabhängigkeitstest ist ein **Hypothesentest** Dabei besteht das Ziel darin, zu bestimmen, ob zwei Zufallsvariablen unabhängig sind oder nicht.

GESCHICHTE

Siehe **Kendall-Rangkorrelationskoeffizient, Spearman-Rangkorrelationskoeffizient,** und **Chi-Quadrat-Test der Unabhängigkeit.**

MATHEMATISCHE ASPEKTE

Seien (X1, *X2,...,Xn)* und (Y1, *Y2,...,Yn)* . zwei Proben der Dimension n. Das Ziel besteht darin, die folgenden zwei Hypothesen zu testen:

Nullhypothese H0: Die beiden **Variablen** *X* und *Y* sind unabhängig.

Alternativhypothese H1: Die beiden Variablen *X* und *Y* sind nicht unabhängig.

Der Unabhängigkeitstest besteht im Vergleich der empirischen Verteilung mit der theoretischen Verteilung durch Berechnung einer Indica tor.

Der auf zwei kontinuierliche Variablen angewendete Unabhängigkeitstest basiert auf den Rängen der Beobachtungen. Dies ist der Fall, wenn die Tests durchgeführt werden basierend auf der **Spearman-Rangkorrelation Koeffizient** oder auf dem **Kendall-Rangkorre Lationskoeffizient.**

Im Fall von zwei kategorialen Variablen ist die Der am weitesten verbreitete Test ist der **Chi-Quadrat-Test der Unabhängigkeit.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Machen Sie bei jedem Paar einen Unabhängigkeitstest *(Xi,* Yi), *i = 1, 2,...,n* muss aus dem stammen gleiche bivariate **Population.**

BEISPIELE

Siehe **Chi-Quadrat-Test der Unabhängigkeit.**

Machine Translated by Google

WEITERLESEN

Chi-Quadrat-Test der Unabhängigkeit

Hypothesentest

Kendall-Rangkorrelationskoeffizient

Nichtparametrischer Test

Rangkorrelationskoeffizient nach Spearman

Zeitfolgen

Eine Zeitreihe ist eine Folge von Beobachtungen zu aufeinanderfolgenden Zeitpunkten gemessen. Zeitreihen sind manchmal monatlich, vierteljährlich oder jährlich wöchentlich, täglich oder stündlich (Untersuchung des Straßenverkehrs, Telefonverkehr) oder alle zwei oder zehn Jahre stattfinden. Die Zeitreihenanalyse besteht aus Methodendas Versuchen Sie, solche Zeitreihen zu verstehen Voraussagen machen.

Zeitreihen können in vier Teile zerlegt werden Komponenten, von denen jede etwas Besonderes ausdrückt Aspekt der Bewegung der Werte der Zeitfolgen.

Diese vier Komponenten sind:

- **Säkularer Trend,** der den Umzug beschreibt ment entlang der Laufzeit;
- **Saisonale Schwankungen,** die darstellen saisonale Veränderungen;
- **Zyklische Schwankungen,** die korrespondieren auf periodische, aber nicht saisonale Schwankungen;
- **Unregelmäßige Variationen,** die anders sind äh nichtzufällige Variationsquellen von Serie.

Die Analyse von Zeitreihen besteht darin, diese Elemente mathematisch zu beschreiben, also die vier getrennt zu schätzen Komponenten.

GESCHICHTE

In einem Artikel von Funkhauser aus dem Jahr 1936 reproduzierte HG ein Diagramm aus dem 10. Jahrhundert zeigt die Neigung der Umlaufbahn von Sev de Planeten im Zeitverlauf. Übereinstimmung Laut Kendall ist dies das älteste in der westlichen Welt bekannte Zeitdiagramm.

Es reicht aus, sich hierfür an die Astronomie zu wenden Ursprünge der Zeitreihe. Die **Beobachtung** der Sterne war in der Antike beliebt.

Im 16. Jahrhundert kam es wieder zur Astronomie eine Quelle wichtiger Entdeckungen. Die Werke von Brahe, Tycho (1546–1601), dessen Sammlung von Daten über die Bewegung der Planeten es Kepler, Johannes (1571–1630) ermöglichte seine Gesetze zur Planetenbewegung zu formulieren.

Eine statistische Analyse von Daten, so empirisch wie es könnte damals gewesen sein, lag dem zugrunde Die Arbeit eines großartigen Wissenschaftlers.

Die Entwicklung der Mathematik im 18 und das 19. Jahrhundert ermöglichte eine Umgehung grafische Visualisierung, um zum ersten zu gelangen Techniken der Zeitreihenanalyse. Verwendung der Frequenzansatz haben Wissenschaftler entwickelt funktioniert zunächst in diesem Bereich.

Frequenzanalyse (auch Harmonische genannt). (Analyse) war ursprünglich dazu gedacht, eine oder mehrere zyklische „Komponenten“ hervorzuheben eine Zeitreihe.

Im Jahr 1919 schlug WM in Persons eine Zerlegung der Zeitreihen in Bezug auf Tendenzen **(säkulare Trends),** zyklische **zyklische Schwankungen),** saisonale **(saisonale Schwankungen) und** und zufällige **(unregelmäßige Variation)** Komponenten.

Viele Werke wurden ihm gewidmet die Feststellung und Beseitigung eines bzw eine weitere dieser Komponenten.

Die Bestimmung von Tendenzen hat oft stattgefunden mit Problemen der Regressionsanalyse in Zusammenhang gebracht wurden, entwickelten auch Forscher diesbezüglich nichtlineare Tendenzen, wie zum Beispiel die Gompertz-Kurve (gegründet 1825 von Gompertz, Benjamin) oder die Logistikkurve (1844 von Verhulst, Pierre eingeführt François). Ein weiterer Ansatz zur Bestimmung von Tendenzen besteht in der Verwendung von Bewegungen Durchschnittswerte, weil sie es einem ermöglichen, das zu behandeln

Machine Translated by Google

Problem von Tendenzen, wenn sie dazu dienen, unregelmäßige oder periodische Schwankungen zu beseitigen. Die Verwendung gleitender Durchschnitte wurde in zwei Richtungen entwickelt: für versicherungsmathematische Probleme und für die Untersuchung von Zeitreihen in der Wirtschaft. Die Verwendung gleitender Durchschnitte wird insbesondere im Rahmen der Methoden der Saisonbereinigungsmethoden entwickelt. Genauer gesagt, die Trennung saisonaler Variablen voneinander

Komponenten der Zeitreihe wurden ursprünglich 1915 von Copeland, MT und 1919 von Persons, WM untersucht und 1930 von Macauley, FR fortgesetzt.

Macauley, FR verwendete die heute noch weit verbreitete Methode der Verhältnisse zum gleitenden Durchschnitt (siehe saisonale Variation), siehe auch Dufour (2006).

MATHEMATISCHE ASPEKTE

Sei *Yt* eine Zeitreihe, die mit Hilfe dieser vier Komponenten zerlegt werden kann
Vorteile:

• **Säkularer Trend** *Tt* •
Saisonale Schwankungen *St* •
Zyklische Schwankungen *Ct* •
Unregelmäßige Schwankungen
It Wir nehmen an, dass die von der **Zufallsvariablen** *Yt* angenommenen Werte durch eine **Beziehung** zwischen den vier vorherigen Komponenten bestimmt werden .
Wir unterscheiden drei Zusammensetzungsmodelle: 1. Additives Modell:

$$Yt = Tt + Ct + St + It .$$

Es wird davon ausgegangen, dass die vier Komponenten unabhängig voneinander sind.

2. Multiplikatives Modell:

$$Yt = Tt \cdot Ct \cdot St \cdot It .$$

Mit Hilfe von Logarithmen gelangen wir vom multiplikativen Modell zum additiven Modell.

3. Gemischtes Modell:

$$Yt = St + (Tt \cdot Ct \cdot It)$$

oder

$$Yt = Ct + (Tt \cdot St \cdot \mathrm{It}).$$

Diese Art von Modell wird sehr wenig verwendet. Saisonal wählen wir das additive Modell
Variationen sind nahezu konstant und ihr Einfluss
Die Wirkung auf die Tendenz hängt nicht von ihrem Niveau ab.

Wenn die saisonalen Schwankungen eine Amplitude haben, die nahezu proportional zu der der säkularen Trends ist, führen wir das multiplikative Modell ein; Dies ist im Allgemeinen der Fall.
Die Analyse von Zeitreihen besteht darin, die von jeder Komponente angenommenen Werte zu bestimmen.
Wir beginnen immer mit den saisonalen Schwankungen und enden mit den zyklischen Schwankungen. Alle Schwankungen, die nicht auf einen der Faktoren zurückzuführen sind
Drei Komponenten werden mit den unregelmäßigen Variationen gruppiert.
Basierend auf dem verwendeten Modell kann die Analyse, nachdem die Komponenten bestimmt wurden, nachfolgende Daten durch Subtraktion oder Division anpassen.
So, wenn wir das Säkulare geschätzt haben
Trend *Tt,* zu jedem Zeitpunkt t wird die Zeitreihe wie folgt angepasst:
– Wenn wir das additive **Modell** verwenden *Yt* = *Tt* + *Ct* + *St* + *It:*

$$Yt \ddot{y} Tt = Ct + St + It .$$

– Für das multiplikative Modell *Yt* = *Tt* ·*Ct* · *St* · *Es:*

$$\frac{Yt}{Tt} = Ct \cdot St \cdot It .$$

Indem wir nacheinander jede der drei Komponenten auswerten und die Zeitreihen zwischen ihnen anpassen, erhalten wir die Werte, die den unregelmäßigen Variationen zugeordnet werden.

T

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Wenn die Zeitreihen nicht über einen ausreichend langen Zeitraum verfügen, gruppieren wir den **säkularen Trend** und die zyklischen Schwankungen in einer Komponente, die als **außersaisonale Bewegung bezeichnet wird.**

Grafische Darstellung Die grafische Darstellung von Zeitreihen hilft dabei, beobachtete Werte in Abhängigkeit von der Zeit zu analysieren.

Um diese Werte zu analysieren, müssen bestimmte Vorsichtsmaßnahmen getroffen

werden: 1. Jeder Punkt muss für eine Beobachtung repräsentativ sein. Wenn der Wert eine Aktie an einem festen Datum darstellt, wird der Punkt am Referenzdatum gefunden. Wenn da ist

B. monatlich, dann wird der Zeitpunkt auf die Monatsmitte festgelegt.

2. Manchmal müssen in derselben Grafik Jahresdaten und Monatsdaten gemeldet werden. Die repräsentativen Punkte der Ergebnisse sollten 12-mal häufiger sein als diejenigen relativ zu den Jahresergebnissen. Also müssen wir

Achten Sie auf die repräsentativen Punkte an der richtigen Stelle.

3. Wenn viele Reihen im selben Diagramm dargestellt werden, müssen sie ihrer Natur nach identisch sein. Andernfalls sollten wir unterschiedliche Maßstäbe verwenden oder Diagramme untereinander platzieren.

Die Hauptziele der Untersuchung der Zeitreihen sind:

- Beschreibung: Bestimmung oder Eliminierung verschiedener Komponenten • Stochastische Modellierung
- **Prognose** • Filterung:

Eliminierung oder

Umrechnung von

bestimmte Eigenschaften

- Kontrolle: **Beobachtung** der Vergangenheit eines Prozesses, um auf zukünftige Entwicklungen zu reagieren

Für die Prognose liegen der Analyse von Zeitreihen folgende Hypothesen zugrunde:

- Die große historisch beobachtete Tendenz

städte werden in der mehr oder weniger nahen Umgebung beibehalten

Zukunft.

- Die messbaren Schwankungen einer **Variablen**

werden in regelmäßigen Abständen reproduziert.

Die Analyse von Zeitreihen ist dann gerechtfertigt, wenn dadurch die Unsicherheit bei der Erstellung von Prognosen verringert werden kann.

Wir unterscheiden: 1.

Langfristige Prognose: Basierend

auf dem säkularen Trend; zyklische Schwankungen werden grundsätzlich nicht berücksichtigt.

2. *Mittelfristige Prognose:* Um den

wahrscheinlichen Effekt der zyklischen Komponente zu berücksichtigen, sollten wir den prognostizierten Wert des säkularen Trends mit einer **Schätzung** der relativen diskontierten Variation multiplizieren, die auf die zyklischen Schwankungen zurückzuführen ist.

3. *Kurzfristige Prognose:* Im

Allgemeinen versuchen wir nicht, unregelmäßige Schwankungen vorherzusagen oder kurzfristige Prognosen zu erstellen

Prognosen über längere Zeiträume. Daher multiplizieren wir die prognostizierten Werte des säkularen Trends mit der Schätzung der zyklischen Schwankungen und mit der saisonalen Variation des entsprechenden Datums.

Bei einer Analyse von Zeitreihen stehen wir vor folgenden Problemen: 1. *Ist das klassische Modell geeignet?*

Wir gehen davon aus, dass die Qualität der Analyseergebnisse proportional zur Präzision und Genauigkeit des **Modells** selbst ist.

Das multiplikative Modell Yt = Tt·St·Ct·Es setzt voraus, dass die vier Komponenten unabhängig sind, was in der Realität selten zutrifft.

2. *Sind unsere Hypothesen der Konstanz und Regelmäßigkeit gültig?*

Die Analyse muss an subjektive und qualitative Faktoren angepasst werden, die in fast jeder Situation vorhanden sind.

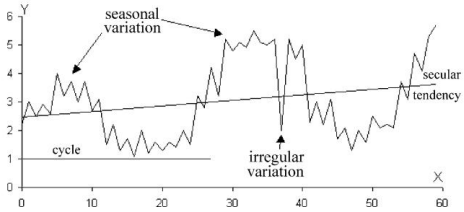
Machine Translated by Google

3. *Können wir den verfügbaren Daten vertrauen?*

Das Problem kann durch Reihenbrüche entstehen, die auf unzureichende Daten oder eine Änderung der Variablenqualität zurückzuführen sind.

BEISPIELE

Eine Zeitreihe wird im Allgemeinen wie folgt dargestellt:



Wir unterscheiden vier Komponenten: • **Säkularer Trend,** leicht steigend im vorliegender Fall • Jahreszeitliche Schwankungen, leicht erkennbar • Zyklische Schwankungen in Form von Zyklen mit einer ungefähren Amplitude von 27 Zeiteinheiten

• Unregelmäßige Schwankungen, im Allgemeinen schwach genug, mit Ausnahme von $t = 38$, was einen sehr abrupten Abfall darstellt, der nicht rational wäre

WEITERLESEN

Zyklische Schwankung
Prognose
Grafische Darstellung
Unregelmäßige Variation
Saisonale Unterschiede
Säkularen Trend

VERWEISE

Copeland, MT: Statistische Indizes der Geschäftsbedingungen. QJ Econ. **29,** 522–562 (1915)

Dufour, IM: Histoire de l'analyse de Séries Chronologues, Université de Montréal, Kanada (2006)

Funkhauser, HG: Eine Anmerkung zu einer Grafik aus dem 10. Jahrhundert. Osiris 1, 260–262 (1936)

Kendall, MG: Zeitreihen. Griffin, London (1973)

Macauley, FR: Die Glättung von Zeitreihen. Nationales Wirtschaftsbüro Forschung, 121–136 (1930)

Personen,WM: Verzeichnis der Allgemeinen Geschäftsbedingungen. Rev. Econom. Stat. 1, 111–205 (1919)

Verfolgen

Für eine quadratische Matrix *A* der Ordnung n definieren wir die Spur von *A* als die Summe der situierten Terme auf der Diagonale. Somit ist die Spur einer Matrix eine skalare Größe.

MATHEMATISCHE ASPEKTE

Sei *A* eine quadratische **Matrix** der Ordnung n, $A =$ wobei $i, j = 1, 2,...,n$. Wir definieren aij, verfolgen *A* durch:

$$\mathrm{tr}(A) = \sum_{i=1}^{N} aii\ .$$

BEREICHE UND EINSCHRÄNKUNGEN

Da wir nur an Diagonalelementen interessiert sind, ist die Spur nur für quadratische Matrizen definiert.

1. Die Spur der Summe zweier Matrizen entspricht der Summe der Spuren der Matri ces:

$$\mathrm{tr}(A + B) = \mathrm{tr}(A) + tr(B).$$

2. Die Spur des Produkts zweier Matrizen ändert sich bei der Kommutierung der Matrizen nicht. Mathematisch: $\mathrm{tr}(A \cdot B) = \mathrm{tr}(B \cdot A)$.

Machine Translated by Google

BEISPIELE

Seien A und B zwei Quadratmatrizen der Ordnung 2:

$$A = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \quad \text{und } B = \begin{matrix} 5 & 2 \\ \ddot{y}2 & 3 \end{matrix} .$$

Die Berechnung der Spuren ergibt:

$$\text{tr}(A) = \text{tr} \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} = 1 + 3 = 4 ,$$

$$\text{tr}(B) = \text{tr} \begin{matrix} 5 & 2 \\ \ddot{y}2 & 3 \end{matrix} = 5 + 3 = 8 ,$$

Und

$$\text{tr}(A + B) = \text{tr} \begin{matrix} 6 & 4 \\ \ddot{y}1 & 6 \end{matrix}$$

$$= 6 + 6 = 12 .$$

Wir überprüfen, dass tr(A) + tr(B) = 12. Auf der andere Hand:

$$A \cdot B = \begin{matrix} 1 & 8 \\ \ddot{y}1 & 11 \end{matrix} \quad \text{Und}$$

$$B \cdot A = \begin{matrix} 7 & 16 \\ 15 & \end{matrix} ,$$

$$\text{tr}(A \cdot B) = \text{tr} \begin{matrix} 1 & 8 \\ \ddot{y}1 & 11 \end{matrix}$$

$$= 1 + 11 = 12 ,$$

$$\text{tr}(B \cdot A) = \text{tr} \begin{matrix} 7 & 16 \\ 15 & \end{matrix}$$

$$= 7 + 5 = 12 .$$

Somit ist tr(A · B) = tr(B · A) = 12.

WEITERLESEN

Matrix

Transformation

Eine Transformation ist eine Veränderung in einem oder mehreren Variablen in einer statistischen Studie.

Wir transformieren Variablen, zum Beispiel indem man sie durch ihre Logarithmen ersetzt **(logarithmische Transformation).**

DOMÄNEN UND EINSCHRÄNKUNGEN

Wir transformieren Variablen aus vielen Gründen:

1. Wenn wir einen Datensatz haben relativ zu vielen Variablen und wir wollen Drücken Sie eine davon, *Y* **(abhängige Variable),** mit Hilfe anderer (sogenannte unabhängige Variablen) aus und wir wollen die **Beziehung** zwischen den Variablen linearisieren. Zum Beispiel, wenn die Beziehung zwischen unabhängige Variablen sind multiplikativ, eine **logarithmische Transformation** der Daten machen es additiv und ermöglichen die Nutzung der linearen Regression zur Schätzung der Parameter des **Modells.**
2. Wenn wir uns in der Gegenwart von Daten befinden relativ zu einer Variablen und ihrer **Varianz** eingestellt nicht konstant ist, ist dies im Allgemeinen möglich Machen Sie es durch Transformation nahezu konstant diese Variable.
3. Wenn eine **Zufallsvariable** einer folgt Verteilung, eine angemessene Transformation davon ermöglicht es, eine neue Zufallsvariable zu erhalten in der Lage, dass ungefähr **anormal** folgt **Verteilung.**

Die Vorteile der Transformation eines Datensatzes relativ zu einer Variablen lassen sich im Großen und Ganzen wie folgt zusammenfassen: folgt:

- Es vereinfacht bestehende Beziehungen mit anderen Variablen.
- Ein Mittel gegen Ausreißer, Linearität und Homoskedastizität.
- Es wird eine Variable erzeugt, die annähernd normalverteilt ist.

WEITERLESEN

Abhängige Variable
Unabhängige Variable
Logarithmische Transformation
Normalverteilung
Regressionsanalyse
Varianz

Machine Translated by Google

VERWEISE

Cox, DR, Box, GEP: Eine Analyse von Transformationen (mit Diskussion). J. Roy. Stat. Soc. Ser. B **26,** 211–243 (1964)

Transponieren

Die (Matrix-)Transponierte einer **MatrixA der** Ordnung *(m* × n) ist eine Matrix *(n* × m), bezeichnet mit *A* , die , durch Schreiben der Zeilen von A in Spalten von *A* und umgekehrt erhalten wird.

Somit bewirken wir eine einfache „Umkehrung“ der Tabelle.

MATHEMATISCHE ASPEKTE

Wenn $A = aij$, $i = 1, 2,...,m$ und $j = 1, 2,...,n$ eine Ordnungsmatrix (m × n) ist , dann ist die Transponierte von A *die* Ordnungsmatrix *A (n* × m) gegeben durch:

$$A = aji \text{ mit } j = 1, 2,...,n \text{ und } i = 1, 2,...,m.$$

Die Transposition ist ein reversibler und reziproker Vorgang; Wenn wir also die Transponierte der Transponierten einer Matrix nehmen, finden wir die Anfangsmatrix.

Domänen und Einschränkungen

Wir verwenden die Transponierte einer **Matrix,** oder genauer gesagt die Transponierte eines Vektors, während wir das **Skalarprodukt berechnen.**

BEISPIELE

Sei *A* die folgende **Ordnungsmatrix** (3 × 2):

$$A = \begin{matrix} 1 & 2 \\ 0 & 3 \\ 2 & 5 \end{matrix} .$$

Die Transponierte von *A* ist die Matrix (2 × 3):

$$A = \begin{matrix} 1 & 0 & 2 \\ 2 & 3 & 5 \end{matrix} .$$

Wir können auch überprüfen, dass die Transponierte von *A* gleich der Anfangsmatrix A ist:

$$A = A .$$

WEITERLESEN

Matrix

Vektor

Behandlung

In einem **experimentellen Design** ist eine Behandlung eine bestimmte Kombination verschiedener Faktoren.

BEISPIELE

Oft werden Experimente durchgeführt, um zwei oder mehr Behandlungen zu vergleichen, zum Beispiel zwei verschiedene Düngemittel für eine bestimmte Pflanzenart oder viele Arten von Arzneimitteln zur Behandlung einer bestimmten Krankheit. Ein weiteres Beispiel besteht in der Messung der Gerinnungszeit von Blutproben von 16 Tieren, die unterschiedliche Regime A, B, C und D durchlaufen haben.

In diesem Fall sind unterschiedliche Beispiele für Behandlungen Düngemittel, Medikamente und Kuren.

Andererseits können wir bei Experimenten, bei denen ein bestimmter Dünger, beispielsweise Azote, an einer Weizenernte getestet wird, unterschiedliche Mengen desselben Düngers als unterschiedliche Behandlungen betrachten. Hier haben wir einen **Faktor** (Azote) auf verschiedenen Ebenen (Mengen), zum Beispiel 30 kg, 50 kg, 100 kg und 200 Jede dieser Ebenen entspricht einem Leckerbissen ment.

T

Machine Translated by Google

Das Ziel eines Experiments besteht darin, festzustellen, ob es einen echten (signifikanten) Unterschied zwischen diesen Behandlungen gibt.

WEITERLESEN

Versuchsplanung
Experiment
Versuchseinheit
Faktor

VERWEISE

Senn, St.: Cross-over-Studien in der Klinik Forschung. Wiley (1993)

Tukey, John Wilder

Tukey, John Wilder wurde 1915 in Bedford, MA, geboren. Er studierte Chemie an der Brown University und promovierte 1939 in Mathematik an der Princeton University. Im Alter von 35 Jahren wurde er Professor für Mathematik an der Princeton University. Er war

1966 war er maßgeblich an der Ausbildung in der Statistikabteilung der Princeton University beteiligt. Von 1966 bis 1969 war er deren Vorsitzender. Er ist Autor von *„Exploratory Data Analysis* “ (1977) (übersetzt ins Russische) und acht Bänden mit Artikeln. Er ist Co-Autor von *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male; Datenanalyse und Regression* (auch ins Russische übersetzt), *Index zu Statistik und Wahrscheinlichkeit.* Er ist Mitherausgeber der folgenden Bücher: *Understanding Robust and Exploratory Data Analysis, Exploring Data* Tables, Trends *Shapes, Configuration Polysampling* und *Fundamentals of Exploratory Analysis of Variance.* Er starb im Juli 2000 in New Jersey.

Ausgewählte Hauptwerke und Artikel von Tukey, John Wilder:

1977 Explorative Datenanalyse. 1. Aufl. Addison-Wesley, Reading, MA.

1980 Wir brauchen sowohl Sondierungen als auch Bestätigungen. Bin. Stat. 34, 23–25.

1986 Datenanalyse und Verhaltenswissenschaft oder lernen, die Last des quantitativen Menschen zu tragen, indem man schlechte Gebote meidet. In: Gesammelte Werke von John W. Tukey, Band III: Philosophie und Prinzipien der Datenanalyse, 1949–1964. (Jones, LV, Hrsg.) Wadsworth Advanced Books & Software, Monterey, CA.

1986 Gesammelte Werke von John W. Tukey, Band III: Philosophie und Prinzipien der Datenanalyse, 1949–1964. (Jones, LV, Hrsg.) Wadsworth Advanced Books & Software, Monterey, CA.

1986 Gesammelte Werke von John W. Tukey, Band IV: Philosophie und Prinzipien der Datenanalyse, 1965–1986. (Jones, LV, Hrsg.) Wadsworth Advanced Books & Software, Monterey, CA.

1988 Gesammelte Werke von John W. Tukey, Band V: Grafiken, 1965–1985. (Cleve Land, WS, Hrsg.) Waldsworth und Brooks/Cole, Pacific Grove, CA.

WEITERLESEN

Box-Plot
Explorative Datenanalyse

Zweiseitiger Test

Ein zweiseitiger Test bezüglich einer **Grundgesamtheit** ist ein **Hypothesentest ,** der angewendet wird, wenn wir eine Schätzung eines Parameters zu einem gegebenen Wert mit der Alternativhypothese vergleichen möchten, die nicht dem angegebenen Wert entspricht.

Machine Translated by Google

MATHEMATISCHE ASPEKTE

Ein zweiseitiger Test ist ein **Hypothesentest** für eine Grundgesamtheit der **folgenden** Art:

Nullhypothese H0 : $\ddot{y} = \ddot{y}0$

Alternativhypothese H1 : $\ddot{y} = \ddot{y}0$,

wobei ÿ ein **Parameter** der Population ist dessen **Wert** unbekannt ist und ÿ0 der angenommene Wert dieses Parameters ist.
Für einen Test zum Vergleich von zwei Populationen lauten die Hypothesen:

Nullhypothese H0 : $\ddot{y}1 = \ddot{y}2$

Alternativhypothese H1 : $\ddot{y}1 = \ddot{y}2$,

wobei ÿ1 und ÿ2 die unbekannten Parameter sind von zwei zugrunde liegenden Populationen.

DOMÄNEN UND EINSCHRÄNKUNGEN

Bei einem zweiseitigen Test beträgt der **Ablehnungsbereich** in zwei Teile geteilt, den linken und den rechten Seiten des betrachteten **Parameters.**

BEISPIELE

Ein Unternehmen produziert Stahlseile. Es will Überprüfen Sie, ob der Durchmesser ÿ eines Großteils des hergestellten Kabels dem Standarddurchmesser entspricht Durchmesser von 0,9 cm.

Es wird eine **Stichprobe** von 100 Kabeln genommen, deren Mittelwert Der Durchmesser beträgt 0,93 cm mit einer **Standardabweichung tion** von $S = 0{,}12$.

Die Hypothesen lauten wie folgt:

Nullhypothese H0 : $\ddot{y} = 0{,}9$

Alternativhypothese H1 : $\ddot{y} = 0{,}9$,

Da die Stichprobengröße groß genug ist, ist die **Stichprobenverteilung** des **Mittelwerts** kann sein durch eine **Normalverteilung** angenähert .
Für ein **Signifikanzniveau** ÿ = 5 % erhalten wir ein **kritischer Wert** zÿ/2 in der **Normaltabelle** von zÿ/2 = 1,96.

Wir finden den **Akzeptanzbereich** der Null Hypothesen (siehe **Ablehnungsbereich):**

$$\ddot{y} \pm z\ddot{y}/2 \cdot \frac{\ddot{y}}{\ddot{y}n} ,$$

wobei ÿ die Standardabweichung des ist Bevölkerung. Da die Standardabweichung ÿ ist unbekannt, wir schätzen es anhand des Standards Abweichung S der Stichprobe $(S = 0{,}12)$, die gibt:

$$0{,}9 \pm 1{,}96 \cdot \frac{0{,}12}{\ddot{y}100} = 0{,}8765 \quad 0{,}9235 .$$

Der Akzeptanzbereich der Nullhypothese entspricht dem **Intervall** [0,8765, 0,9235].
Da der Mittelwert der Stichprobe (0,93 cm) außerhalb dieses Intervalls liegt, müssen wir das ablehnen Nullhypothese für die Alternativhypothese. Wir kommen zu dem Schluss, dass dies auf einem Signifikanzniveau liegt von 5 % und abhängig von der untersuchten Stichprobe, Der Durchmesser der Kabel stimmt nicht überein zur Norm.

Es wird eine Studie durchgeführt, um den Mittelwert zu vergleichen Lebensdauer von zwei Reifentypen. Eine Probe von 50 Reifen der Marke 1 ergeben eine durchschnittliche Lebensdauer von 72.000 km und eine Probe von 40 Reifen der Marke 2 hat eine mittlere Lebensdauer von 74400 km. Angenommen, die Standardabweichung für Marke 1 beträgt ÿ1 = 3200 km und die der Marke 2 beträgt ÿ2 = 2400 km wollen wir wissen, ob es einen signifikanten Unterschied in der Lebensdauer zwischen den beiden gibt zwei Marken, auf einem Signifikanzniveau ÿ = 1 %.

Die Hypothesen lauten wie folgt:

Nullhypothese H0 : $\ddot{y}1 = \ddot{y}2$

Alternativhypothese H1 : $\ddot{y}1 = \ddot{y}2$.

Die Größe der Probe ist groß genug, und die Standardabweichungen der Populationen sind bekannt, so dass wir uns dem nähern können

Machine Translated by Google

Stichprobenverteilung unter Verwendung einer Normalverteilung.

Der Akzeptanzbereich der Nullhypothese ist gegeben durch (siehe Ablehnungsbereich):

$$ÿ1 ÿ ÿ2 ± zÿ/2 · ÿx¯1ÿ¯x2 ,$$

Dabei ist ÿx¯1ÿ¯x2 die Standardabweichung der Stichprobenverteilung der Differenz zweier Mittelwerte oder der **Standardfehler** zweier Mittelwerte.

Der Wert zÿ/2 findet sich in der **Normaltabelle.** Für einen zweiseitigen Test mit ÿ = 1 % erhalten wir zÿ/2 = 2,575, was Folgendes ergibt:

$$0 ± 2,575 · \sqrt{\frac{ÿ2_1}{n1}ÿ2 + \frac{2}{n2}}$$

$$= 0 ± 2,575 · \sqrt{\frac{32002}{50}\frac{24002}{40} +}$$

$$= 0 ± 2,575 · 590,59$$

$$= ±1520,77 .$$

Der Akzeptanzbereich der Nullhypothese ist durch das **Intervall** [ÿ1520,77, *1520,77] gegeben.* Da die Mittelwertdifferenz der Stichproben (72000 ÿ 74400 = ÿ2400) außerhalb des Akzeptanzbereichs liegt, müssen wir die Nullhypothese für die Alternativhypothese ablehnen. Wir können daher den Schluss ziehen, dass es einen erheblichen Unterschied zwischen der Lebensdauer der Reifen gibt

geprüfte Marken.

WEITERLESEN

- **Akzeptanzregion**
- **Alternative Hypothese**
- **Hypothesentest**
- **Nullhypothese**
- **Einseitiger Test**
- **Ablehnungsbereich**
- **Stichprobenverteilung**
- **Signifikanzniveau**

Zwei-Wege-Varianzanalyse

Die zweiseitige Varianzanalyse ist eine Erweiterung der einseitigen Varianzanalyse, bei der es zwei unabhängige Faktoren (Variablen) gibt. Jeder Faktor hat zwei oder mehr Ebenen und Behandlungen werden durch die Erstellung aller möglichen Kombinationen von Ebenen zweier Faktoren erstellt.

GESCHICHTE Es ist **Fisher, RA** (1925), der komplexen Experimenten den Namen **„Fakultätsexperimente“** gab.

Yates, F. (1935, 1937) entwickelte das Konzept und die Analyse dieser faktoriellen Experimente. Siehe **Varianzanalyse.**

MATHEMATISCHE ASPEKTE

Gegeben sei ein **Experiment** mit zwei **Faktoren,** wobei Faktor A a - **Niveaus** und Faktor B *b*- Niveaus hat.

Wenn der mit diesem **faktoriellen Experiment** verbundene Versuchsplan ein **vollständig randomisierter Versuchsplan ist,** sieht das **Modell** wie folgt aus:

$$Yijk = ÿ + ÿi + ÿj + (ÿÿ)ij + ÿijk , i = 1, 2,...,a$$

(Stufen von **Faktor** A) $j = 1, 2,...,b$ **(Stufen** von **Faktor** B) $k = 1, 2,...,c$ (Anzahl der **Beobachtungen** Erhalt der **Behandlung** *ij)*

Dabei ist ÿ der allgemeine **Mittelwert** aller **Behandlungen,** *ÿi* die Auswirkung der **Stufe** *i* von **Faktor** A, *ÿj* die Auswirkung der Stufe *j* von Faktor B, (ÿÿ)ij die Auswirkung der Wechselwirkung zwischen

Machine Translated by Google

ÿi und *ÿj,* und *ÿijk* ist der experimentelle **Fehler** der **Beobachtung** *Yijk.*
Dieses Modell unterliegt **den Grundannahmen** der **Varianzanalyse,** wenn
Man geht davon aus, dass die Fehler *ÿijk* unabhängige Zufallsvariablen sind, **die** einer **Normalverteilung** N(0, ÿ2) folgen.

Anschließend können drei **Hypothesen** getestet werden:

1.

H0 : ÿ1 = ÿ2 = ··· = *ÿa*

H1 : Mindestens ein *ÿi* ist unterschiedlich aus *ÿj, i* = j.

2.

H0 : ÿ1 = ÿ2 = H1 : ··· = *ÿb*

Mindestens ein *ÿi* ist unterschiedlich aus *ÿj, i* = j.

3.

H0 : (ÿÿ)11 = (ÿÿ)12 = ··· = (ÿÿ)1b
= (ÿÿ)21 = ··· = (ÿÿ)ab

H1 : Mindestens einer der Interaktionen sind unterschiedlich von den Anderen.

Zum Testen wird die **Fisher-Verteilung** verwendet erste **Hypothese.** Diese Verteilung erfordert die Schaffung eines Verhältnisses, dessen Zähler ist eine **Schätzung** der Varianz von **Faktor** A und dessen Nenner eine Schätzung der Vari ist

Fehler innerhalb von Behandlungen werden auch als **Fehler** bezeichnet oder **Rest.**

Dieses mit F bezeichnete Verhältnis folgt einem **Fisher Verteilungmitaÿ1** undab(cÿ1)Grad **der Freiheit.**

Nullhypothese H0 : ÿ1 = ÿ2 = ··· = *ÿa*

wird auf dem **signifikanten Niveau** ÿ abgelehnt, wenn Das Verhältnis Fi ist größer oder gleich dem Wert in der **Fisher-Tabelle,** also wenn

$$F \ddot{y} F a\ddot{y}1,ab(c\ddot{y}1),\ddot{y} .$$

Um die zweite **Hypothese zu testen,** erstellen wir ein Verhältnis, dessen Zähler eine **Schätzung** von ist **Varianz** des **Faktors** B und dessen Nennwert Toris ist eine Schätzung der Variation innerhalb der Behandlung gen.

Dieses mit F bezeichnete Verhältnis folgt einem **Fisher Verteilungmitbÿ1** undab(cÿ1)Grad **der Freiheit.**

Nullhypothese H0 : ÿ1 = ÿ2 = ··· = *ÿb*

wird abgelehnt, wenn das F -Verhältnis größer als oder ist gleich dem **Wert** der **Fisher-Tabelle,** d. h. if

$$F \ddot{y} Fb\ddot{y}1,ab(c\ddot{y}1),\ddot{y} .$$

Um die dritte **Hypothese zu testen,** erstellen wir ein Verhältnis dessen Zähler eine **Schätzung** des ist **Varianz** der Wechselwirkung zwischen fac toren A und B und deren Nenner ein ist Schätzung der Varianz innerhalb der Behandlungen.

Dieses mit F bezeichnete Verhältnis folgt einem **Fisher** Verteilungmit(aÿ1)(bÿ1)undab(cÿ1) **Freiheitsgrade.**

Nullhypothese H0 :
(ÿÿ)11 = (ÿÿ)12 = ··· = (ÿÿ)1b
= (ÿÿ)21 = ··· = (ÿÿ)ab

wird abgelehnt, wenn das F -Verhältnis größer als oder ist gleich dem **Wert** in **der Fisher-Tabelle,** d. h. if

$$F \ddot{y} F(a\ddot{y}1)(b\ddot{y}1),ab(c\ddot{y}1),\ddot{y} .$$

Varianz von Faktor A

Für die **Varianz** von **Faktor** A ist die Summe von Quadrate müssen für Faktor A *(SSA) berechnet werden.* was man wie folgt erhält:

$$SSA = b \cdot c \sum_{i=1}^{A} (\bar{Y}i.. \ddot{y} \bar{Y}...)^2 ,$$

wobei $\bar{Y}i..$ **der Mittelwert aller Beobachtungen** ist der **Ebene** *i* des Faktors A und $\bar{Y}...$ ist das Allgemeine Mittelwert aller Beobachtungen.

T

Machine Translated by Google

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** ist gleich *a* ÿ 1.

Die **Varianz** von **Faktor** A ist dann gleich:

$$s_A^2 = \frac{SSA}{a \text{ ÿ } 1}.$$

Varianz von Faktor B

Für die **Varianz** von **Faktor** B ist die Summe von Quadrate müssen für Faktor B *(SSB) berechnet werden.* was man wie folgt erhält:

$$SSB = a \cdot c \sum_{j=1}^{B} (\bar{Y}.j. \text{ ÿ } \bar{Y}...)^2,$$

wobei $\bar{Y}i..$ **der Mittelwert aller Beobachtungen** ist der **Ebene** *i* des Faktors B und $\bar{Y}...$ ist das Allgemeine Mittelwert aller Beobachtungen.

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** beträgt *b* ÿ 1.

Die **Varianz** des **Faktors** B ist dann gleich:

$$s_B^2 = \frac{SSB}{b \text{ ÿ } 1}.$$

Varianz der Interaktion AB

Für die **Varianz der Interaktion AB** ist die Summe Für die Wechselwirkung AB *(SSAB) muss die Anzahl der Quadrate berechnet werden,* die sich wie folgt ergibt:

$$SSAB = c \sum_{i=1}^{A} \sum_{j=1}^{B} (\bar{Y}ij. \text{ ÿ } \bar{Y}i.. \text{ ÿ } \bar{Y}.j. + \bar{Y}...)^2,$$

wo $\bar{Y}ij.$ **ist der Mittelwert aller Beobachtungen** oflevelijofinteractionAB,$\bar{Y}$i..isthemeane of alle Beobachtungen der Ebene *i* des Faktors A, $\bar{Y}.j.$ ist der Mittelwert aller Beobachtungen der Ebene *j* des Faktors B und $\bar{Y}...$ ist der allgemeine Mittelwert von allen die Beobachtungen.

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** ist gleich *(a* ÿ 1)(b ÿ 1).

Die **Varianz** der **Wechselwirkung** AB ist dann gleich:

$$s_{AB}^2 = \frac{SSAB}{(a \text{ ÿ } 1)(b \text{ ÿ } 1)}.$$

Abweichung innerhalb der Behandlungen oder Fehler

Für die **Varianz** innerhalb der Behandlungen die Summe der Quadrate innerhalb der Behandlungen wird erhalten als folgt:

$$SSE = \sum_{i=1}^{A} \sum_{j=1}^{B} \sum_{k=1}^{C} (Yijk \text{ ÿ } \bar{Y}ij.)^2,$$

wobei *a* und *b* jeweils die Zahl sind der **Stufen** der **Faktoren** A und B ist *c* die Zahl Von den Beobachtungen, die die Behandlung *ij* (Konstante für *anyiandj) erhalten,* ist Yijk die Beobachtungk von Stufe *i* von Faktor A und Stufe *j* von Faktor B, und $\bar{Y}ij.$ ist der Mittelwert aller Beobachtungen von Ebene *ij* der Interaktion AB.

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** ist gleich ab(c ÿ 1).

Die **Varianz** innerhalb der Behandlungen ist somit gleich Zu:

$$s_E^2 = \frac{SSE}{ab(c \text{ ÿ } 1)}.$$

Die Gesamtquadratsumme *(SST)* ist gleich:

$$SST = \sum_{i=1}^{A} \sum_{j=1}^{B} \sum_{k=1}^{C} Y2_{ijk}.$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** ist gleich N, also die Gesamtzahl der **Beobachtungen.**

Die Summe der Quadrate für den **Mittelwert** ist gleich:

$$SSM = NY\bar{}2_{...} \cdot$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** ist gleich 1.

Die Gesamtquadratsumme *(SST)* kann sein ausgedrückt mit allen anderen Quadratsummen auf die folgende Weise:

$$SST = SSM + SSA + SSB + SSAB + SSE.$$

Fisher-Tests

Wir können nun die verschiedenen *F*- Verhältnisse berechnen.

Machine Translated by Google

Um die **Nullhypothese** zu testen

$$H0 : \ddot{y}1 = \ddot{y}2 = \quad \cdots \quad = \ddot{y}a ,$$

Das erste *F* -Verhältnis wird gebildet, dessen Zähler ist die **Schätzung** der **Varianz** von **fac Tor** A und Nenner ist die Schätzung von die Varianz innerhalb der Behandlungen:

$$F = \frac{s^2_A}{s^2_E} .$$

WennF größer oder gleich dem **Wert** in ist **Fisher-Tabelle** für aÿ1 und *ab(cÿ1)* **Grad der Freiheit** wird die **Nullhypothese** abgelehnt und daraus wird geschlossen, dass mindestens ein *ÿi* von ÿj verschieden ist , *i* = j.
Um die **Nullhypothese** zu testen

$$H0 : \ddot{y}1 = \ddot{y}2 = \quad \cdots \quad = \ddot{y}b ,$$

Das zweite *F*- Verhältnis wird gebildet, dessen Zähler ist die **Schätzung** der **Varianz von Faktor B** und Nenner ist die Schätzung der Vari ance innerhalb der Behandlungen:

$$F = \frac{s^2_B}{s^2_E} .$$

WennF größer oder gleich dem **Wert** in ist **Fisher-Tabelle** für bÿ1 und *ab(cÿ1)* **Grad der Freiheit** wird die **Nullhypothese** abgelehnt und es wird gefolgert, dass H, wahr ist.
Um die **Nullhypothese** zu testen

$$H0 : (\ddot{y}\ddot{y})11 = (\ddot{y}\ddot{y})12 = = (\ddot{y}\ddot{y})21 \quad \cdots \quad = (\ddot{y}\ddot{y})1b$$
$$= \quad \cdots \quad = (\ddot{y}\ddot{y})ab$$

Das dritte *F* -Verhältnis wird erstellt, dessen Zähler ist die **Schätzung** der **Varianz** von **Interac tion** AB und deren Nenner die esti ist mation der Varianz innerhalb von Behandlungen:

$$F = \frac{s^2_{AB}}{s^2_E} .$$

WennF größer oder gleich dem **Wert** in ist **Fisher-Tabelle** für (aÿ1)(bÿ1) und *ab(cÿ1)* **Freiheitsgrade** lautet die **Nullhypothese** abgelehnt und es wird gefolgert, dass H wahr ist.

Tabelle der Varianzanalyse

Alle zur Berechnung erforderlichen Informationen Zwei *F*- Verhältnisse können in einer Tabelle zusammengefasst werden **Varianzanalyse:**

Quelle von Varia tion	Abschlüsse von kostenlos dom	Die Summe von Quadrate	Bedeutung von Quadrate	F
Mittel 1		*SSM*		
Faktor A *a* ÿ 1		*SSA*	s^2_A	$\frac{s^2_A}{s^2_E}$
Faktor B *b* ÿ 1		*SSB*	s^2_B	$\frac{s^2_B}{s^2_E}$
Inter Aktion AB	*(a* ÿ 1) · *(b* ÿ 1)	*SSAB* s2	*AB*	$\frac{s^2_{AB}}{s^2_E}$
Innerhalb behandeln gen	*ab(cÿ1) SSI*		s^2_E	
Gesamt	*ABC*	*SST*		

Wenn *c* = 1, bedeutet dies, dass es nur einen gibt **Beobachtung** pro Behandlung *ij* (oder pro Zelle ij), Es ist nicht möglich, einen Teil des **Fehlers** zuzuordnen „innerhalb der Gruppen“ zur **Interaktion.**

T

BEISPIELE

Der Direktor einer Schule will vier testen Marken von Schreibmaschinen. Er bittet fünf professionelle Sekretärinnen, jede Marke auszuprobieren und zu wiederholen die Übung am nächsten Tag. Sie müssen alle tippen der gleiche Text und nach 15 Minuten der Durchschnitt Es wird die Anzahl der pro Minute eingegebenen Wörter aufgezeichnet. Hier sind die Ergebnisse:

Machine Translated by Google

Schreibmaschine	Sekretär				
	1	2	3	4	5
1	33 31		34 34 31		
	36 31		36 33 31		
2	32 37 40 33 35				
	35 35 36 36 36				
3	37 35 34 31				37
	39 35 37 35 40				
4	29 31		33 31		33
	31	33 34 27 33			

Lassen Sie uns eine **bidirektionale Varianzanalyse** durchführen , Der erste Faktor sind die Marken der Schreibmaschinen und der zweite die Sekretärinnen.

Die **Nullhypothesen** lauten wie folgt:

1.

H0 : ÿ1 = ÿ2 = ÿ3 = ÿ4

H1 : Mindestens ein *ÿi* ist unterschiedlich aus *ÿj, i* = j.

2.

H0 : ÿ1 = ÿ2 = ÿ3 = ÿ4 = ÿ5

H1 : Mindestens ein *ÿi* ist unterschiedlich aus *ÿj, i* = j.

3.

H0 : (ÿÿ)11 = (ÿÿ)12 = = (ÿÿ)21 ··· = (ÿÿ)15 = ··· = (ÿÿ)45

H1 : Mindestens eine der Interaktionen ist anders als die anderen.

Varianz von Faktor A

„Marken von Schreibmaschinen“

Varianz von Faktor A

Die Quadratsumme des **Faktors** A ist gleich:

$$SSA = 5 \cdot 2 \sum_{i=1}^{4} (\bar{Y}_{i..} - \bar{Y}_{...})^2$$

$$= 5 \cdot 2 (33 - 34)^2 + (35{,}5 - 34)^2 + (36 - 34)^2 + (31{,}5 - 34)^2$$

$$= 135 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** beträgt 4 ÿ 1 = 3.

Die **Varianz des Faktors** A **(Mittelwert** der Quadrate) ist dann gleich:

$$s_A^2 = \frac{SSA}{a - 1} = \frac{135}{3} = 45 .$$

Varianz des Faktors B „Sekretärinnen“

Die Quadratsumme des **Faktors** B ist gleich:

$$SSB = 4 \cdot 2 \sum_{j=1}^{5} (\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

$$= 4 \cdot 2 (34 - 34)^2 + (33{,}5 - 34)^2 + ... + (35{,}5 - 34)^2 + (34{,}5 - 34)^2 = 40 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** beträgt 5 ÿ 1 = 4.

Die **Varianz** von **FaktorB (Mittelwert** der Quadrate) ist dann gleich:

$$s_B^2 = \frac{SSB}{b - 1} = \frac{40}{4} = 10 .$$

Varianz der Interaktion AB

„Marken von Schreibmaschinen – Sekretäre“

Die Summe der **Wechselwirkungsquadrate** AB ist gleich:

$$SSAB = 2 \sum_{i=1}^{4} \sum_{j=1}^{5} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$$

$$= 2 (34{,}5 - 33 - 34 + 34) 2 + (31 - \textit{31,5} - \textit{34,5} + 34)^2 + ... + (33 - \textit{31,5} - \textit{34,5} + 34)^2$$

Machine Translated by Google

$$= 2 \cdot 1{,}52 + (\ddot{y}1{,}5)^2 + \ldots + 12$$

$$= 83 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** beträgt (4ÿ1)(5ÿ1) = 12.

Die **Varianz** der **Wechselwirkung** AB ist dann gleich:

$$^2AB = \frac{SSAB}{(a \ddot{y} 1)(b \ddot{y} 1)\ 83}$$

$$= \frac{}{12}$$

$$= 6{,}92 .$$

Abweichung innerhalb der Behandlungen (Fehler)

Die Summe der Quadrate innerhalb der Behandlungen ist gleich:

$$SSE = \sum_{i=1}^{4} \sum_{j=1}^{5} \sum_{k=1}^{2} (Yijk \ddot{y} \bar{Y}ij.)^2$$

$$2\ 2 = (33 \ddot{y} 34{,}5) + (36 \ddot{y} 34{,}5)^2 + (31 \ddot{y} 31)^2 + \ldots + (33 \ddot{y} 33)^2$$

$$= 58 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** beträgt 4 ·5(2ÿ1) = 20:

$$^2E = \frac{SSI}{ab(c \ddot{y} 1)\ 58}$$

$$= \frac{}{20}$$

$$= 2{,}9 .$$

Die Gesamtsumme der Quadrate ist gleich:

$$SST = \sum_{i=1}^{4} \sum_{j=1}^{5} \sum_{k=1}^{2} (Yijk)^2$$

$$= 332 + 362 + 312 + \ldots + 272 + 332 + 332$$

$$= 46556 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** entspricht der Anzahl *N* der Beobachtungen, also 40.

Die Summe der Quadrate für den **Mittelwert** ist gleich:

$$SSM = NY\bar{}2..$$

$$= 40 \cdot 342$$

$$= 46240 .$$

Die Anzahl der dieser Summe zugeordneten **Freiheitsgrade** ist gleich 1.

Fisher-Tests

Um die **Nullhypothese** zu testen

$$H0 : \ddot{y}1 = \ddot{y}2 = \ddot{y}3 = \ddot{y}4 ,$$

Wir bilden das erste *F* -Verhältnis, dessen Zähler ist die **Schätzung** der **Varianz** von **fac Tor** A und Nenner ist die Schätzung von die Varianz innerhalb der Behandlungen:

$$F = \frac{s^2_A}{s^2}$$

$$= \frac{E\,45}{2.9}$$

$$= 15{,}5172 .$$

Der **Wert** der **Fisher-Tabelle** mit 3 und 20 **Freiheitsgrade** für ein **signifikantes Niveau** ÿ = 0,05 sind gleich 3,10.

Da *F* ÿ *F3,20,0,05,* lehnen wir die H0- **Hypothesen** ab und kommen zu dem Schluss, dass die Marken von Schreibmaschinen einen signifikanten Einfluss auf die Anzahl der in 1 Minute getippten Wörter haben.

Um die **Nullhypothese** zu testen

$$H0 : \ddot{y}1 = \ddot{y}2 = \ddot{y}3 = \ddot{y}4 = \ddot{y}5 ,$$

Wir bilden die *zweiteFratio,* deren Zähler ist die **Schätzung** der **Varianz von Faktor B** und dessen Nenner die Schätzung von ist

T

Machine Translated by Google

die Varianz innerhalb der Behandlungen:

$$F = \frac{s^2_B}{s^2_E} = \frac{10}{2.9} = 3{,}4483 .$$

Der **Wert** der **Fisher-Tabelle** mit 4 und 20 **Freiheitsgrade** für ÿ = 0,05 sind gleich bis 2,87.

Da *F* ÿ *F4,20,0,05,* lehnen wir H0 ab **Hypothesen** und schlussfolgern, dass die Sekretäre einen erheblichen Einfluss auf die Anzahl haben In 1 Minute eingegebene Wörter.

Um die **Nullhypothese** zu testen

$$H0 : (ÿÿ)11 = (ÿÿ)12 = = (ÿÿ)21 \cdots = (ÿÿ)15 = \cdots = (ÿÿ)45 ,$$

Wir bilden das dritte *F*- Verhältnis, dessen Zähler ist die **Schätzung** der **Varianz von interac tion** AB und deren Nenner die esti ist mation der Varianz innerhalb von Behandlungen:

$$F = \frac{s^2_{AB}}{s^2_E} = \frac{6{,}92}{2.9} = 2{,}385 .$$

Der **Wert** der **Fisher-Tabelle** mit 12 und 20 **Freiheitsgrade** für ÿ = 0,05 sind gleich bis 2,28.

Da *F* ÿ *F4,20,0,05,* lehnen wir H0 ab **Hypothesen** und kommen zu dem Schluss, dass die Wirkung von die Interaktion zwischen der Sekretärin und dem Der Typ der Schreibmaschine ist von Bedeutung für die Anzahl der Wörter, die in einer Minute getippt werden.

Tabelle der Varianzanalyse

Alle diese Informationen sind in der zusammengefasst Tabelle der **Varianzanalyse:**

Quelle von Variation	Abschlüsse von kostenlos dom	Die Summe von Quadrate	Bedeutung von Quadrate	F
Bedeuten	1	46240		
Faktor A	3	135 45		15.5172
Faktor B	4	40 10		3.4483
Inter Aktion AB	12	83	6,92	2,385
Innerhalb behandeln gen	20	58	2.9	
Gesamt	40	46556		

WEITERLESEN

- Varianzanalyse
- Kontrast
- Fisher-Tisch
- Fisher-Test
- Interaktion
- Test auf den geringsten signifikanten Unterschied
- Einseitige Varianzanalyse

VERWEISE

Cox, DR, Reid, N.: Theorie des Designs von Experimenten. Chapman & Hall, London (2000)

Fisher, RA: Statistische Methoden für Forschungsmitarbeiter. Oliver & Boyd, Edinburgh (1925)

Yates, F.: Komplexe Experimente. Zus. J. Roy. Stat. Soc. 2, 181–247 (1935)

Yates, F.: Das Design und die Analyse faktorieller Experimente. Technische Kommunikation des Commonwealth Bureau of Soils **35,** Commonwealth Agricultural Bureaux, Farnham Royal (1937)

Fehler vom Typ I

Wenn **Hypothesentests** durchgeführt werden, Ein Fehler vom Typ I tritt auf, wenn man das ablehnt

Machine Translated by Google

Nullhypothese , wenn sie wahr ist und ÿ die ist
Wahrscheinlichkeit , die **Nullhypothese** abzulehnen
H0 , wenn es wahr ist:

ÿ = P(H0 ablehnen | H0 ist wahr) .

Die Wahrscheinlichkeit eines Fehlers vom Typ I ÿ ist gleich
das **Signifikanzniveau des Hypothesentests.**

GESCHICHTE

Im Jahr 1928 waren Neyman, J. und Pearson, ES
Die ersten Autoren erkannten, dass eine rationale Auswahl
beim **Testen von Hypothesen** getroffen werden muss
Berücksichtigen Sie nicht nur die **Hypothese ,** dass
Man will aber auch die **Alternative** verifizieren
Hypothese. Sie führten den Typ-I **-Fehler** ein
und der **Typ-II-Fehler.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Fehler vom Typ I ist einer der beiden **Fehler**
Statistiker wird in der **Hypothese** konfrontiert
Testen: Es handelt sich um die Art des Fehlers, der auftreten kann
bei der Entscheidungsfindung, wenn die **Nullhypothese** lautet
WAHR.

Wenn die **Nullhypothese** falsch ist, ein anderer Typ
Es entsteht ein Fehler vom Typ **II ,** der das Akzeptieren der
Nullhypothese H0 bedeutet
ist falsch.

Die verschiedenen Fehlerarten können in der folgenden
Tabelle dargestellt werden:

Situation	Entscheidung	
	Akzeptieren H_0	Ablehnen H_0
H0 wahr	1 ÿ ÿ	ÿ
H0 falsch	ÿ	1 ÿ ÿ

wobei ÿ die Wahrscheinlichkeit des Typ-I-Fehlers ist
und ÿ ist die Wahrscheinlichkeit des **Typ-II-Fehlers.**

BEISPIELE

Wir veranschaulichen dies anhand eines Beispiels aus
Alltag, die verschiedenen Fehler, die man machen kann
sich bei der Entscheidungsfindung verpflichten.

Wenn wir morgens das Haus verlassen, wir
Ich frage mich, wie das Wetter sein wird. Wenn wir
Da wir denken, dass es regnen wird, nehmen wir einen Regenschirm.
Wenn wir glauben, dass es sonnig sein wird, tun wir das nicht
Nimm irgendetwas gegen schlechtes Wetter mit.
Deshalb stehen wir vor Folgendem
ing **-Hypothesen:**

Nullhypothese H0: Es wird regnen.

Alternativhypothese H1: Es wird passieren
Sei sonnig.

Nehmen wir an, wir akzeptieren die Regenhypothese und
nimm einen Regenschirm.

Sollte es tatsächlich regnen, haben wir Vorkehrungen getroffen
Entscheidung, aber wenn es sonnig ist, haben wir einen Fehler gemacht:
der Fehler, eine falsche Hypothese zu akzeptieren.
Im umgekehrten Fall, wenn wir den Regen ablehnen
Hypothese, wir haben eine gute Entscheidung getroffen
Wenn es sonnig ist, haben wir jedoch einen Fehler gemacht
wenn es regnet: der Fehler, eine wahre Hypothese abzulehnen.

Wir können diese verschiedenen Arten von darstellen
Fehler in der folgenden Tabelle:

Situation	Entscheidung	
	Akzeptieren H_0 (Ich nehme eine Regenschirm)	Ablehnen H_0 (Ich nehme nicht ein Regenschirm)
H0 wahr (Es regnet)	Gut Entscheidung	Fehler vom Typ I
H0 falsch (es ist sonnig)	Fehler vom Typ II ÿ Gut	Entscheidung

Die Wahrscheinlichkeit, eine wahre Hypothese abzulehnen ,
wird als Signifikanzniveau bezeichnet.
Die **Wahrscheinlichkeit** ÿ, ein Falsches zu akzeptieren
Hypothese ist die Wahrscheinlichkeit von **Typ II**
Fehler.

WEITERLESEN

- **Hypothesentest**
- **Signifikanzniveau**
- **Fehler vom Typ II**

T

Machine Translated by Google

VERWEISE

Neyman, J., Pearson, ES: Zur Verwendung und Interpretation bestimmter Testkriterien für statistische Schlussfolgerungen, Teile I und II. Biometrie 20A, 175–240, 263–294 (1928)

Fehler vom Typ II

Bei der Durchführung eines **Hypothesentests** tritt ein **Fehler** vom Typ II auf, wenn die Nullhypothese akzeptiert wird, sofern sie falsch ist, und ÿ die **Wahrscheinlichkeit** ist, die **Nullhypothese** H0 zu akzeptieren , wenn sie falsch ist:

$$ÿ = P(\text{akzeptiere H0} \mid \text{H0 ist falsch}) .$$

GESCHICHTE

Siehe **Fehler vom Typ I.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Wahrscheinlichkeit eines **Fehlers** vom Typ II wird mit ÿ bezeichnet, die **Wahrscheinlichkeit ,** die **Nullhypothese** abzulehnen, wenn sie falsch ist, wird mit 1 ÿ ÿ angegeben:

$$ÿ = P(\text{H0 ablehnen} \mid \text{H0 ist falsch}) .$$

Die **Wahrscheinlichkeit** 1ÿÿ wird als Trennschärfe des Tests bezeichnet.

Daher ist es immer notwendig, das Risiko eines Typ-II-**Fehlers zu minimieren,** was bedeutet, dass die Aussagekraft des **Hypothesentests erhöht wird.**

BEISPIELE

Siehe **Fehler vom Typ I.**

WEITERLESEN

Hypothesentest

Fehler vom Typ I

Machine Translated by Google

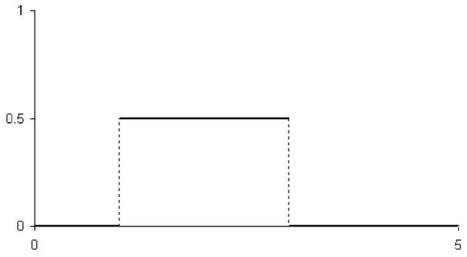
Gleichmäßige Verteilung

Zur Beschreibung wird eine Gleichverteilung verwendet eine **Population ,** die in einem Intervall gleichmäßig verteilt ist .

Eine Zufallsvariable X soll im **Intervall [a,** b] gleichmäßig verteilt sein, wenn ihre **Dichtefunktion** gegeben ist durch:

$$f(x) = \begin{cases} \frac{1}{b - a} & \text{wenn } a \le x \le b \\ 0 & \text{sonst .} \end{cases}$$

Aufgrund der rechteckigen Form ihrer Dichtefunktion wird die Gleichverteilung manchmal auch als Rechteckverteilung bezeichnet.



Gleichmäßige Verteilung, a = 1, b = 3

Die Gleichverteilung ist eine **stetige Wahrscheinlichkeitsverteilung.**

MATHEMATISCHE ASPEKTE

Die **Verteilungsfunktion** einer Uniform

Die Verteilung ist wie folgt:

$$F(x) = \begin{cases} 0, \text{ wenn } x < a \\ \frac{x - a}{b - a} & \text{wenn } a < x < b \\ 1 & \text{sonst .} \end{cases}$$

Der **erwartete Wert** einer **Zufallsvariablen ,** die gleichmäßig in einem **Intervall** *[a,* b] verteilt ist, ist gleich:

$$E[X] = \int_A^B \frac{x}{b - a} \, dx = \frac{b^2 - a^2}{2(b - a)} = \frac{a + b}{2} .$$

Dieses Ergebnis ist intuitiv sehr leicht zu verstehen, da die **Dichtefunktion** symmetrisch um den Mittelpunkt des **Intervalls** *[a,* b] ist.

Wir können rechnen

$$E[X^2] = \int_A^B x^2 f(x) \, dx = \frac{b^2 + ab + a^2}{3} .$$

Die **Varianz** ist daher gleich:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \frac{(b - a)^2}{12} .$$

Die Gleichverteilung ist ein Sonderfall der **Betaverteilung ,** wenn = = 1.

DOMÄNEN UND EINSCHRÄNKUNGEN

Es gibt zahlreiche Anwendungen der Gleichverteilung bei der Konstruktion von **Modellen** für

physikalische, biologische und soziale Phänomene. Die Gleichverteilung wird oft als verwendet

Machine Translated by Google

ein **Modell** für **Approximationsfehler** bei Messungen. Wenn beispielsweise die Länge einiger Objekte mit einer Genauigkeit von 1 cm gemessen wird, kann eine Länge von 35 cm alle Objekte mit einer Länge zwischen 34,5 und 35,5 cm darstellen. Die Approximationsfehler folgen einer gleichmäßigen Verteilung im Intervall

[ÿ0,5, 0,5].

Die Gleichverteilung wird häufig verwendet, um **Zufallszahlen** einer beliebigen diskreten oder kontinuierlichen **Wahrscheinlichkeitsverteilung zu generieren.** Die Methode zur Generierung einer kontinuierlichen Verteilung ist die folgende. Betrachten Sie *X* als kontinuierliche **Zufallsvariable** und F(x)seine **Verteilungsfunktion.** Nehmen wir an, dass *F* stetig und streng steigend ist. Dann ist *U = F(X)* eine gleichmäßige und kontinuierliche Zufallsvariable. *U* nimmt seine **Werte** an im Intervall [0,1]. Wir haben dann *X* = Fÿ1(U), wobei Fÿ1 die Umkehrfunktion ist der **Verteilungsfunktion.**

Wenn also u1, *u2,...,un* eine Reihe von Zufallszahlen ist, dann gilt:

$$x1 = F\ddot{y}1(u1)\ ,\ x2$$

$$= F\ddot{y}1(u2)\ ,$$

$$\ldots\ ,$$

$$xn = F\ddot{y}1(un)$$

ist eine Reihe von Zufallszahlen, die aus der Verteilung der **Zufallsvariablen** X generiert werden.

BEISPIELE

Wenn **die Zufallsvariable** *X* einer negativen **Exponentialverteilung** folgt , ist der Mittelwert gleich zu 1, dann ist seine **Verteilungsfunktion**

$$F(x) = 1\ \ddot{y}\ \exp(\ddot{y}x)\ .$$

Da $F(x) = u$, gilt $u = 1\ \ddot{y}\ e\ddot{y}x$.
Um x zu finden, suchen wir nach Fÿ1:

$$u = 1\ \ddot{y}\ e\ddot{y}x\ ,$$

$$u\ \ddot{y}\ 1 = \ddot{y}e\ddot{y}x\ ,$$

$$1\ \ddot{y}\ u = e\ddot{y}x\ ,$$

$$\ln(1\ \ddot{y}\ u) = \ddot{y}x\ ,$$

$$\ddot{y}\ \ln(1\ \ddot{y}\ u) = x\ .$$

Beachten Sie, dass das gilt, wenn *u* eine **Zufallsvariable** ist einheitlich in [0,1], dann ist (1 ÿ u) auch eine **Zufallsvariable ,** die einheitlich in [0,1] ist. Indem wir das Vorstehende auf eine Reihe von **Zufallszahlen anwenden,** finden wir eine **Stichprobe** aus X:

U	X
0,14	1,97
0,97	0,03
0,53	0,63
0,73	0,31
⋮	⋮

WEITERLESEN

Beta-Verteilung

Kontinuierliche Wahrscheinlichkeitsverteilung

Diskrete Gleichverteilung

Unimodale Verteilung

Siehe **Häufigkeitsverteilung.**

Machine Translated by Google

V

Wert

Ein Wert ist ein quantitatives oder qualitatives Maß, das einer **Variablen zugeordnet ist.**

Wir sprechen von einer Reihe von Werten, die von einer Variablen übernommen werden. Die Werte können Kardinalzahlen sein, wenn Sie sind einer **quantitativen Variablen** oder Ordnungszahlen zugeordnet, wenn sie einer **qualitativen kategorialen Variablen** zugeordnet sind .

BEISPIELE

Die Werte, die eine bestimmte **Variable** annimmt, können folgender Art sein: •

Betrachten Sie die **quantitative Variable** „Anzahl der Kinder“. Die Werteassozi

Dieser Variable sind die natürlichen Zahlen zugeordnet

Ber 0, 1, 2, 3, 4 usw.

• Betrachten Sie die **qualitative kategoriale Variable** „Geschlecht“. Die zugehörigen Werte sind die Kategorien „männlich“ und „weiblich“.

WEITERLESEN

Qualitative kategoriale Variable
Quantitative Variable
Variable

Variable

Eine Variable ist ein messbares Merkmal für die auf viele verschiedene Variablen zurückzuführen sind.

Wir unterscheiden viele Arten von Variablen: • **Quantitative Variable,** von denen wir unterscheiden:

– Diskrete Variable, z. B. das Ergebnis von ein Test

– Kontinuierliche Variable, z. B. ein Gewicht oder Einnahmen

• **Qualitative kategoriale Variable,** z. B. Familienstand

WEITERLESEN

Daten
Dichotome Variable
Qualitative kategoriale Variable
Quantitative Variable
Zufällige Variable
Wert

Varianz

Varianz ist ein **Maß für die Streuung** einer Verteilung einer **Zufallsvariablen.**

Empirisch ist die Varianz einer **quantitativen Variablen** *X* als die Summe der quadratischen Abweichungen jeder **Beobachtung** relativ zu definiert das **arithmetische Mittel** dividiert durch die Anzahl der Beobachtungen.

GESCHICHTE

Die **Varianzanalyse ,** wie wir sie heute verstehen und praktizieren, war grundsätzlich

Machine Translated by Google

Entwickelt von **Fisher, Ronald Aylmer** (1918, 1925, 1935). Er führte auch die Begriffe Varianz und Varianzanalyse ein.

MATHEMATISCHE ASPEKTE

Die Varianz wird im Allgemeinen mit S2 bezeichnet , wenn sie relativ zu einer **Stichprobe** ist , und mit ÿ2 , wenn sie relativ zu einer **Grundgesamtheit ist.**

Wir bezeichnen die Varianz auch mit Var(X), wenn wir über die **Varianz einer Zufallsvariablen sprechen.**

Eine Grundgesamtheit von Beobachtungen sei relativ zu einer **quantitativen Variablen** X. Per Definition wird die Varianz einer Grundgesamtheit wie folgt berechnet:

$$\ddot{y}2 = \frac{\sum_{i=1}^{N} (xi \ddot{y} \ddot{y})^2}{N},$$

wobei *N* die Dimension der Population und ÿ der **Mittelwert** der Beobachtung ist:

$$\ddot{y} = \frac{\sum_{i=1}^{N} xi}{N}.$$

Wenn die Beobachtungen in der Reihenfolge angeordnet sind

In Form einer **Häufigkeitsverteilung** erfolgt die Berechnung der Varianz im Folgenden

leise Art:

$$\ddot{y}2 = \frac{\sum_{i=1}^{k} fi \cdot (xi \ddot{y} \ddot{y})2}{\sum_{i=1}^{k} fi},$$

Dabei sind *xi* die verschiedenen Werte der **Variablen, fi** die diesen Werten zugeordneten Häufigkeiten und *k* die Anzahl der verschiedenen Werte ues.

Um die Varianz der Häufigkeitsverteilung eines quantitativen X zu berechnen, bei dem die Werte in Klassen gruppiert sind, berücksichtigen wir Folgendes

Alle Beobachtungen, die zu einer bestimmten Klasse gehören, nehmen die Werte des Klassenmittelpunkts an.

Sie ist nur dann richtig, wenn die **Hypothese ,** dass die Beobachtungen innerhalb jeder Klasse gleichmäßig verteilt sind, bestätigt wird. Wenn dies Wird die Hypothese nicht verifiziert, ist der erhaltene Wert der Varianz nur näherungsweise.

Für die in Klassen gruppierten Werte gilt:

$$\ddot{y}2 = \frac{\sum_{i=1}^{k} fi \cdot (\ddot{y}i \ddot{y} \ddot{y})^2}{\sum_{i=1}^{k} fi},$$

Dabei sind *ÿi* die Zentren der Klassen, *fi* die jeder Klasse zugeordneten Häufigkeiten und *k* die Anzahl der Klassen.

Die Formel zur Berechnung der Varianz kann geändert werden, um die Berechnungszeit zu verkürzen tion und zur Erhöhung der Präzision. Daher ist es besser, die Varianz mit fol. zu berechnen

Lowing-Formel:

$$\ddot{y}2 = \frac{N \cdot \sum_{i=1}^{N} x2_i - \left(\sum_{i=1}^{N} xi\right)^2}{N2}.$$

Die Varianz entspricht dem zentrierten **Moment** zweiter Ordnung.

Die an einer Stichprobe gemessene Varianz **ist ein Maß** für die Varianz der Grundgesamtheit.

Betrachten Sie eine Stichprobe von *n* Beobachtungen relativ zu einer quantitativen Variablen *X* und bezeichnen Sie mit $\bar{x}$ das **arithmetische Mittel** dieser Stichprobe:

$$\bar{x} = \frac{\sum_{i=1}^{N} xi}{N}.$$

Damit die Varianz der Stichprobe ein Schätzwert ohne Verzerrung der Varianz der Grundgesamtheit sein kann, muss sie durch Division durch berechnet werden

Machine Translated by Google

(*n* ÿ 1) und nicht durch n:

$$S2 = \frac{\sum_{i=1}^{N} (xi \ddot{y} \bar{x})^2}{n \ddot{y} 1}.$$

BEISPIELE

Fünf Studenten bestanden zwei Prüfungen und erhielten die Auszeichnung folgende Noten:

Prüfung 1:

3.544,5 *3,5 4,5*

Prüfung 2:

2,5 5,5 3,5 4,5 4

Beachten Sie, dass das arithmetische **Mittel** $\bar{x}$ dieser beiden ist Beobachtungsreihen ist identisch:

$$\bar{x}1 = \bar{x}2 = 4 \quad .$$

Die Streuung der Beobachtungen um der **Mittelwert** ist nicht derselbe.

Um die Varianz zu ermitteln, berechnen wir zunächst die Abweichungen jeder Beobachtung von der arithmetisches Mittel und quadrieren Sie dann die Abweichungen:

Prüfung 1:

Grad	$(xi \ddot{y} x_1)$	$(xi \ddot{y} x_1)2$
3.5	*ÿ0,5*	0,25
4	0,0	0,00
4.5	0,5	0,25
3.5	*ÿ0,5*	0,25
4.5	0,5	0,25
$\sum_{i=1}^{5} xi \ddot{y} \bar{x}1^2$	1,00	

Prüfung 2:

Grad	$xi \ddot{y} x_2$	$xi \ddot{y} x_2^2$
2.5	*ÿ1,5*	2,25
5.5	1.5	2,25
3.5	*ÿ0,5*	0,25
4.5	0,5	0,25
4	0,0	0,00
$\sum_{i=1}^{5} xi \ddot{y} \bar{x}2^2$	5.00	

Die Varianz der Noten jeder Prüfung entspricht:

Prüfung 1:

$$S2_1 = \frac{\sum_{i=1}^{5} (xi \ddot{y} \bar{x}1)^2}{n \ddot{y} 1} = \frac{1}{4} = 0{,}25$$

Prüfung 2:

$$S2_2 = \frac{\sum_{i=1}^{5} (xi \ddot{y} \bar{x}2)^2}{n \ddot{y} 1} = \frac{5}{4} = 1{,}25$$

Da die Varianz der Noten der zweiten Prüfung größer ist, wird bei der zweiten Prüfung die Die Notenstreuung liegt stärker um das arithmetische Mittel herum als bei der ersten Prüfung.

WEITERLESEN

Varianzanalyse

Maß der Streuung
Standardabweichung

Varianz einer Zufallsvariablen

VERWEISE

Fisher, RA: Die Korrelation zwischen Verwandten nach der Mendelschen Annahme Nachlass. Trans. Roy. Soc. Edinb. **52,** 399–433 (1918)

Fisher, RA: Statistische Methoden für Forschungsmitarbeiter. Oliver & Boyd, Edinburgh (1925)

Fisher, RA: Das Design von Experimenten. Oliver & Boyd, Edinburgh (1935)

Machine Translated by Google

Varianzanalysen

Siehe **Varianzanalyse.**

Varianz einer Zufallsvariablen

Die **Varianz** einer quantitativen **Zufallsvariablen** misst den Wert des Mittelwerts
Abweichungen der Werte dieser Zufallsvariable
aus dem **mathematischen Erwartungswert** (d. h. aus seinem **Mittelwert) abgeleitet werden kann.**

MATHEMATISCHE ASPEKTE

Je nachdem, ob die **Zufallsvariable** diskret oder stetig ist, spricht man von der Varianz einer diskreten oder stetigen
zufällige Variable.
Für den Fall, dass die Zufallsvariable dis ist
creteX nimmt *n* Werte x1, *x2,...,xn* mit relativen Häufigkeiten f1, *f2,...,fn an,* die **Varianz** ist definiert durch:

$$\ÿ2 = \text{Var}(X) = E\,(X \ÿ \ÿ)^2 = \sum_{i=1}^{N} fi\,(xi \ÿ \ÿ)^2 ,$$

wobei ÿ den **mathematischen Erwartungswert** von X darstellt.
Wir können eine andere Formel dafür aufstellen
Die Berechnung der Varianz erfolgt folgendermaßen:

$$\begin{aligned}\text{Var}(X) &= E\,(X \ÿ \ÿ)^2\\ &= E\,X2 \ÿ 2\ÿX + \ÿ2\\ &= E\,X2 \ÿ E\,[2\ÿX] + E\,\ÿ2\\ &= E\,X2 \ÿ 2\ÿE\,[X] + \ÿ2\\ &= E\,X2 \ÿ 2\ÿ2 + \ÿ2\\ &= E\,X2 \ÿ \ÿ2\\ &= E\,X2 \ÿ (E\,[X])^2\\ &= \sum_{i=1}^{N} fix2 \ÿ \Big(\sum_{i=1}^{N} fixi\Big)^2 .\end{aligned}$$

In der Praxis eignet sich diese Formel im Allgemeinen besser für die Berechnung.
Wenn die Zufallsvariable *X* stetig ist in
Intervall D, der Ausdruck der Varianz wird:

$$\ÿ2 = \text{Var}(X) = E\,(X \ÿ \ÿ)^2 = \int_D (x \ÿ \ÿ)^2 f\,(x)\,dx .$$

Wir können diese Formel auch durch die gleichen Operationen wie im diskreten Fall transformieren, sodass die Varianz gleich ist:

$$\text{Var}(X) = \int_D x2f(x)dx \ÿ \Big(\int_D xf(x)dx\Big)^2 .$$

Eigenschaften der Varianz

1. Seien *a* und *b* zwei Konstanten und *X* a ran
   Dom-Variable:

$$\text{Var}(aX + b) = a2 \cdot \text{Var}(X) .$$

2. Seien *X* und *Y* zwei Zufallsvariablen:

$$\begin{aligned}\text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y)\\ &\quad + 2\text{Cov}\,(X, Y),\\ \text{Var}(X \ÿ Y) &= \text{Var}(X) + \text{Var}(Y)\\ &\quad \ÿ 2\text{Cov}\,(X, Y),\end{aligned}$$

   wobei Cov *(X,* Y) die **Kovarianz** zwischen *X* und Y ist. Insbesondere wenn *X* und *Y* unabhängig sind:

$$\begin{aligned}\text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y)\\ \text{Var}(X \ÿ Y) &= \text{Var}(X) + \text{Var}(Y) .\end{aligned}$$

Machine Translated by Google

BEISPIELE

Wir betrachten zwei Beispiele, von denen eines besorgniserregend ist eine diskrete **Zufallsvariable,** eine andere eine kontinuierliche Zufallsvariable.

Wir würfeln mehrmals hintereinander. Angenommen, wir gewinnen 1 Euro, wenn das Ergebnis ausgeglichen ist und 2 Euro, wenn das Ergebnis 1 oder 3 ist und wir verlieren 3 Euro, wenn das Ergebnis 5 ist.

Die Zufallsvariable *X* beschreibt die Anzahl von Euro gewonnen oder verloren. Die folgende Tabelle stellt verschiedene Werte von *X* und ihre jeweiligen **Wahrscheinlichkeiten dar:**

X	ÿ3 1 2		
P (X)	1/6	3/6	2/6

Die Varianz der diskreten Zufallsvariablen *X* entspricht:

$$\text{Var}(X) = \sum_{i=1}^{3} fix2 - \sum_{i=1}^{3} fixi$$

$$2\,\frac{2}{6} = (ÿ3)\,\frac{1}{6} + (1)\,\frac{3}{6} + (2)\,\frac{2}{6}$$

$$ÿ\,ÿ3\,\frac{1}{6} + 1\,\frac{3}{6} + 2\,\frac{2}{6}$$

$$= \frac{26}{9}.$$

Betrachten Sie eine kontinuierliche Zufallsvariable *X* **deren Dichtefunktion** auf (0, 1) einheitlich ist:

$$f(x) = \begin{cases} 1 \text{ für } 0 < x < 1 \\ 0 \text{ sonst} \end{cases}.$$

Der **mathematische Erwartungswert** dieser Zufallsvariablen ist gleich $\frac{1}{2}$.
Wir können die Varianz des Zufalls berechnen Variable X:

$$\text{Var}(X) = \int_0^1 x2f(x)\,dx ÿ \left(\int_0^1 xf(x)\,dx\right)^2$$

$$= \int_0^1 x2 \cdot 1\,dx ÿ \left(\int_0^1 x \cdot 1\,dx\right)^2$$

$$= \frac{x3}{3}\Big|_0^1 - \left(\frac{x2}{2}\Big|_0^1\right)^2$$

$$= \frac{1}{3} - \left(\frac{1}{2}\right)^2$$

$$= \frac{1}{12}.$$

WEITERLESEN

Erwarteter Wert
Zufällige Variable

Varianz-Kovarianz-Matrix

Die Varianz-Kovarianz-Matrix **V** zwischen *n* Variablen *X1,...,Xn* ist eine Ordnungsmatrix *(n* × n), wobei die Diagonale die enthält **Varianzen** jeder VariablenXi und außerhalb der Diagonalen sind die **Kovarianzen** zwischen den Variablenpaare *(Xi, Xj)* für jedes *i* = j. Für *n* Variablen berechnen wir *n* Varianzen und *n(nÿ* 1) Kovarianzen. Das Ergebnis ist die Matrix **V** ist eine Matrix der Dimension *n* × n. Im Liter Natur, wir stoßen oft auf diese Matrix unten ein anderer Name: Dispersionsmatrix oder Kovarianzmatrix.

Wenn die Variablen standardisiert sind (siehe **standardisierte Daten),** sind die Varianzen der Variablen gleich 1 und die Kovarianzen werden Korrelationen, und so wird die Matrix eine Korrelationsmatrix.

V

GESCHICHTE

Siehe **Varianz** und **Kovarianz.**

MATHEMATISCHE ASPEKTE

Seien *X1,...,Xn n* Zufallsvariablen .
Die Varianz-Kovarianz-Matrix ist eine Matrix der Ordnung(n × n);ist symmetrisch und enthält

Machine Translated by Google

auf der Diagonale die Varianzen jeder Variablen und als andere Terme die Kovarianzen zwischen den Variablen. Matrix **V** ist die fol

Lowing-Matrix:

$$\begin{pmatrix} \mathrm{Var(X1)} & \mathrm{Cov\ (X1,\ X2)} & \cdots & \mathrm{Cov\ (X1,\ } Xn) \\ \mathrm{Cov\ (X2,\ X1)} & \mathrm{Var(X2)} & \cdots & \mathrm{Cov\ (X2,\ } Xn) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov\ (}Xn,\ \mathrm{X1)} & \mathrm{Cov\ (}Xn,\ \mathrm{X2)} & \cdots & \mathrm{Var(Xn)} \end{pmatrix},$$

mit

$$\mathrm{Var(Xi)} = E\,\mathrm{X2} - (E\,[Xi])^2 \quad \text{Und}$$

$$\mathrm{Cov}\ Xi,\ Xj = E\,(Xi - E[Xi])\,Xj - E[Xj]\,.$$

BEISPIELE

Betrachten Sie die folgenden drei Variablen X1, X2 und X3 in der folgenden Tabelle zusammengefasst:

X 1	X 2	X 3
1	2	3
2	3	4
1	2	3
5	4	3
4	4	4

Im Fall $n = 3$ besteht jede Zufallsvariable aus fünf Termen, die wir mit $N = 5$ bezeichnen. Wir wollen die Varianz berechnen – Kovarianzmatrix **V** dieser Variablen. Wir

Führen Sie die folgenden Schritte aus.

Wir schätzen den Erwartungswert jeder Variablen anhand ihres empirischen Mittelwerts $\overline{xi}$ und erhalten die folgenden Ergebnisse:

$$\overline{\mathrm{X1}} = 2{,}6$$

$$\overline{\mathrm{X2}} = 3$$

$$\overline{\mathrm{X3}} = 3{,}4\,.$$

Die Varianz für eine Variable Xj ist gegeben durch:

$$\frac{\sum_{i=1}^{N} (xij - \overline{xj})^2}{4}\,.$$

Die Kovarianz zwischen zwei Variablen Xr und Xs wird wie folgt berechnet:

$$\frac{\sum_{i=1}^{N} (xir - \overline{xr})\,(xis - \overline{xs})}{4}\,.$$

Dann sollten wir den Mittelwert Xi jeder Variablen Xi *subtrahieren.* Wir erhalten eine neue Variable $Yi = Xi - \overline{Xi}$.

Y 1	Y 2	Y 3
−1,6	−1	−0,4
−0,6	0	0,6
−1,6	−1	−0,4
2.4	1	−0,4
1.4	1	0,6

Beachten Sie, dass der Mittelwert jeder neuen Variablen Yi gleich 0 ist. Wir finden diese Ergebnisse in Matrixform:

$$\mathbf{Y} = \begin{pmatrix} -1{,}6 & -1 & -0{,}4 \\ -0{,}6 & 0 & 0{,}6 \\ -1{,}6 & -1 & -0{,}4 \\ 2.4 & 1 & -0{,}4 \\ 1.4 & 1 & 0{,}6 \end{pmatrix}.$$

Aus dieser Matrix berechnen wir die Matrix $\mathbf{Y} \cdot \mathbf{Y}$, eine symmetrische Matrix, in der folgenden Form:

$$\begin{pmatrix} \sum_{i=1}^{N} y_{i1}^2 & \sum_{i=1}^{N} yi2 \cdot yi1 & \sum_{i=1}^{N} yi3 \cdot yi1 \\ \sum_{i=1}^{N} yi2 \cdot yi1 & \sum_{i=1}^{N} y_{i2}^2 & \sum_{i=1}^{N} yi3 \cdot yi2 \\ \sum_{i=1}^{N} yi3 \cdot yi1 & \sum_{i=1}^{N} yi3 \cdot yi2 & \sum_{i=1}^{N} y_{i3}^2 \end{pmatrix}.$$

Diese Matrix enthält die Werte der Summe aus Produkten von Quadraten jeder Variablen Yi (auf der Diagonale) und dem Kreuzprodukt jedes Variablenpaares $(Yi,\ Yj)$ auf dem Rest der Matrix. Für die Originalvariablen Xi enthält die folgende Matrix auf der Diagonale die Summe der quadrierten Abweichungen vom Mittelwert für

Machine Translated by Google

jede Variable und auf die externen Terme die Kreuzprodukte dieser Abweichungen. Numerisch erhalten wir:

$$\begin{matrix} & 13.270.8 & \\ ÿ & 7\ 41 & \\ ÿ & 0{,}8\ 11{,}2 & ÿ\ ÿ\ . \end{matrix}$$

Der Mittelwert der Summe der quadrierten Abweichungen vom Mittelwert einer Zufallsvariablen *Xi* von *N* Termen wird als Varianz bezeichnet . Der Mittelwert der gekreuzten Produkte dieser Abweichungen wird **Kovarianz genannt.** So erhält man die Varianz-Kovarianz-Matrix *V* der Vari

Wenn *X1,...,Xn* möglich ist , sollten wir die Elemente der Matrix **Y** · **Y** durch *N* ÿ 1 dividieren, was mathematisch dasselbe ist wie:

$$\begin{matrix} ÿ & \frac{\sum_{i=1}^{N} y_{i1}^2}{Nÿ1} & \frac{\sum_{i=1}^{N} yi2 \cdot yi1}{Nÿ1} & \frac{\sum_{i=1}^{N} yi3 \cdot yi1}{Nÿ1} & ÿ \\ & \frac{\sum_{i=1}^{N} yi2 \cdot yi1}{Nÿ1} & \frac{\sum_{i=1}^{N} y_{i2}^2}{Nÿ1} & \frac{\sum_{i=1}^{N} yi3 \cdot yi2}{Nÿ1} & \\ ÿ & \frac{\sum_{i=1}^{N} yi3 \cdot yi1}{Nÿ1} & \frac{\sum_{i=1}^{N} yi3 \cdot yi2}{Nÿ1} & \frac{\sum_{i=1}^{N} y_{i3}^2}{Nÿ1} & ÿ \end{matrix} .$$

Die Tatsache, dass zur Ermittlung des Mittelwerts eine Division *durch Nÿ1* und nicht durch N vorgenommen wird, hängt mit der Tatsache zusammen, dass wir im Allgemeinen mit einer Stichprobe von Zufallsvariablen arbeiten, ohne die tatsächliche Population zu kennen und für die wir eine unverzerrte Schätzung vornehmen mation der Varianz und der Kovarianz.

Numerisch erhalten wir:

$$\mathbf{V} = \begin{matrix} & 3{,}3\ 1{,}75\ 1{,}75 & \\ ÿ & 1{,}75 \quad 1 \quad 0{,}2 & \\ ÿ & 0{,}2\ 0{,}25\ 0{,}3 & ÿ\ ÿ\ . \end{matrix}$$

WEITERLESEN

- **Korrelationskoeffizient**
- **Kovarianz**
- **Standardisierte Daten**
- **Varianz**

Vektor

Wir nennen einen Vektor der Dimension *n* eine *(n* × 1) **-Matrix** , das heißt, eine Tabelle mit nur einer Spalte und *n* Zeilen. Die Elemente darin Tabelle werden Koeffizienten oder Vektorcom genannt Komponenten.

MATHEMATISCHE ASPEKTE

Seien **x** = *(xi)* und **y** = *(yi)* für *i = 1, 2,...,n* zwei Vektoren der Dimension n.

Wir definieren:

- Die Summe zweier Vektoren, die Komponente für Komponente gebildet werden:

$$\mathbf{x} + \mathbf{y} = (xi) + (yi) = (xi + yi) .$$

- Das Produkt des Vektors **x** mit dem Skalar k:

$$k \cdot \mathbf{x} = (k \cdot \text{xi}),$$

wobei jede Komponente mit k multipliziert wird .

- Das **Skalarprodukt** von **x** und y:

$$\mathbf{x} \cdot \mathbf{y} = (xi) \cdot (yi) = \sum_{i=1}^{N} xi \cdot yi .$$

- Die **Norm** von x:

$$\mathbf{x} = ÿ\ \mathbf{x} \cdot \mathbf{x}$$

$$\mathbf{x} = \sum_{i=1}^{N} x2_{ich} .$$

Wir nennen einen Einheitsvektor einen Vektor der Norm (oder Länge) 1. Wir erhalten einen Einheitsvektor, indem wir einen Vektor durch seine Norm dividieren. So zum Beispiel

$$\frac{1}{\mathbf{x}} (xi)$$

ist ein Einheitsvektor mit derselben Richtung als x.

V

Machine Translated by Google

- Der i-te Vektor der kanonischen Basen in

  Der Raum mit *n* Dimensionen wird dadurch definiert, dass der Vektor der Dimension *n* eine „1“ auf Zeile *i* und „0“ für andere Komponenten hat nen.

- Ein Vektor ist eine **Matrix** einer bestimmten Dimension

  sion *(n* × 1). Es ist möglich, einen Vektor mit einer (m×n)Matrix gemäß den für Matri definierten Multiplikationsregeln zu multiplizieren ces.

DOMÄNEN UND EINSCHRÄNKUNGEN

Ein Vektor kann als eine Spalte von num betrachtet werden bers, daher der Begriff Vektorspalte, so wie wir den Begriff Vektorlinie verwenden, um die **Transponierte** eines Vektors zu bezeichnen.

Somit entspricht das **innere Produkt** zweier Vektoren einer Vektorzeile multipliziert mit einer Vektorspalte.

BEISPIELE

In einem dreidimensionalen Raum betrachten wir die Vektoren:

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \quad \text{und} \quad \mathbf{y} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

wobei **y** der zweite Vektor der kanonischen Basis ist.

– Die **Norm** von **x** ist gegeben durch:

$$\begin{aligned} \|\mathbf{x}\|^2 &= \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \\ &= [1\ 2\ 3] \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \\ &= 12 + 22 + 32 = 1 + 4 + 9 \\ &= 14 \quad , \end{aligned}$$

woraus $\|\mathbf{x}\| = \sqrt{14} \approx 3{,}74$.

– Auf die gleiche Weise erhalten wir $\|\mathbf{y}\| = 1$.

– Die Summe zweier Vektoren **x** und **y** ist gleich:

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1+0 \\ 2+1 \\ 3+0 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix}. \end{aligned}$$

– Die Multiplikation des Vektors **x** mit dem Skalarprodukt 3 ergibt:

$$\begin{aligned} 3 \cdot \mathbf{x} &= 3 \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \\ &= \begin{pmatrix} 3 \cdot 1 \\ 3 \cdot 2 \\ 3 \cdot 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix}. \end{aligned}$$

– Das **Skalarprodukt** von **x** und **y** ist gleich:

$$\begin{aligned} \mathbf{x} \cdot \mathbf{y} &= \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \\ &= [1\ 2\ 3] \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \\ &= 1 \cdot 0 + 2 \cdot 1 + 3 \cdot 0 = 2 \quad . \end{aligned}$$

WEITERLESEN

Matrix

Transponieren

Von Mises, Richard

Von Mises, Richard wurde 1883 in Lemberg im Österreichisch-Ungarischen Reich (heute Lemberg, Ukraine) geboren und starb 1953. Von

Machine Translated by Google

Mises studierte bis 1906 Maschinenbau an der Technischen Universität Wien.
Er wurde Assistent von Georg Hamel, Professor für Mechanik, in Brünn (heute Brünn, Tschechische Republik), wo er 1908 die Venia Legendi erhielt. Anschließend arbeitete er als außerordentlicher Professor für Angewandte Mathematik. Nach dem Ersten Weltkrieg wurde er 1920 Direktor des Instituts für Angewandte Mathematik in Berlin.

Von Mises ist Richard vor allem für seine Arbeiten zu den Grundlagen der Wahrscheinlichkeitsrechnung und Statistik bekannt, die 1991 rehabilitiert wurden die 1960er Jahre. Er gründete die angewandte Schule für Mathematik und schrieb seine ersten Arbeiten darüber Philosophischer Positivismus im Jahr 1939.

Ausgewählte Hauptwerke und Artikel von Mises, Richard:

1928 Wahrscheinlichkeit, Statistik und Wahrheit. Springer, Berlin Heidelberg New York.

1939 Wahrscheinlichkeit, Statistik und Wahrheit, trans. Neyman, J., Scholl, D. und Rabinowitsch, E. Hodge, Glasgow.

Machine Translated by Google

Gewichtetes arithmetisches Mittel

Das **gewichtete arithmetische Mittel** ist ein **Maß für die zentrale Tendenz** einer Reihe quantitativer Beobachtungen, wenn nicht alle Beobachter

ationen haben die gleiche Bedeutung.

Wir müssen jeder Beobachtung ein Gewicht zuweisen, abhängig von ihrer Bedeutung im Verhältnis zu anderen Beobachtungen.

Das gewichtete arithmetische Mittel entspricht der Summe der Beobachtungen multipliziert mit ihren Gewichtungen dividiert durch die Summe ihrer Gewichtungen.

GESCHICHTE Das gewichtete arithmetische Mittel wurde 1712 von Cotes, Roger, eingeführt. Sein Werk wurde 1722, sechs Jahre nach seinem Tod, veröffentlicht.

MATHEMATISCHE ASPEKTE Sei

x1, *x2,...,xn* eine Menge von *n* Größen oder *n* Beobachtungen relativ zu einer **quantitativen Variablen** *X* , der wir die Gewichte zuweisen

w1, *w2,...,wn.*

Das gewichtete arithmetische Mittel beträgt:

$$\bar{x} = \frac{\sum_{i=1}^{N} xi \cdot wi}{\sum_{i=1}^{N} wi}.$$

BEREICHE UND GRENZEN Das gewichtete arithmetische Mittel wird heute in der Wirtschaftswissenschaft verwendet, insbesondere in Verbraucher- und Produzentenpreisindizes usw.

BEISPIELE

Nehmen wir an, dass sich das Management während eines Kurses über das Unternehmen aus drei relativ unterschiedlichen Anforderungen unterschiedlicher Bedeutung zusammensetzte:

Einzelprojekt	30 %
Zwischenprüfung	20 %
Abschlussprüfung	50 %

Jeder Schüler erhält eine Note auf einer Skala von 1 bis 10. Ein Schüler erhält eine 8 für sein Indi

Einzelprojekt, eine 9 für die Zwischenprüfung und eine 4 für die Abschlussprüfung.

Seine Abschlussnote errechnet sich aus dem gewichteten arithmetischen Mittel:

$$\bar{x} = \frac{(8 \cdot 30) + (9 \cdot 20) + (4 \cdot 50) = 6{,}2 \,.\, 30}{+ 20 + 50}$$

In diesem Beispiel entsprechen die Gewichtungen der relativen Bedeutung der verschiedenen Anforderungen des Kurses.

WEITERLESEN

- **Arithmetisches Mittel**
- **Bedeuten**
- **Maß für die zentrale Tendenz**

Machine Translated by Google

VERWEISE

Cotes, R.: Aestimatio Errorum in Mixta Mathesi, per Variationes partium Trian guli plani et sphaerici. In: Smith, R. (Hrsg.) Opera Miscellania, Cambridge (1722)

Gewichtete kleinste Quadrate Methode

Die gewichtete Methode der kleinsten Quadrate wird verwendet, wenn die Fehlervarianz nicht konstant ist, d. h. wenn die folgende Hypothese der Methode der kleinsten Quadrate verletzt wird: Die Fehlervarianz ist für jede Beobachtung konstant (gleich dem unbekannten Wert ÿ2) . *i* (d. h. unabhängig vom Wert des betreffenden *xij)*. Anstatt also für jedes *i = 1,...,n* Var(ÿi) = ÿ2 zu haben, gilt:

$$\text{Var}(ÿi) = ÿ2wi \,,$$

wobei die *Gewichte wi* > 0 für jedes *i* = 1,...,n unterschiedlich sein können .

MATHEMATISCHE ASPEKTE

In der Matrixform haben wir das Modell

$$\mathbf{Y} = \mathbf{X}ÿ + ÿ \quad ,$$

Dabei ist **Y** der Vektor *(n* × 1) der Beobachtungen relativ zur **abhängigen Variablen** *(n* Beobachtungen), **ÿ** der Vektor *(p* × 1) der zu schätzenden Parameter und **ÿ** der Vektor *(n* × 1) der Fehler , Und

$$\mathbf{X} = \begin{matrix} ÿ \\ ÿ \\ ÿ \end{matrix} \begin{matrix} 1 \; X11 \; \ldots \; X1(pÿ1) \\ \vdots \quad \vdots \qquad\qquad \vdots \\ 1 \; Xn1 \; \ldots \; Xn(pÿ1) \end{matrix} \begin{matrix} ÿ \\ ÿ \\ ÿ \end{matrix}$$

ist die **Matrix** *(n* × p) des Plans bezüglich der unabhängigen Variablen. Darüber hinaus haben wir:

$$\text{Var}(ÿ) = ÿ2V \quad ,$$

Wo

$$\mathbf{V} = \begin{matrix} ÿ \\ ÿ \\ ÿ \\ ÿ \end{matrix} \begin{matrix} 1/w1 \; 0 \; \cdots \; 0 \; 1/w2 \\ \ldots \qquad\qquad\qquad \ldots \\ \vdots \qquad \vdots \qquad \ldots \qquad \vdots \\ 0 \qquad 0 \cdots 1/Woche \end{matrix} \begin{matrix} 0 \\ 0 \\ \vdots \\ ÿ \end{matrix} \begin{matrix} \\ ÿ \\ \\ \end{matrix} .$$

Angabe

$$Yw = \mathbf{WY} \quad ,$$

$$Xw = \mathbf{WX} \quad ,$$

$$ÿw = \mathbf{W}ÿ \quad ,$$

mit

$$\mathbf{W} = \begin{matrix} ÿ \\ ÿ \\ \\ ÿ \end{matrix} \begin{matrix} ÿw1 \; 0 \; 0 \qquad \cdots \\ ÿw2 \; \ldots \qquad — \\ \vdots \qquad \vdots \qquad \ldots \\ 0 \qquad 0 \qquad \ldots \; ÿwn \end{matrix} \begin{matrix} 0 \\ 0 \\ \vdots \\ — \end{matrix} \begin{matrix} \\ ÿ \\ \\ ÿ \end{matrix} ,$$

wie z. **B. WW** = Vÿ1, erhalten wir das Äquiv Alent-Modell:

$$Yw = Xwÿ + ÿw \,,$$

Wo

$$\text{Var}(ÿw) = \text{Var}(W ÿ) = W\text{Var}(ÿ)\,\mathbf{W} = ÿ2 \cdot \mathbf{WVW} = \mathbf{ÿ2In} \,,$$

wobei *In* die **Identitätsmatrix** der Dimension n ist . Da bei diesem neuen Modell die Fehlervarianz konstant ist, wird die Methode der kleinsten Quadrate verwendet. Der Vektor der Schätzer ist:

$$ÿ\hat{}\, w = \mathbf{X}\,\mathbf{w}\mathbf{X}\mathbf{w}^{ÿ1}\,\mathbf{X}\,\mathbf{w}\mathbf{Y}\mathbf{w} = \mathbf{XW}\,WXÿ1\,\mathbf{XW}\,\mathbf{WY} = \mathbf{X}\,Vÿ1X^{ÿ1}\,\mathbf{X}\,Vÿ1Y \,.$$

Die Vektoren der geschätzten Werte für *Yw* = **WY** wird:

$$Yw = \mathbf{Xw}ÿ\hat{}\, w = \mathbf{WX}\,\mathbf{X}\,Vÿ1X^{ÿ1}\,\mathbf{X}\,Vÿ1Y \,.$$

Machine Translated by Google

Wir können einen Vektor geschätzter Werte erhalten für **Y** = Wÿ1Yw Einstellung:

$$\begin{aligned}\mathbf{Y} &= \mathrm{W}\ddot{y}1\mathrm{Yw} \\ &= \mathrm{W}\ddot{y}1\mathrm{WX}\,\mathbf{X}\,\mathrm{V}\ddot{y}1\mathrm{X}\qquad \ddot{y}1\qquad \mathbf{X}\,\mathrm{V}\ddot{y}1\mathrm{Y} \\ &= \mathbf{X}\ddot{y}\hat{}\,w\,.\end{aligned}$$

Die Varianz ÿ2 wird geschätzt durch:

$$\begin{aligned}\hat{}^{2}_{w} &= \frac{Yw\,\ddot{y}\,Yw\,Yw\,\ddot{y}\,Yw}{n\,\ddot{y}\,p} \\ &= \frac{\mathbf{Y}\,\ddot{y}\,\mathbf{YWWY}\,\ddot{y}\,\mathbf{Y}}{n\,\ddot{y}\,p} \\ &= \frac{\sum_{i=1}^{N} wi\,(yi\,\ddot{y}yi)^{2}}{n\,\ddot{y}\,p}\,.\end{aligned}$$

Im Fall einer einfachen Regression ergibt sich aus einer Stichprobe von *n* Beobachtungen *(xi, yi)* und Gewichtungen *wi :*

$$\mathbf{X}\,\mathrm{V}\ddot{y}1\mathrm{X} = \begin{pmatrix} \ddot{y} & \sum wi & \sum wixi & \ddot{y} \\ \ddot{y} & \sum Wixi & \sum wix2 & \ddot{y} \end{pmatrix},$$

$$\mathrm{XV}\ddot{y}1\mathrm{Y} = \begin{pmatrix} \ddot{y} & \sum wiyi & \ddot{y} \\ \ddot{y} & \sum wixiyi & \ddot{y} \end{pmatrix},$$

und wenn wir also ÿ^ *w* setzen $= \begin{matrix} \ddot{y}\,0w \\ \ddot{y}\,1w \end{matrix}$, wir finden:

$$\ddot{y}\,0w = \frac{\sum wiyi \sum wix2 \;\ddot{y}n\; \sum Wixi \sum wixiyi}{\sum wi \sum wix2 \;-\; \sum Wixi},$$

$$\ddot{y}\,1w = \frac{\sum wi \sum wixiyi \;\ddot{y}\; \sum Wixi \sum wiyi}{\sum wi \sum wix2 \;-\; \sum Wixi}.$$

Unter Berücksichtigung der überlegten Gewichtung bedeutet:

$$\bar{x}w = \frac{\sum wixi}{\sum wi}, \qquad \bar{y}w = \frac{\sum Wiyi}{\sum wi},$$

wir haben die Schätzer:

$$\ddot{y}\,1w = \frac{\sum wixiyi \;\ddot{y}\; \sum wixwyw}{\sum wix2 \;-\; \sum wix2_{w}} = \frac{\sum wi\,(xi\,\ddot{y}\,\bar{x}w)\,yi\,\ddot{y}\,\bar{y}w}{\sum wi\,(xi\,\ddot{y}\,\bar{x}w)^{2}}$$

$$\ddot{y}\,0w = \bar{y}w\,\ddot{y}\,\ddot{y}\,1w\bar{x}w$$

die den Kleinste-Quadrate-Schätzern stark ähneln, mit der Ausnahme, dass wir den Beobachtungen mit einem großen *wi* mehr Gewicht beimessen (d. h. Beobachtungen, bei denen die Fehler im ursprünglichen Modell eine geringere Varianz aufweisen). Die geschätzten Werte ergeben sich aus:

$$yi = \ddot{y}\,0w + \ddot{y}\,1wxi = \bar{y}w + \ddot{y}\,1w\,(xi\,\ddot{y}\,\bar{x}w)$$

und die Varianz ÿ2 wird geschätzt durch:

$$\begin{aligned}s^{2}_{w} &= \frac{\sum wi\,(yi\,\ddot{y}yi)^{2}}{n\,\ddot{y}\,2} \\ &= \frac{\sum wi\,\mathrm{yi}\ddot{y}\mathrm{yw}^{2} \;2\,\ddot{y}\ddot{y}\,1w\; \sum wi(xi\ddot{y}xw)^{2}}{n\,\ddot{y}\,2}.\end{aligned}$$

Beispiele für in der einfachen Regression gewählte Gewichtungen *wi* sind *wi = 1/xi* in Fällen, in denen die Varianz von *ÿi* proportional zu xi ist , oder *wi = 1/x2* , wenn die Varianz von *ÿi* proportional zu x2 ist

W

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Der größte Nachteil der gewichteten
Methode der kleinsten Quadrate, die viele Menschen verwenden
Lieber nichts davon wissen, ist wohl das
Tatsache, dass es auf der Annahme basiert, dass
das Gewicht ist genau bekannt. Das ist fast
Dies ist in der Praxis jedoch nie der Fall und stattdessen
werden die geschätzten Gewichte verwendet. Im Allgemeinen ist es so
Es ist schwierig, die Wirkung der Verwendung von zu beurteilen
geschätzte Gewichte, aber die Erfahrung lässt darauf schließen
dass die Ergebnisse der meisten Regressionsanalysen
reagieren nicht sehr empfindlich auf die verwendeten Gewichte
Vorteile einer gewichteten Analyse sind oft
in großem Maßstab erhalten, wenn auch nicht immer
komplett, mit den ungefähren Gewichten. Es ist
Es ist wichtig, sich dieses Problems bewusst zu sein und
es zu vermeiden und nur Gewichte zu verwenden, die
relativ zueinander präzise abgeschätzt werden können.
Die gewichtete Regression der kleinsten Quadrate, as
Andere Methoden der kleinsten Quadrate sind ebenfalls empfindlich
zu Ausreißern.

BEISPIELE

Betrachten Sie einen Datensatz mit 10 Beobachtungen,
wobei die erklärende Variable *X* die Werte *Xi* = *i* mit *i* =
1, 2, ..., 10 annimmt . Variable *Y*
wird mit dem Modell generiert:

$$yi = 3 + 2xi + \ddot{y}i ,$$

wobei die $\ddot{y}i$ normalverteilt sind
$E(\ddot{y}i) = 0$ und $Var(\ddot{y}i) = (0,2xi)\ 2$. Somit
die Fehlervarianz der ersten Beobachtung
entspricht $[(1)\ (0,2)]^2 = 0,04$, genauso
denen der zehnten Beobachtung entsprechen
$[(10)\ (0,2)]^2 = 4$. Die so generierten Daten
sind in der folgenden Tabelle dargestellt:

Vom Modell generierte Werte xi und yi yi = 3 +
2xi + ÿi

ich	xi	yi
1	1	4,90
2	2	6.55
3	3	8.67
4	4	12.59
5	5	17.38
6	6	13.81
7	7	14.60
8	8	32.46
9	9	18.73
10	10	20.27

Zuerst führen wir eine ungewichtete Regression durch.
Wir erhalten die Regressionsgleichung:

$$yi = 3,49 + 2,09xi ,$$

$\ddot{y}\ 0 = 3,49$ mit $tc = 0,99$ $\ddot{y}\ 1 = 2,09$ mit ,

$tc = 3,69$.

Varianzanalyse

Quelle von variation	Abschlüsse von kostenlos dom	Die Summe von Quadrate	Bedeutung von Quadrate	F
Regression	1	360,68	360,68	13,59
Resi Dual	8	212,39	26,55	
Gesamt	9	573.07		

Der R2 dieser Regression beträgt 62,94 %. Anfangs
Auf den ersten Blick scheinen die Ergebnisse gut zu
sein. Der geschätzte Koeffizient, der der erklärenden
Variablen zugeordnet ist, ist eindeutig signifikant. Das
Das Ergebnis ist überhaupt nicht überraschend, denn das
Tatsache, dass die Varianz der Fehler nicht con

Die Konstante hat selten Einfluss auf die geschätzten
Koeffizienten. Das mittlere Quadrat der Residuen lässt
sich nur schwer erklären, da es kein eindeutiges Ergebnis gibt
Varianz zu schätzen. Darüber hinaus der Standard
Fehler und die Längen der Konfidenzintervalle bei 95 %
für die geschätzte Bedingung

Die Mittelwerte sind für alle Beobachtungen relativ
konstant. Bei der ungewichteten Analyse ist dies nicht der Fall

Machine Translated by Google

Berücksichtigen Sie die beobachteten Ungleichheiten zwischen den Varianzen der Fehler.

Wir führen nun eine gewichtete Regression durch, wobei wir als Gewichte die Werte *1/x2 verwenden.* Die Gewichte sind bekannt, da sie proportional zu den tatsächlichen Varianzen sind, die (0,2x) 2 entsprechen. Die Ergebnisse sind die folgenden:

$$yi = 2{,}53 + 2{,}28xi \ ,$$

$$\ddot{y}\ 0 = 2{,}53 \text{ mit } tc = 3{,}07 \quad ,$$

$$\ddot{y}\ 1 = 2{,}28 \text{ mit } tc = 7{,}04 \ .$$

Varianzanalyse

Quelle von Variation	Abschlüsse von kostenlos dom	Summe Quadrate	Bedeutung von Quadrate	F
Regression	1	23.29	23.29	49,6
Rest	8	3,75	0,47	
Gesamt	9	27.05		

Der R2 dieser Regression beträgt 86,11 %. Auf dieser Grundlage können wir folgende Anmerkungen machen:

- Alle Zahlen, die sich auf die Quadratsumme der abhängigen Variablen beziehen, werden von den Gewichten beeinflusst und sind nicht mit denen vergleichbar, die durch nichtgewichtete Regression erhalten werden.
- Der Fisher-Test und der R2 können zwischen zwei Modellen verglichen werden. In unserem Beispiel sind zwei Werte größer als die nicht gewichtete Regression, aber das ist nicht immer der Fall.
- Die zusammengefassten Koeffizienten sind rela denen der ungewichteten Regression sehr nahe kommen. Das ist oft der Fall, aber keine allgemeine Regel.
- Die Standardfehler und das Konfidenzintervall Werte für die bedingten Mittel spiegeln die wider Tatsache, dass die Präzision der Schätzungen nimmt mit steigenden Werten von x ab.

Dieses Ergebnis ist der erste Grund, das zu verwenden gewichtete Regression.

WEITERLESEN

- Verallgemeinerte lineare Regression
- Kleinsten Quadrate
- Regressionsanalyse

REFERENZEN

Seber, GAF (1977) Lineare Regressionsanalyse. Wiley, New York

Wilcoxon, Frank

Wilcoxon, Frank (1892–1965) erhielt 1921 seinen Master-Abschluss in Chemie an der Rutgers University und promovierte 1924 in Physik an der Cornell University, wo er von 1924 bis 1925 auch als Postdoktorand tätig war.

Von 1928 bis 1929 war er bei der Nichols Copper Company in Maspeth in Queens, New York, beschäftigt und wurde dann vom Boyce Thompson Institute for Plant Research als Leiter einer Gruppe angestellt, die die Wirkung von Insektiziden und Fungiziden untersuchte. Er bleibt bis zum Beginn des Studiums am Institut

Zweiter Weltkrieg; Zu dieser Zeit und für die nächsten zwei Jahre wechselte er zur Atlas Powder Company in Ohio, wo er das Kontrolllabor leitete.

Im Jahr 1943 wurde Frank Wilcoxon zum Leiter der Gruppe im Labor der American Cyanamid Company (in Stamford, CT) ernannt, die Insektizide und Fungizide untersuchte. Anschließend entwickelte er eine interessante Statistik: 1950 wurde er in die Abteilung Lederle Laboratories versetzt

sion von Cyanamid in Pearl River, NY, wo er eine statistische Beratungsgruppe aufbaute.

Von 1960 bis zu seinem Tod 1965 lehrte er

Machine Translated by Google

Angewandte Statistik im Department of Statistics für naturwissenschaftliche Hauptfächer an der Florida State University Universität.

Ausgewählte Hauptwerke und Artikel von Frank Wilcoxon:

1945 Einzelvergleiche nach Rangfolge Methoden. Biometrie 1, 80–83.

1957 Einige schnelle näherungsweise statistische Verfahren. Amerikanisches Cyanamid, Stamford Research Laboratories, Stamford, CT.

Von Wilcoxon signierter Tisch

Die von Wilcoxon signierte Tabelle gibt theoretische Informationen Werte **der Statistik** *T* des **Wilcoxon-Signaturen** Der **Test** wurde auf zwei gepaarte Proben angewendet **Hypothese ,** dass zwei Populationen dem folgen gleiche Verteilung.

GESCHICHTE

Owen, DB (1962) veröffentlichte einen Wilcoxon Vorzeichentabelle für *n* ÿ 20 (wobei *n* die Anzahl der gepaarten Beobachtungen ist) für den von Wilcoxon, F. entwickelten **Wilcox-On-Vorzeichentest.** (1945).

Die allgemein verwendete Tabelle ist die von Harter, HL und Owen, DB (1970) gegründet für *n* ÿ 50.

MATHEMATISCHE ASPEKTE

Betrachten Sie eine Menge von *n* Beobachtungspaaren *((X1, Y1),...,(Xn,* Yn)). Wir berechnen die absolute Differenzen /*Xi* ÿ *Yi*/ zu dem wir Weisen Sie einen Rang zu (das heißt, wir ordnen Rang 1 zu der kleinste Wert, Rang 2 zum nächsthöheren Wert usw. bis zum Rang *n* für den Großen Schätzwert). Dann geben wir diesem Rang das Zeichen entsprechend der DifferenzXiÿYi. Daher Wir berechnen die Summe der positiven Ränge:

$$T = \sum_{i=1}^{k} Ri ,$$

wobei *k* die Anzahl der darstellenden Paare ist eine positive Differenz *(Xi* ÿ *Yi* > 0). Für unterschiedliche Werte von *m* und ÿ (Signifikanzniveau) die Wilcoxon-Vorzeichentabelle gibt die theoretischen Werte der **Statistik** *T* von an Der **Wilcoxon-Signed-Test** basiert auf der **Hypothese ,** dass die beiden zugrunde liegenden Populationen identisch sind identisch verteilt.

DOMÄNEN UND EINSCHRÄNKUNGEN

Die vorzeichenbehaftete Wilcoxon-Tabelle wird in nichtparametrischen Tests verwendet, die Rang und insbesondere verwenden im **Wilcoxon-signierten Test.**

BEISPIELE

Unten finden Sie einen Auszug aus einem von Wilcoxon unterzeichneten Dokument Tabelle *m* geht von 10 bis 15 und ÿ = 0,05 bis ÿ = 0,975:

M	ÿ = 005.	ÿ = 0975
10	11	46
11	14	55
12	18	64
3	22	73
14	26	83
15	31	94

Hier finden Sie ein Beispiel dafür, wie die Wilcoxon unterschrieben hat Tabelle verwendet wird, siehe **Wilcoxon-signierte Tabelle.**

WEITERLESEN

Statistische Tabelle

Von Wilcoxon unterzeichneter Test

Machine Translated by Google

VERWEISE

Harter, HL, Owen, DB: Selected Tables in Mathematical Statistics, vol. 1.Markham, Chicago (5.7 Anhang) (1970)

Owen, DB: Handbuch der statistischen Tabellen. (3.3, 5.7, 5.8, 5.10). Addison-Wesley, Reading, MA (1962)

Wilcoxon, F.: Einzelvergleiche nach Ranking-Methoden. Biometrics 1, 80–83 (1945)(5.1, 5.7).

Von Wilcoxon signierter Test

Der vorzeichenbehaftete Wilcoxon-Test ist ein **nichtparametrischer** Test. Als **vorzeichenbehafteter Test** dient er dazu, Unterschiede zwischen zwei Populationen festzustellen. Es gibt an, ob die Differenz zwischen zwei Beobachtungspaaren **positiv** oder negativ ist, wobei die Amplitude dieser Differenz berücksichtigt wird enz.

GESCHICHTE Das Wilcoxon-Zeichen wurde 1945 von Wilcoxon, F. erstellt. 1949 schlug er eine andere Methode zur Behandlung der Nulldifferenzen vor das zwischen zwei Beobachtungen der betrachteten Proben auftreten kann.

MATHEMATISCHE ASPEKTE

Seien (X1, *X2,...,Xn)* und (Y1, *Y2,...,Yn)* zwei Stichproben der Größe n. Wir betrachten *n* Beobachtungspaare (x1, y1),(x2, *y2),...,(xn,* yn). Wir bezeichnen mit */di/* der Unterschied im absoluten Wert zwischen *xi* und *yi:*

$$|di| = |yi \ddot{y} xi| , \qquad \text{für } i = 1, 2,...,n.$$

Wir subtrahieren von der Dimension alle Beobachtungspaare, die den Unterschied nicht darstellen *(di* = 0). Mit *m* bezeichnen wir die Anzahl der verbleibenden Paare. Wir weisen dann zu

Die Beobachtungspaare haben einen Rang von 1 bis *m ,* abhängig von der Größe von */di/,* das heißt, wir geben dem kleinsten Wert von */di/* den Rang 1 und ordne *m* dem größten Wert von */di/ zu.* Den so definierten Rang nennen wir *Ri, i = 1,...,m .* Wenn viele Variablenpaare die gleiche absolute Differenz */di/ darstellen,* weisen wir ihnen den mittleren Rang zu. Wenn mittlere Ränge vorhanden sind, wird die **Statistik** des Tests anhand des folgenden Ausdrucks berechnet:

$$T1 = \frac{\sum_{i=1}^{M} Ri}{\sum_{i=1}^{M} R2} ,$$

Dabei ist Ri der Rang des Paares (xi, yi) ohne das Vorzeichen der Differenz *di.*

Wenn es keine mittleren Ränge gibt, ist es korrekter, die Summe der Ränge zu verwenden, die der positiven Differenz zugeordnet sind:

$$T = \sum_{(\text{für } i \text{ wie } di > 0)} Ri .$$

Hypothesen

Die zugrunde liegenden Hypothesen im Wilcoxon-Vorzeichentest basieren auf *dk,* dem **Median** von *di.* Je nachdem, ob es sich um einen **zweiseitigen** oder einen **einseitigen Test handelt,** haben wir Folgendes Fälle:

A: Zweiseitiger Test:

H0 : *dk* = 0

H1 : *dk* = 0

B: Einseitiger Test:

H0 : *dk* ÿ 0

H1 : *dk* > 0

C: Einseitiger Test:

H0 : *dk* ÿ 0

H1 : *dk* < 0

W

Machine Translated by Google

Im Fall A stellt die Hypothese die Situation dar, in der es keine Unterschiede zwischen ihnen gibt die beiden Populationen.

Im Fall B stellen wir die Hypothese auf, dass die Werte der Grundgesamtheit von X tendenziell größer sind als die der Grundgesamtheit von Y.

Im Fall C stellen wir die **gegenteilige Hypothese** auf , dass die Werte der **Grundgesamtheit** von *X* tendenziell kleiner sind als die Werte der Grundgesamtheit von Y.

Entscheidungsregeln Wenn *T* verwendet wird und *m* < 50 ist, wird die **Wilcox-Tabelle mit Vorzeichen** zum Testen der **Nullhypothese** H0 verwendet .
Die **Zufallsvariable** *T* ist eine lineare Kombination Nation von *m* unabhängigen Bernoulli-Zufallsvariablen (aber nicht identisch verteilt). Als **Mittelwert** und **Standardabweichung** gilt :

$$ÿ = \frac{m\,(m+1)\,4}{} \quad \text{Und}$$

$$ÿ = \sqrt{\frac{m\,(m+1)\,(2m+1)\,24}{}}\ .$$

Wir leiten die **Zufallsvariable** Z ab:

$$Z = \frac{T ÿ\ ÿ}{ÿ}\ .$$

Wir bezeichnen mit wÿ den Wert der Wilcoxon-Signaturtabelle mit den Parametern m und ÿ (wobei ÿ das **Signifikanzniveau ist).**
Wenn m > 20, dann kann wÿ mit Hilfe der entsprechenden Werte der **Normalverteilung angenähert werden:**

$$wÿ = \frac{m\,(m+1)\,4}{} + zÿ \sqrt{\frac{m\,(m+1)\,(2m+1)\,24}{}}\ ,$$

wobei zÿ die Werte in der **Normaltabelle** auf Ebene ÿ sind.

Wenn T1 verwendet wird, basieren wir unsere Berechnung direkt auf der Normaltabelle, um die Null zu testen Hypothese H0. Die Entscheidungsregeln sind je nach Hypothese unterschiedlich. Wir haben die Entscheidungsregeln *A, B* und *C* relativ zu den vorherigen Fällen *A, B* und C. Wenn T1 verwendet wird, sollten wir in den folgenden Regeln *T* durch T1 und wÿ durch zÿ ersetzen.

Fall A

Wir lehnen H0 auf dem Signifikanzniveau ÿ ab, wenn *T* größer als w1ÿÿ/2 oder kleiner als wÿ/2 ist, d. h. wenn

$$T < wÿ/2 \text{ oder } T > w1ÿÿ/2\ .$$

Fall B

Wir lehnen H0 auf dem Signifikanzniveau ÿ ab, wenn *T* größer als w1ÿÿ ist, d. h. wenn

$$T > w1ÿÿ\ .$$

Fall C

Wir lehnen H0 auf dem **Signifikanzniveau** ÿ ab, wenn *T* kleiner als wÿ ist, d. h. wenn

$$T < wÿ\ .$$

BEREICHE UND EINSCHRÄNKUNGEN

Um den Wilcoxon-Vorzeichentest verwenden zu können, müssen die folgenden Kriterien erfüllt sein: 1. Die Verteilung von *di* muss symmetrisch sein. 2. *di* muss unabhängig sein.
3. *di* muss in wahren Werten gemessen werden.
Als Test des **Medians wird auch der Wilcoxon-Signed-Test verwendet.** Die Daten X1, *X2,...,Xn* bestehen aus einer einzelnen **Stichprobe** der Dimension n.
Die den vorherigen Fällen *A, B* und *C* entsprechenden Hypothesen lauten wie folgt:

A: Zweiseitige Fälle:

H0: Der Median von *X* entspricht einem angenommenen **Wert** Md

H1: Der Median von *X* ist nicht gleich Md

Machine Translated by Google

B: One-Tail-Fall:

H0: Der Median von *X* ist ÿ Md

H1: Der Median von *X* ist < Md

C: One-Tail-Fall:

H0: Der Median von *X* ist ÿ Md

H1: Der Median von *X* ist > Md

Beachten Sie, dass „Median“ durch ersetzt werden könnte **„Mittelwert“ ,** weil wir die Hypothese der Symmetrie der Verteilung von X akzeptiert haben.
Was die Berechnungen betrifft, bilden wir ab Paare (X1, Md),(X2, *Md),...,(Xn,* Md) und Behandeln Sie sie auf die gleiche Weise wie zuvor. Der Rest des Tests ist identisch.

BEISPIELE

Wir führen Tests an acht Zwillingspaaren durch. Der Ziel ist es herauszufinden, ob der Erstgeborene intelligenter ist als der Zweitgeborene.
Die Daten werden in der folgenden Tabelle dargestellt: Je höher die Punktzahl, desto besser der Test Ergebnisse.

Ein Paar Zwillinge	Erste geboren Xi	Zweite geboren Yi	Rang von di	di	Ri
1	90	88	ÿ2 2 ÿ2		
2	75	79	4 4		4
3	99	98	ÿ1 1		ÿ1
4	60	66	6 5		5
5	72	64	ÿ8 6 ÿ6		
6	83	83	0 –		–
7	86	86	0 –		–
8	92	95	3 3		3

Da die Paare 6 und 7 keinen Unterschied darstellen *(d6* = d7 = 0), die Anzahl der verbleibenden Paare gleich *m* = 6. Da die *di* alle unterschiedlich sind, müssen die mittleren Ränge nicht berechnet werden, und wir verwenden die Summe der positiven Ränge:

$$T = \sum Ri ,$$

(denn *ich* so als *di* > 0)

Dabei ist *Ri* der Rang des Paares (xi, *yi)* , das ein positives *Di* darstellt . Somit haben wir:

$$T = 4 + 5 + 3 = 12 \quad .$$

Die Hypothesen lauten wie folgt:

H0: Es gibt keinen Unterschied im Niveau Intelligenz zwischen dem ersten und dem zweiten Zwilling *(dm* = 0)

H1: Es gibt einen Unterschied *(dm* = 0)

Diese Hypothesen entsprechen Fall *A*
und die Entscheidungsregel sind die folgenden:
Lehnen Sie H0 auf der Ebene ÿ ab, wenn *T* < wÿ/2 oder *T* > w1ÿÿ/2.
Wenn wir das **Signifikanzniveau** ÿ = wählen 0,05 erhalten wir aus der **Wilcoxon-Signatur Tabelle** für *m* = 6:

$$w1ÿÿ/2 = 20 \text{ und } wÿ/2 = 1 \quad ,$$

wobei T weder größer als 20 noch kleiner ist als 1, und wir können die **Nullhypothese** H0 nicht ablehnen.
Nun nehmen wir an, dass die Ergebnisse für Paare 7 und 8 haben sich mit der neuen Hypothese geändert in der folgenden Tabelle:

Ein Paar Zwillinge	Erste geboren Xi	Zweite geboren Yi	Rang von di	di	Ri
1	90	88	ÿ2 4 ÿ4		
2	75	79	4 5		5
3	99	98	ÿ1 2 ÿ2		
4	60	66	6 6		6
5	72	64	ÿ8 7 ÿ7		
6	83	83	0 –		–
7	86	87	1 2		2
8	92	91	ÿ1 2 ÿ2		

Nun stellt Paar 6 keinen Unterschied dar *(d6* = 0); also ist *m* gleich 7.
Die Paare 3, 7 und 8 repräsentieren dasselbe Absolute Differenz: |d3| = |d7| = |d8| = 1. Wir

W

Machine Translated by Google

Weisen Sie den Paaren, die gleich sind, einen mittleren Rang zu
Rang des ex aequo, zu dem wir die Hälfte hinzufügen
Anzahl ex aequo minus eins. Daher:

$$\text{mittlerer Rang} = \text{Rang von } |d3| + \frac{1}{2}(3 - 1) = 2 .$$

In diesem Fall verwenden wir den statistischen Test T1 ,
berechnet nach:

$$T1 = \frac{\sum_{i=1}^{M} Ri}{\sum_{i=1}^{M} R2} = \frac{2}{11.75} = 0,17 .$$

Die Entscheidungsregel lautet wie folgt:
H0 auf Ebene ablehnen , wenn T1 < z/2 oder T1 > z1/2.
Wenn wir das Niveau = 0,05 wählen , dann ist das
Werte der **Normaltabelle** sind:

$$z1-\alpha/2 = 1,96 \text{ und } z\alpha/2 = 1,96 .$$

Folglich ist T1 nicht größer als
1,96 nicht kleiner als 1,96.

Wir erhalten das gleiche Ergebnis wie im vorherigen
Fall und lehnen Sie die NullhypotheseH0 nicht ab.

WEITERLESEN

Hypothesentest
Nichtparametrischer Test
Von Wilcoxon signierter Tisch

VERWEISE

Wilcoxon, F.: Einzelvergleiche von
Ranking-Methoden. Biometrie 1, 80–83
(1945)(5.1, 5.7).

Wilcoxon, F.: Einige schnelle ungefähre statistische
Verfahren. Amerikanisches Cyanamid,
Stamford Research Laboratories, Stamford, CT
(1957)

Wilcoxon-Tisch

Die Wilcoxon-Tabelle gibt die theoretischen Werte der **Statistik** *T* des unten aufgeführten **Wilcoxon-Tests** an die **Hypothese ,** dass es keinen Unterschied gibt zwischen der Verteilung von zwei verglichen Populationen.

GESCHICHTE

Die kritischen Werte für *N* 20 (wobei *N* ist (die Gesamtzahl der Beobachtungen) wurden von Wilcoxon, F. (1947) berechnet.

MATHEMATISCHE ASPEKTE

Sei (X1, *X2,...,Xn)* eine **Stichprobe** der Dimension *n* , die aus **Grundgesamtheit** 1 stammt und (Y1, *Y2,...,Ym)* eine Stichprobe der Dimension *m* stammt aus Population 2.
Somit erhalten wir *N* = n+m Beobachtungen
Wir möchten die Klassen in aufsteigender Reihenfolge anordnen, ohne ihre Zugehörigkeit zu berücksichtigen die Beispiele. Wir weisen jedem Wert einen Rang zu: Rang 1 zum kleinsten Wert, Rang 2 zum nächsten Wert usw. bis Rang *N* zum größten Wert Wert.
Im Folgenden definieren wir den statistischen Test *T* Art und Weise:

$$T = \sum_{i=1}^{N} R(Xi) ,$$

wobei R(Xi) der der Beobachtung zugewiesene Rang istXi, *i = 1, 2,...,n* relativ zur Menge von zwei Proben *(X1,...,Xn)* und *(Y1,...,Ym).*
Für verschiedene Werte von n, m und
(Signifikanzniveau) liefert die Wilcoxon-Tabelle die theoretischen Werte der **Statistik** *T* des **Wilcoxon-Test, unter der Hypothese, dass die** zwei zugrunde liegende Populationen sind identisch verteilt.

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Die Wilcoxon-Tabelle wird in nichtparametrischen Tests verwendet, die den Rang verwenden, insbesondere im **Wilcoxon-Test.**

BEISPIELE

Hier ist ein Auszug aus einer Wilcoxon-Tabelle für n = 14, m = 11 und ÿ = 0,05 gleich 0,95:

N	M	ÿ	kritisch
14	11	0,05	152
14	11	0,95	212

Ein Beispiel für die Verwendung des Wilcoxon Tabelle siehe **Wilcoxon-Test.**

WEITERLESEN

- **Statistische Tabelle**
- **Wilcoxon-Test**

VERWEISE

Wilcoxon, F.: Wahrscheinlichkeitstabellen für Einzelvergleiche nach Ranking-Methoden. Biometrik 3, 119–122 (1947)

Wilcoxon-Test

Der Wilcoxon-Test ist ein **nichtparametrischer Test.** Es wird verwendet, wenn wir zwei Proben haben aus zwei Populationen. Ziel ist die Verifizierung wenn sich aufgrund der Zufallsstichproben aus diesen Populationen ein Unterschied zwischen den Populationen ergibt.

GESCHICHTE

Der Wilcoxontest ist nach seinem Autor benannt und wurde 1945 eingeführt.

Siehe auch **Mann-Whitney-Test.**

MATHEMATISCHE ASPEKTE

Sei (X1, *X2,...,Xn)* eine **Stichprobe** der Dimension n , die aus einer **Grundgesamtheit** 1 stammt, und sei (Y1, *Y2,...,Ym)* sei eine Stichprobe der Dimension m aus einer Bevölkerung stammend 2.
Somit erhalten wir N = n+m Beobachtungen

Wir werden in aufsteigender Reihenfolge klassifizieren, ohne ihre Zugehörigkeit zu den Stichproben zu berücksichtigen. Dann weisen wir dem kleinsten **Schätzwert einen Rang von 1 zu, dem** nächsthöheren Wert einen Ran und so weiter bis zum Rang N, der zugeordnet ist der höchste Wert. Wenn viele Beobachtungen vorliegen genau den gleichen Wert, dann werden wir ihn zuweisen ein mittlerer Rang. Mit *R(Xi)* bezeichnen wir den Rang Xi zugeordnet , $i = 1,...,n$.

Die Statistik W des Tests ist definiert durch:

$$W = \sum_{i=1}^{N} R(Xi) .$$

Wenn die Größe der Stichprobe groß ist *(N* ÿ 12), dann verwenden wir nach Gibbons, JD (1971), eine Näherung unter Verwendung der Statistik W1, die einer **Standardnormalverteilung** $N(0, 1)$ folgen soll :

$$W1 = \frac{W \text{ ÿ } \ddot{y}}{\ddot{y}} ,$$

wobei ÿ und ÿ jeweils der **Mittelwert** sind und die **Standardabweichung** des **Zufalls Variable** W:

$$\ddot{y} = \frac{n(N+1)}{2}$$

Und

$$\ddot{y} = \sqrt{\frac{mn(N+1)}{12}} .$$

Wenn wir die Näherung für das Große verwenden Proben W1, und wenn wir die Tatsache nutzen, dass dort Sind die Ränge ex aequo unter den N Beobachtungen, ersetzen wir die Standardabweichung durch:

$$\ddot{y} = \sqrt{\frac{mn}{12}\left(N+1 \text{ ÿ } \frac{\sum_{j=1}^{G} tj\,t_J^2 \text{ ÿ } 1}{N(N \text{ ÿ } 1)}\right)} ,$$

Machine Translated by Google

wobei *g* die Anzahl der Gruppen gleicher (Unentschieden) ist Ränge undtj die Dimension der Gruppej. (Wenn dort Es gibt keine gleichen (gleichen) Ränge, die Beobachtungen sind es gesehen als viele Gruppen der Dimension 1. Folglich gilt $g = N$ und $tj = 1$ fürj = 1,...,N, und ÿ reduziert sich auf $\frac{mn(N+1)}{12}$.)

Hypothesen

Auf dieser Grundlage kann der Wilcoxon-Test durchgeführt werden eines **zweiseitigen Tests** oder **eines einseitigen Tests,** je nach Art der **Hypothese:**

A: Zweiseitiger Test:

$$H0 : P(X < Y) = \frac{1}{2}$$

$$H1 : P(X < Y) = \frac{1}{2}$$

B: One-Tail-Test:

$$H0 : P(X < Y) \ddot{y} \frac{1}{2}$$

$$H1 : P(X < Y) > \frac{1}{2}$$

C: Einseitiger Test:

$$H0 : P(X < Y) \ddot{y} \frac{1}{2}$$

$$H1 : P(X < Y) < \frac{1}{2}$$

Fall A drückt die Hypothese aus, dass dies der Fall ist kein Unterschied zwischen zwei Populationen.
Fall B stellt die Hypothese dar, dass Population 1 (von der wir die Stichprobe genommen haben).
X) nimmt im Allgemeinen größere Werte an als dies der Fall ist Population 2 (von wo wir die Probe entnommen haben von Y).
Fall C drückt die gegenteilige Hypothese aus:
Das heißt, dass die Werte von Population 1 tendenziell kleiner sind als die von Population 2.

Entscheidungsregeln

Fall A

Wir lehnen die **Nullhypothese** H0 auf dem **Signifikanzniveau ÿ** ab , wenn *W* kleiner als der Wert ist ue der **Wilcoxon-Tabelle** mit Parametern n, m und $\frac{\ddot{y}}{2}$ bezeichnet mit *tn,m,ÿ/2* orifW größer ist dann der Wert in der Tabelle für n, m und 1ÿÿ $\frac{\ddot{y}}{2}$, durch tn,m,1ÿÿ/2, das heißt, wenn

$$W < tn,m,\ddot{y}/2 \text{ oder } W > tn,m,1\ddot{y}\ddot{y}/2 .$$

Wenn der Test die Statistik W1 verwendet, erfolgt ein Vergleich gemacht mit der **normalen Tabelle:**

$$W1 < z\ddot{y}/2 \text{ oder } W1 > z1\ddot{y}\ddot{y}/2 ,$$

mit zÿ/2 und z1ÿÿ/2 die Werte der Normaltabelle mit jeweils den Parametern $\frac{\ddot{y}}{2}$ und 1 ÿ $\frac{\ddot{y}}{2}$.

Fall B

Wir lehnen die Nullhypothese H0 auf dem Signifikanzniveau ÿ ab, wenn *W* kleiner als der Wert ist der Wilcoxon-Tabelle mit den Parametern n, m, und ÿ, bezeichnet mit *tn,m,ÿ,* also wenn

$$W < tn,m,\ddot{y} ,$$

und für den Fall, dass die Statistik W1 verwendet wird:

$$W1 < z\ddot{y} ,$$

wobei zÿ der Wert der Normaltabelle ist mit Parameter ÿ.

Fall C

Wir lehnen die Nullhypothese H0 auf dem Signifikanzniveau ÿ ab, wenn *W* größer als der Wert ist in der Wilcoxon-Tabelle mit den Parametern n, m, und 1 ÿ ÿ, bezeichnet mit *tn,m,1ÿÿ,* also wenn

$$W > tn,m,1\ddot{y}\ddot{y} ,$$

und mit Statistik W1:

$$W1 > z1\ddot{y}\ddot{y} ,$$

wobei z1ÿÿ der Wert der Normaltabelle ist mit Parameter ÿ.

Machine Translated by Google

DOMÄNEN UND EINSCHRÄNKUNGEN

Um den Wilcoxon-Test nutzen zu können, müssen folgende Kriterien erfüllt sein:

1. Die beiden Stichproben müssen zufällig sein
   Proben aus ihren jeweiligen **Populationen.**
2. Zusätzlich zur **Unabhängigkeit** im Inneren
   Bei jeder Probe muss eine gegenseitige Unabhängigkeit
   zwischen den beiden Proben bestehen.
3. Der Maßstab muss mindestens ordi sein
   nal.

BEISPIELE

In einer Klasse zählen wir 25 Schüler: 14 Jungen und 11 Mädchen. Wir testen sie auf mentale Berechnung um zu sehen, ob die Jungen in dieser Klasse eine Tendenz haben besser sein als die Mädchen.

Die Daten sind in der folgenden Tabelle aufgeführt, wobei die höchsten Ergebnisse dem besten Test entsprechen Ergebnisse.

Jungen (Xi)		Mädchen (Yi)	
19.8	17.5	17.7	23.6
12.3	17.9	7.1	11.1
10.6	21.1	21.0	20.3
11.3	16.4	10.6	15.6
14.0	7.7	13.3	
9.2	15.2	8.6	
15.6	16.0	14.1	

Wir ordnen die Ergebnisse in aufsteigender Reihenfolge und einen Rang zuweisen:

Partituren	Probe	Rang	R (Xi)
7.1	*Y*	1	–
7.7	*X*	2	2
8.6	*Y*	3	–
9.2	*X*	4	4
10.6	*Y*	5.5	–
10.6	*X*	5.5	5.5
11.1	*Y*	7	–
11.3	*X*	8	8
12.3	*X*	9	9
13.3	*Y*	10	–
14.0	*X*	11	11
14.1	*Y*	12	–
15.2	*X*	13	13
15.6	*X*	14.5	14.5
15.6	*Y*	14.5	–
16.0	*X*	16	16
16.4	*X*	17	17
17.5	*X*	18	18
17.7	*Y*	19	–
17.9	*X*	20	20
19.8	*X*	21	21
20.3	*Y*	22	–
21.0	*Y*	23	–
21.1	*X*	24	24
23.6	*Y*	25	–

Wir berechnen **die Statistik** W:

$$W = \sum_{i=1}^{14} R(Xi)$$

$$= 2 + 4 + 5{,}5 + 8 + 9 + 11 + 13 + 14{,}5 + 16 + 17 + 18 + 20 + 21 + 24 = 183 .$$

Wenn wir die Näherung für das Große wählen Stichproben unter Berücksichtigung der mittleren Ränge, Wir nehmen die Anpassung vor:

$$W1 = \frac{W - \frac{n(N+1)}{2}}{\sqrt{\frac{mn}{12}\left(m+n+1 - \frac{\sum_{j=1}^{G} tj\,t_j^{2} - 1}{(m+n)(m+n-1)}\right)}} .$$

Hier gilt $g = 2$ und $tj = 2$ für $j = 1, 2$. Also wir haben:

$$W1 = \frac{183 - \frac{14(25+1)}{2}}{\sqrt{\frac{11\cdot 14}{12}\left(11+14+1 - \frac{2(2^2-1)+2(2^2-1)}{(11+14)(11+14-1)}\right)}}$$

$$= 0{,}0548 .$$

W

Machine Translated by Google

Die Hypothesen, die wir testen wollen, sind:

H0: Die Jungen dieser Klasse haben keine Chance, im Kopfrechnen besser zu sein als die Mädchen dieser Klasse.

H1: Die Jungen in dieser Klasse neigen dazu, im Kopfrechnen besser zu sein als die Mädchen in dieser Klasse.

Dies wird durch H0 ausgedrückt : $P(X < Y) = \frac{1}{2}$ und 2 entspricht Fall C. Die Entscheidungsregel lautet wie folgt: Verwerfen
Sie H0 auf dem **Signifikanzniveau** ÿ, wenn

$$W > tn,m,1ÿÿ \text{ oder } W1 > z1ÿÿ .$$

Wenn wir ÿ = 0,05 wählen , beträgt der theoretische **Wert** in der **Wilcoxon-Tabelle** $tn,m,1ÿÿ = 212$.
Der berechnete Wert von W ($W = 183$) ist nicht größer als $tn,m,1ÿÿ = 212$, und wir können die **Nullhypothese** H0 nicht ablehnen.
Der Wert in der **Normaltabelle** z1ÿÿ beträgt 1,6449.
Folglich ist W1 nicht größer als z1ÿÿ (0,0548 < 1,6449), und wir ziehen auf der Grundlage der Statistik die gleiche Schlussfolgerung

Tick W: Die Jungen in dieser Klasse haben nicht die Absicht, im Kopfrechnen besser zu sein als die Mädchen in dieser Klasse.

WEITERLESEN

- **Hypothesentest**
- **Mann-Whitney-Test**
- **Nichtparametrischer Test**
- **Von Wilcoxon unterzeichneter Test**
- **Wilcoxon-Tisch**

REFERENZEN

Gibbons, JD: Nichtparametrische statistische Inferenz. McGraw-Hill, New York (1971)

Wilcoxon, F.: Einzelvergleiche nach Ranking-Methoden. Biometrie 1, 80–83 (1945)

Willcox, Walter Francis

Willcox, Walter Francis (1861–1964) schloss sein Studium am Amherst College in Massachusetts ab (1884 erhielt er seinen Bachelor-Abschluss und 1888 seinen Master-Abschluss). Ende der 1880er Jahre setzte er sein Studium an der Columbia University in New York fort. Er erhielt seinen Arzt

1891 Absolvent der Columbia University.
Sein Ph.D. Die Dissertation hieß *„Die Scheidung“. Problem: Eine Studie in Statistik.* Ab 1891 wurde Willcox Professor für Wirtschaftswissenschaften und Statistik an der Cornell University

Stadt, wo er bis zu seiner Pensionierung im Jahr 1931 blieb.

Seine wichtigsten Beiträge zur Statistik lagen im Bereich der Demographie sowie bei der Entwicklung des Systems der Bundesstatistik.

Von 1899 bis 1902 leitete er die 12. Volkszählung der Vereinten Nationen und war Vorsitzender der American Statistical Association

1912, die American Economic Association 1915 und das International Statistical Institute 1947.

Ausgewählte Hauptwerke und Artikel von Willcox, Walther Francis:

1897 Das Scheidungsproblem: Eine Studie zur Statistik. Studium der Geschichte, der Wirtschaftswissenschaften und des öffentlichen Rechts. Columbia versity Press, New York.

1940 Studium der amerikanischen Demographie, Cornell University Press, Ithaca, NY.

Machine Translated by Google

Yates-Algorithmus

Der Yates- **Algorithmus** ist ein Verfahren zur Berechnung der Schätzungen **der** Haupteffekte **und** der **Wechselwirkungen** in einem **faktoriellen Experiment.**

GESCHICHTE

Der Yates- **Algorithmus** wurde 1937 von Yates und Franck entwickelt.

MATHEMATISCHE ASPEKTE

Betrachten Sie ein **faktorielles Experiment** mit *n* **Faktoren,** von denen jeder zwei Ebenen hat . Wir bezeichnen dieses **Experiment** mit *2n.*

Die **Faktoren** werden mit Großbuchstaben bezeichnet.

Wir verwenden einen Kleinbuchstaben für die hohe **Stufe** des Faktors und lassen den Buchstaben für die untere weg eben. Wenn alle Faktoren niedriger sind

Ebene verwenden wir das Symbol (1).

Der **Algorithmus** wird in einer Tabelle dargestellt. In der ersten Spalte finden wir alle **Kombinationen** der **Ebenen** verschiedener **Faktoren** in einer Standardreihenfolge (nach der Einführung eines Buchstabens folgen Kombinationen mit allen vorherigen Kombinationen). Die zweite Spalte enthält die Beobachtungen aller Kombinationen.

Für die folgenden *n* Spalten ist die Vorgehensweise dieselbe: Die erste Hälfte der Spalte ist

erhalten durch Summieren der Paare aufeinanderfolgender Werte der vorherigen Spalte; Die zweite Hälfte der Spalte wird durch Subtrahieren erhalten

Ermitteln des ersten Werts aus dem zweiten Wert jedes Paares.

Spalte *n* enthält **die Kontraste** der in der ersten Spalte angegebenen **Haupteffekte und Wechselwirkungen .** Um die **Schätzer** dieser Effekte zu finden, dividieren wir die n-te Spalte durch *m·2nÿ1,* wobei *m* die Anzahl der Wiederholungen des **Experiments** und *n* die Anzahl **der Faktoren ist.** Anliegen

In der ersten Zeile finden wir in Spalte *n* die Summe aller **Beobachtungen.**

DOMÄNEN UND EINSCHRÄNKUNGEN

Der Yates-Algorithmus kann auf **Experimente** mit **Faktoren** angewendet werden , die mehr als haben zwei **Ebenen.**

Mit diesem **Algorithmus** kann man auch die Summe der Quadrate ermitteln, die den **Haupteffekten** und Wechselwirkungen der *2n-* **Fakultät** entsprechen **Experiment.** Sie erhält man, indem man die n-te Spalte quadriert und das Ergebnis durch m· *2n* dividiert *(wobei mis* die Anzahl der Replikationen und *n* die Anzahl der **Faktoren ist).**

BEISPIELE

Um die **Schätzungen** der **Haupteffekte** und der drei **Wechselwirkungen zu berechnen,** betrachten wir ein faktorielles Experiment 23 (d. h. drei Faktoren mit zwei Ebenen) relativ zur Zuckerrübenproduktion. Der erste Faktor, mit A bezeichnet, betrifft die Art des Düngemittels (Gülle oder Nicht-Gülle); Die

Machine Translated by Google

Zweitens, mit B bezeichnet, betrifft die Tiefe von Pflügen (20 oder 30 cm); und der letzte Faktor ist C, die Zuckerrübensorte (Sorte 1 oder 2). Da jeder Faktor zwei Ebenen hat, ist dies möglich Man spricht von einem 2 × 2 × 2-Experiment. Beispielsweise wird die **Wechselwirkung** zwischen **Faktor** A und Faktor B mit A×B und bezeichnet Das **Zusammenspiel** aller drei Faktoren ist geschrieben A × B × C.

Die acht **Beobachtungen** geben das Gewicht an der Ernte in Kilogramm für die Hälfte pro Hektar, was die folgende Tabelle ergibt:

A	Düngen		Nichtmist	
B	20 cm	30 cm	20 cm	30 cm
Var. 1	80	96	84	88
C				
Var. 2	86	96	89	93

Mit diesem Experiment wollen wir nicht nur die Wirkung von Düngung, Pflügen, und die Vielfalt der Rüben (einfache Effekte), aber auch die Auswirkung kombinierter Effekte (effects der **Interaktion)** bei der Ernte.
Um die Tabelle des Yates-Algorithmus zu erstellen, Die acht **Beobachtungen** müssen eingeordnet werden Yates' Befehl. Die Klassifikationstabelle ist in diesem Fall praktisch; die Symbole „ÿ“ und „+“ repräsentieren jeweils die untere und die obere Ebene der Faktoren. Zum Beispiel die erste Zeile gibt die Beobachtung „80“ an, die den niedrigeren Werten für alle drei Faktoren entspricht (A: Mist, B: 20 cm, C: Sorte 1).

Beobachtungsklassifizierungstabelle in Yates-Reihenfolge

Obs.	Faktor ABC	Dazugehörigen Überwachung
1 ÿÿÿ	80	
2 +ÿÿ	84	
3 ÿ+ÿ	96	
4 ++ÿ	88	
5 ÿÿ+	86	
6 +ÿ+	89	
7 ÿ++	96	
8 +++	93	

Wir können nun die Tabelle des Yates-Algorithmus erstellen:

Tabelle entsprechend dem Yates-Algorithmus

Kombi-Nation	Obs.	Säulen 123			Durchschnitt Einschätzung	Wirkung
(1)	80	164	348	712	89	—
A	84	184	364	ÿ4	ÿ1	A
B	96	175	ÿ4	34	8,5	B
ab	88	189		0	ÿ18	*ÿ4,5* A × B
C	86	4	20	16	4	C
ac	89	ÿ8	14		4	1 A × C
v. Chr	96	3	ÿ12	ÿ6	*ÿ1,5*	B × C
abc	93	ÿ3	ÿ6	6		1,5 A × B × C

Die Terme in Spalte 2 werden beispielsweise wie folgt berechnet:

- Erste Hälfte der Spalte (Summe der Elemente von In diesem Fall jedes Paar der vorherigen Spalte Spalte 1):

$$348 = 164 + 184$$

$$364 = 175 + 189$$

$$ÿ4 = 4 + (ÿ8)$$

$$0 = 3 + (ÿ3)$$

- Zweite Hälfte der Spalte (Subtraktion der Elemente jedes Paares in umgekehrter Reihenfolge):

$$20 = 184 ÿ 164$$

$$14 = 189 ÿ 175$$

$$ÿ12 = ÿ8 ÿ 4$$

$$ÿ6 = ÿ3 ÿ 3$$

Machine Translated by Google

Die Schätzung für den **Haupteffekt** von **fac**

tor A ist gleich ÿ1. Wie man sieht, wurde der Wert ÿ4 (Spalte 3) durch $m \cdot 2n\ddot{y}1 = 1 \cdot 23\ddot{y}1 = 4$ geteilt, wobei m gleich 1 ist, weil das Experiment nur einmal (ohne Wiederholung) durchgeführt wurde und n entspricht der Anzahl der Faktoren, die hier drei beträgt

Fall.

WEITERLESEN

Varianzanalyse
Kontrast
Versuchsplanung
Faktorielles Experiment
Interaktion
Haupteffekt

REFERENZEN

Yates, F.: Das Design und die Analyse faktorieller Experimente. Technische Kommunikation des Commonwealth Bureau

of Soils **35,** Commonwealth Agricultural Bureaux, Farnham Royal (1937)

Youden, William John

Youden, William John (1900–1971) erwarb 1924 seinen Master-Abschluss in Chemie an der Columbia University. Von 1924 bis 1948 arbeitete er am Boyce Thompson Institute FOR Plant Research in

Yonkers, NY. Von 1948 bis zu seiner Pensionierung im Jahr 1965 war Youden, WJ, Mitglied der Abteilung für Angewandte Mathematik im National Bureau of Standards in Washington, DC. Während dieser Zeit entwickelte Youden einen Test und schrieb Statistical Methods for Chemists .

Hauptwerk von Youden, William John:

1951 Statistische Methoden für Chemiker.

Wiley, New York.

Yule- und Kendall-Koeffizient

Zur Messung wird der Yule-Koeffizient verwendet Schiefe einer **Häufigkeitsverteilung.** Es berücksichtigt die relativen Positionen der **Quartile** in Bezug auf den Median und vergleicht die Ausbreitung der Kurve rechts und links vom Median.

MATHEMATISCHE ASPEKTE

Bei einer symmetrischen Verteilung liegen die **Quartile** auf beiden Seiten des **Medians im gleichen Abstand.** Das bedeutet, dass

$$(Q3 \ddot{y} M) \ddot{y} (M \ddot{y} Q1) = 0 \quad ,$$

Dabei ist M der Median, Q1 das erste Quartil und Q3 das dritte Quartil.
Wenn die Verteilung asymmetrisch ist, ist die vorherige Gleichheit nicht wahr.
Der linke Teil der Gleichung kann folgendermaßen umsortiert werden:

$$Q3 + Q1 \ddot{y} 2M \quad .$$

Um einen Schiefekoeffizienten zu erhalten

Unabhängig von der Maßeinheit sollte die Gleichung durch den Wert $(Q3 \ddot{y} Q1)$ dividiert werden.
Der Weihnachtskoeffizient ist dann:

$$CY = \frac{Q3 + Q1 \ddot{y} 2M}{Q3 \ddot{y} Q1} \quad .$$

Wenn CY positiv ist, breitet sich die Verteilung nach rechts aus; Ist CY negativ, breitet sich die Verteilung nach links aus.

DOMÄNEN UND EINSCHRÄNKUNGEN

Wenn der Koeffizient positiv ist, stellt er eine Kurve dar, die sich nach rechts ausbreitet. Ist sie positiv, dann stellt sie eine Kurve dar, die sich nach links ausbreitet. Ein Koeffizient also

Machine Translated by Google

Nahe Null bedeutet, dass die Kurve annähernd symmetrisch ist.
Der Yule- und Kendall-Koeffizient kann variieren zwischen ÿ1 und +1. Tatsächlich liegt das **Medium** *M* immer zwischen dem ersten und dem ersten drittes **Quartil,** die folgenden beiden Extreme Möglichkeiten ergeben sich:

1. $M = Q1$
2. $M = Q3$

Im ersten Fall wird der Koeffizient:

$$CY = \frac{Q3 + Q1 \text{ ÿ } 2Q1}{Q3 \text{ ÿ } Q1} = 1 \quad ,$$

und im zweiten Fall wird es:

$$CY = \frac{Q3 + Q1 \text{ ÿ } 2Q3}{Q3 \text{ ÿ } Q1} = \text{ÿ}1 \quad .$$

Unter diesen Bedingungen sollte der Koeffizient sein nahe Null, damit die Verteilung con ist seitensymmetrisch.
Dieser Koeffizient sowie die anderen Mea Parameter **der Schiefe** sind nur dann von Interesse, wenn sie kann die Formen von zwei oder mehreren vergleichen Verteilungen. Die Ergebnisse variieren erheblich von einer Formel zur anderen. Es ist offensichtlich das Vergleiche müssen anhand derselben durchgeführt werden Formel.

BEISPIELE

Angenommen, wir möchten die Form vergleichen der Verteilung des Tagesumsatzes im Jahr 75 Bäckereien über 2 Jahre. Wir rechnen dann ein in beiden Fällen der Yule- und der Kendall-Koeffizient.

Die **Daten** sind folgende:

Umsatzhäufigkeit	Jahr 1	Frequenz Jahr 2
215–235	4	25
235–255	6	15
255–275	13	9
275–295	22	8
295–315	15	6
315–335	6	5
335–355	5	4
355–375	4	3

Für Jahr 1 das **erste Quartil, der Median** und das dritte Quartil sind jeweils:

$$Q1 = 268{,}46 \quad ,$$
$$M = 288{,}18 \quad ,$$
$$Q3 = 310 \; .$$

Der Koeffizient ist dann folgender:

$$CY = \frac{Q3 + Q1 \text{ ÿ } 2M}{Q3 \text{ ÿ } Q1}$$
$$= \frac{310 + 268{,}46 \text{ ÿ } 2(288{,}18)}{310 \text{ ÿ } 268{,}46}$$
$$= 0{,}05 \; .$$

Für das Jahr 2 lauten diese **Werte** jeweils:

$$Q1 = 230 \quad ,$$
$$M = 251{,}67 \quad ,$$
$$Q3 = 293{,}12 \; .$$

Der Yule- und Kendall-Koeffizient beträgt dann gleich:

$$CY = \frac{Q3 + Q1 \text{ ÿ } 2M}{Q3 \text{ ÿ } Q1}$$
$$= \frac{293{,}12 + 230 \text{ ÿ } 2(251{,}67)}{293{,}12 \text{ ÿ } 230}$$
$$= 0{,}31 \; .$$

Für Jahr 1, da der Koeffizient ein Ergebnis liefert Das ist nahe Null, wir können zugeben, dass die Verteilung des Tagesumsatzes im 75 Bäckereien kommt einer symmetrischen Verteilung sehr nahe. Für Jahr 2 ist der Koeffizient höher; das bedeutet, dass sich die Verteilung ausbreitet nach rechts.

Machine Translated by Google

WEITERLESEN

Formmaß
Maß für die Schiefe

REFERENZEN

Yule, GV, Kendall, MG: An Introduction to the Theory of Statistics, 14. Aufl. Griffflosse, London (1968)

Yule, George Udny

Yule, George Udny wurde 1871 in Beech Hill in der Nähe von Haddington (Ecosse) geboren und starb 1951 in Cambridge.
Der Yule-Assoziationskoeffizient, das Yule-Paradoxon (später auch Simp-Sohn-Paradoxon genannt) und das Yule-Verfahren tragen in der Statistik seinen Namen, die ebenfalls dazu beitrageted zur Mendel-Theorie und der Zeitreihe.

Ausgewählte Hauptwerke von Yule, George Udny:

1897 Zur Korrelationstheorie. J. Roy. Stat. Soc. 60, 812–854.

1900 Über die Assoziation von Attributen in Statistik: mit Illustration aus dem Material der Kindergesellschaft.Phi los. Trans. Roy. Soc. London. Ser. A 194, 257–319.

1926 Warum erhalten wir manchmal unsinnige Korrelationen zwischen Zeitreihen? Eine Studie über Stichproben und die Natur von Zeitreihen. J. Roy. Stat. Soc. (2) 89, 1–

64. **1968** (mit Kendall, MG) Eine Einführung in die Theorie der Statistik. 14. Aufl. Griffin, London.

WEITERLESEN

Maß für die Schiefe
Yule- und Kendall-Koeffizient

Machine Translated by Google

Anhang

Machine Translated by Google

Anhang A: Tabelle der Zufallszahlen, Dezimalstellen von ÿ

1415926535 3305727036 5024459455 8583616035 8164706001 9465764078
8979323846 5759591953 3469083026 6370766010 6145249192 9512694683
2643383279 9218611739 4252230825 4710181942 1732172147 9835259570
5028841971 8193261179 3344685035 9555961989 7235014144 9825822620
6939937510 3105118548 2619311881 4676783744 1973568548 5224894077
5820974944 7446237999 7101000313 9448255379 1613611573 2671947826
5923078164 6274956735 7838752886 7747268471 5255213347 8482601476
6286208995 1885752724 5875332083 4047534640 5741849468 9909026401
8628034825 8912279381 8142061717 6208046684 4385233239 3639443745
3421170679 8301194912 7669147303 2590694912 7394143337 5305068203
8214808651 9833673362 5982534904 3313677028 4547762416 4962524517
3282306647 4406566430 2875546873 8989152104 8625189835 4939965143
9384460952 8602139494 1159562863 7521620569 6948556209 1429809190
5058223172 6395224737 8823537875 6602405803 9219222184 6592509372
5359408128 1907021798 9375195778 8150193511 2725502542 2169646151
4811174502 6094370277 1857780532 2533824300 5688767179 5709858387
8410270193 5392171760 1712268066 3558764024 4946016530 4105978859
8521105559 2931767523 1300192787 7496473263 4668049886 5977297549
6446229489 8467481846 6611195909 9141992726 2723279178 8930161753
5493038196 7669405132 2164201989 4269922792 6085784383 9284681382
4428810975 5681271702 3809525720 6782354781 8279679766 6868386894
6659334461 4526356082 1065485863 6360093417 8145410095 2774155991
2847564823 7785771342 2788659361 2164121992 3883786360 8559252459
3786783165 7517896091 5338182796 4586315030 9506800642 5395943104
2712019091 7363717872 8230301952 2861829745 2512520511 9972524680
4564856692 1468440901 3530185292 5570674983 7392984896 8459872736
3460348610 2249534301 6899577362 8505494588 8412848862 4469584865
4543266482 4654958537 2599413891 5869269956 2694560424 3836736222
1339360726 1050792279 2497217752 9092721079 1965285022 6260991246
2491412733 6892589235 8347913151 7509302955 2106611863 8051243888
7245870066 4201995611 5574857242 3211653449 6744278629 4390451244
6315588170 2129021960 4541500959 8720275596 2039194945 1365497627
4881520920 8640344181 5082953311 2364806657 4712371373 8079771569
9628292540 5981362977 6861727855 4991198818 8696095636 1435997700
9171536436 4771309960 8890750983 3479775356 4371917287 1296160894
7892590360 5187072113 8175463746 6369807426 4677646575 4169486855
1133053052 4999999837 4939319255 5425278625 7396241389 5848406353
4882046652 2978049951 6040092774 5181841757 8658326451 4220722258
1384146951 5973173280 1671139001 4672890977 9958133904 2848864815
9415116094 1609631859 9848824012 7727938000 7802759009 8456028506

Machine Translated by Google

Anhang B: Binomialtabelle

Nummer von Versuchen		Erfolgswahrscheinlichkeit								
		0,01	0,02	0,03	0,04	0,05	0,1	0,2	0,3	0,5
	0	0,9510	0,9039	0,8587	0,8154	0,7738	0,5905	0,3277	0,1681	0,0313
	1	0,9990	0,9962	0,9915	0,9852	0,9774	0,9185	0,7373	0,5282	0,1875
	2	1,0000	0,9999	0,9997	0,9994	0,9988	0,9914	0,9421	0,8369	0,5000
5	3	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9933	0,9692	0,8125
	4	1	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9976	0,9688
	5	1	1	1	1	1	1	1	1	1
	0	0,9044	0,8171	0,7374	0,6648	0,5987	0,3487	0,1074	0,0282	0,0010
	1	0,9957	0,9838	0,9655	0,9418	0,9139	0,7361	0,3758	0,1493	0,0107
	2	0,9999	0,9991	0,9972	0;9938	0,9885	0,9298	0,6778	0,3828	0,0547
	3	1,0000	1,0000	0,9999	0,9996	0,9990	0,9872	0,8791	0,6496	0,1719
	4	1,0000	1,0000	1,0000	1,0000	0,9999	0,9984	0,0672	0,8497	0,3770
	5	1	1,0000	1,0000	1,0000	1,0000	0,9999	0,9936	0,9527	0,6230
10	6	1	1	1,0000	1,0000	1,0000	1,0000	0,9991	0,9894	0,8281
	7	1	1	1	1	1,0000	1,0000	0,9999	0,9984	0,9453
	8	1	1	1	1	1	1,0000	1,0000	0,9999	0,9893
	9	1	1	1	1	1	1	1,0000	1,0000	0,9990
	10	1	1	1	1	1	1	1	1	1
	0	0,8601	0,7386	0,6333	0,5421	0,4633	0,2059	0,0352	0,0047	0,0000
	1	0,9904	0,9647	0,9270	0,8809	0,8290	0,5490	0,1671	0,0353	0,0005
	2	0,9996	0,9970	0,9906	0,9797	0,9638	0,8159	0,3980	0,1268	0,0037
	3	1,0000	0,9998	0,9992	0,9976	0,9945	0,9444	0,6482	0,2969	0,0176
	4	1,0000	1,0000	0,9999	0,9998	0,9994	0,9873	0,8358	0,5155	0,0592
	5	1,0000	1,0000	1,0000	1,0000	0,9999	0,9978	0,9389	0,7216	0,1509
	6	1	1,0000	1,0000	1,0000	1,0000	0,9997	0,9819	0,8689	0,3036
15	7	1	1	1,0000	1,0000	1,0000	1,0000	0,9958	0,9500	0,5000
	8	1	1	1	1,0000	1,0000	1,0000	0,9992	0,9848	0,6964
	9	1	1	1	1	1	1,0000	0,9999	0,9963	0,8491
	10	1	1	1	1	1	1,0000	1,0000	0,9993	0,9408
	11	1	1	1	1	1	1	1,0000	0,9999	0,9824
	12	1	1	1	1	1	1	1,0000	1,0000	0,9963
	13	1	1	1	1	1	1	1,0000	1,0000	0,9995
	14	1	1	1	1	1	1	1	1,0000	1,0000
	15	1	1	1	1	1	1	1	1	1

Machine Translated by Google

Nummer von Versuchen		Erfolgswahrscheinlichkeit								
		0,01	0,02	0,03	0,04	0,05	0,1	0,2	0,3	0,5
	0	0,8179	0,6676	0,5438	0,4420	0,1585	0,1216	0,0115	0,0008	0,0000
	1	0,9831	0,9401	0,8802	0,8103	0,7358	0,3917	0,0692	0,0076	0,0000
	2	0,9990	0,9929	0,9790	0,9561	0,9245	0,6769	0,2061	0,0355	0,0002
	3	1,0000	0,9994	0,9973	0,9926	0,9841	0,8670	0,4114	0,1071	0,0013
	4	1,0000	1,0000	0,9997	0,9990	0;9974	0,9568	0,6296	0,2375	0,0059
	5	1,0000	1,0000	1,0000	0,9999	0,9001	0,9887	0,8042	0,4164	0,0207
	6	1,0000	1,0000	1,0000	1,0000	1,0000	0,9976	0,9133	0,6080	0,0577
	7	1	1,0000	1,0000	1,0000	1,0000	0,9996	0,9679	0,7723	0,1316
	8	1	1	1,0000	1,0000	1,0000	0,9999	0,9900	0,8867	0,2517
	9	1	1	1	1,0000	1,0000	1,0000	0,9974	0,9520	0,4119
20	10	1	1	1	1	1,0000	1,0000	0,9994	0,9829	0,5881
	11	1	1	1	1	1	1,0000	0,9999	0,9949	0,7483
	12	1	1	1	1	1	1,0000	1,0000	0,9987	0,8684
	13	1	1	1	1	1	1	1,0000	0,9997	0,9423
	14	1	1	1	1	1	1	1,0000	1,0000	0,9793
	15	1	1	1	1	1	1	1,0000	1,0000	0,9941
	16	1	1	1	1	1	1	1,0000	1,0000	0,9987
	17	1	1	1	1	1	1	1	1,0000	0,9998
	18	1	1	1	1	1	1	1	1,0000	1,0000
	19	1	1	1	1	1	1	1	1	1,0000
	20	1	1	1	1	1	1	1	1	1

Machine Translated by Google

Anzahl von Versuchen	Erfolgswahrscheinlichkeit								
	0,01	0,02	0,03	0,04	0,05	0,1	0,2	0,3	0,5
0	0,7397	0,5455	0,4010	0,2939	0,2146	0,0424	0,0012	0,0000	0,0000
1	0,9639	0,8795	0,7731	0,6612	0,5535	0,1837	0,0105	0,0003	0,0000
2	0,9967	0,9783	0,9399	0,8831	0,8122	0,4114	0,0442	0,0021	0,0000
3	0,9998	0,9971	0,9881	0,9694	0,9392	0,6474	0,1227	0,0093	0,0000
4	1,0000	0,9997	0,9982	0,9937	0,9844	0,8245	0,2552	0,0302	0,0000
5	1,0000	1,0000	0,9998	0,9989	0,9967	0,9268	0,4275	0,0766	0,0002
6	1,0000	1,0000	1,0000	0,9999	0,9994	0,9742	0,6070	0,1595	0,0007
7	1	1,0000	1,0000	1,0000	0,9999	0,9922	0,7608	0,2814	0,0026
8	1	1,0000	1,0000	1,0000	1,0000	0,9980	0,8713	0,4315	0,0081
9	1	1	1,0000	1,0000	1,0000	0,9995	0,9389	0,5888	0,0214
10	1	1	1,0000	1,0000	1,0000	0,9999	0,9744	0,7304	0,0494
11	1	1	1	1,0000	1,0000	1,0000	0,9905	0,8407	0,1002
12	1	1	1	1	1,0000	1,0000	0,9969	0,9155	0,1808
13	1	1	1	1	1	1,0000	0,9991	0,9599	0,2923
14	1	1	1	1	1	1,0000	0,9998	0,9831	0,4278
30 15	1	1	1	1	1	1,0000	0,9999	0,9936	0,5722
16	1	1	1	1	1	1	1,0000	0,9979	0,7077
17	1	1	1	1	1	1	1,0000	0,9994	0,8192
18	1	1	1	1	1	1	1,0000	0,9998	0,8998
19	1	1	1	1	1	1	1,0000	1,0000	0,9506
20	1	1	1	1	1	1	1,0000	1,0000	0,9786
21	1	1	1	1	1	1	1	1,0000	0,9919
22	1	1	1	1	1	1	1	1,0000	0,9974
23	1	1	1	1	1	1	1	1,0000	0,9993
24	1	1	1	1	1	1	1	1,0000	0,9998
25	1	1	1	1	1	1	1	1	1,0000
26	1	1	1	1	1	1	1	1	1,0000
27	1	1	1	1	1	1	1	1	1,0000
28	1	1	1	1	1	1	1	1	1,0000
29	1	1	1	1	1	1	1	1	1,0000
30	1	1	1	1	1	1	1	1	1

Machine Translated by Google

Anhang C: Fisher-Tabelle für ÿ = 005

	ÿ1									
ÿ2	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10.13	9.55		9.28	9.12	9.01	8,94	8,89	8,85	8.81	8,79
4	7.71	6,94		6.59	6.39	6.26	6.16	6.09	6.04	6.00	5,96
5	6.61	5,79		5.41	5.19	5.05	4,95	4,88	4,82	4,77	4,74
6	5,99	5.14		4,76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4,74		4.35	4.12	3,97	3,87	3,79	3,73	3,68	3,64
8	5.32	4.46		4.07	3,84	3,69	3,58	3,50	3.44	3.39	3.35
9	5.12	4.26		3,86	3,63	3.48	3.37	3.29	3.23	3.18	3.14
10	4,96	4.10		3,71	3.48	3.33	3.22	3.14	3.07	3.02	2,98
11	4,84	3,98		3,59	3.36	3.20	3.09	3.01	2,95	2,90	2,85
12	4,75	3,89		3.49	3.26	3.11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81		3.41	3.18	3.03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74		3.34	3.11	2,96	2,85	2,76	2,70	2,65	2,60
15	4.54	3,68		3.29	3.06	2,90	2,79	2,71	2,64	2,59	2,54
16	4.49	3,63		3.24	3.01	2,85	2,74	2,66	2,59	2,54	2,49
17	4.45	3,59		3.20	2,96	2,81	2,70	2.61	2,55	2,49	2,45
18	4.41	3,55		3.16	2,93	2,77	2,66	2,58	2.51	2.46	2.41
19	4.38	3,52		3.13	2,90	2,74	2,63	2,54	2,48	2.42	2,38
20	4.35	3.49		3.10	2,87	2,71	2,60	2.51	2,45	2.39	2,35
21	4.32	3.47		3.07	2,84	2,68	2,57	2,49	2.42	2.37	2.32
22	4.30	3.44		3.05	2,82	2,66	2,55	2.46	2,40	2.34	2.30
23	4.28	3.42		3.03	2,80	2,64	2,53	2.44	2.37	2.32	2.27
24	4.26	3.40		3.01	2,78	2,62	2.51	2.42	2.36	2.30	2,25
25	4.24	3.39		2,99	2,76	2,60	2,49	2,40	2.34	2.28	2.24
26	4.23	3.37		2,98	2,74	2,59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35		2,96	2,73	2,57	2.46	2.37	2.31	2,25	2.20
28	4.20	3.34		2,95	2,71	2,56	2,45	2.36	2.29	2.24	2.19
29	4.18	3.33		2,93	2,70	2,55	2.43	2;35	2.28	2.22	2.18
30	4.17	3.32		2,92	2,69	2,53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23		2,84	2.61	2,45	2.34	2,25	2.18	2.12	2.08
60	4.00	3.15		2,76	2,53	2.37	2,25	2.17	2.10	2.04	1,99
120	3,92	3.07		2,68	2,45	2.29	2.17	2.09	2.02	1,96	1,91
ÿ 3,84		3,00		2,60	2.37	2.21	2.10	2.01	1,94	1,88	1,83

Machine Translated by Google

	ÿ1								
ÿ2	12	15	20	24	30	40	60	120 ÿ	
1	243,9	245,9	248,0	249.1	250.1	251.1	252,2	253,3	254,3
2	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8,74	8,70	8.66	8,64	8.62	8.59	8.57	8.55	8.53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5.63
5	4,68	4.62	4.56	4.53	4,50	4.46	4.43	4.40	4.36
6	4.00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2,97	2,93
9	3.07	3.01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2.61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2.51	2.47	2.43	2,38	2.34	2.30
13	2,60	2,53	2.46	2.42	2,38	2.34	2.30	2,25	2.21
14	2,53	2.46	2.39	2,35	2.31	2.27	2.22	2.18	2.13
15	2,48	2,40	2.33	2.29	2,25	2.20	2.16	2.11	2.07
16	2.42	2,35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2,38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1,96
18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1,97	1,92
19	2.31	2.23	2.16	2.11	2.07	2.03	1,98	1,93	1,88
20	2.28	2.20	2.12	2.08	2.04	1,99	1,95	1,90	1,84
21	2,25	2.18	2.10	2.05	2.01	1;96	1,92	1,87	1,81
22	2.23	2.15	2.07	2.03	1,98	1,94	1,89	1,84	1,78
23	2.20	2.13	2.05	2.01	1,96	1,91	1,86	1,81	1,76
24	2.18	2.11	2.03	1,98	1,94	1,89	1,84	1,79	1,73
25	2.16	2.09	2.01	1,96	1,92	1,87	1,82	1,77	1,71
26	2.15	2.07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2.13	2.06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2.12	2.04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2.10	2.03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2.09	2.01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
ÿ 1,75		1,67	1,57	1,52	1,46	1,39	1,32	1.22	1,00

Machine Translated by Google

Anhang D: Kruskal-Wallis-Tabelle

N_1	N_2	N_3	Kritischer Wert ÿ	
2	1	1	2,7000	0,500
2	2	1	3.6000	0,267
2	2	2	4.5711	0,067
			3.7143	0,200
3	1	1	3.2000	0,300
3	2	1	4.2857	0,100
			3,8571	0,133
3	2	2	5.3572	0,029
			4.7143	0,048
			4.5000	0,067
			4.4643	0,105
3	3	1	5.1429	0,043
			4.5714	0,100
			4.0000	0,129
3	3	2	6.2500	0,011
			5.3611	0,032
			5.1389	0,061
			4.5556	0,100
			4.2500	0,121
3	3	3	7.2000	0,004
			6.4889	0,011
			5,6889	0,029
			5.0667	0,086
			4.6222	0,100
4	1	1	3.5714	0,200
4	2	1	4.8214	0,057
			4.5000	0,076
			4.0179	0,114
4	2	2	6.0000	0,014
			5.3333	0,033
			5.1250	0,052
			4,3750	0,100
			4.1667	0,105
4	3	1	5.8333	0,021
			5.2083	0,050
			5.0000	0,057
			4.0556	0,093
			3,8889	0,129

N_1	N_2	N_3	Kritischer Wert	ÿ
4	3	2	6.4444	0,009
			6.4222	0,010
			5.4444	0,047
			5.4000	0,052
			4.5111	0,098
			4.4666	0,101
4	3	3	6.7455	0,010
			6.7091	0,013
			5.7909	0,046
			5.7273	0,050
			4.7091	0,094
			4,7000	0,101
4	4	1	6.6667	0,010
			6.1667	0,013
			4,9667	0,046
			4,8667	0,054
			4.1667	0,082
			4.0667	0,102
4	4	2	7.0364	0,006
			6,8727	0,011
			5.4545	0,046
			5.2364	0,052
			4.5545	0,098
			4.4455	0,103
4	4	3	7.1739	0,010
			7.1364	0,011
			5,5985	0,049
			5.5758	0,051
			4.5455	0,099
			4.4773	0,102
4	4	4	7.6538	0,008
			7,5385	0,011
			5.6923	0,049
			5.6538	0,054
			4.6539	0,097
			4.5001	0,104

Machine Translated by Google

N_1	N_2	N_3	Kritischer Wert	ÿ
5	1	1	3,8571	0,143
5	2	1	5,2500	0,036
			5.0000	0,048
			4,4500	0,071
			4.2000	0,095
			4.0500	0,119
5	2	2	6.5333	0,008
			6.1333	0,013
			5.1600	0,034
			5.0400	0,056
			4.3733	0,090
			4.2933	0,122
5	3	1	6.4000	0,012
			4,9600	0,048
			4.8711	0,052
			4.0178	0,095
			3,8400	0,123
5	3	2	6.9091	0,009
			6.8606	0,011
			5.4424	0,048
			5.3455	0,050
			4.5333	0,097
			4.4121	0,109
5	3	3	6.9818	0,010
			6.8608	0,011
			5.4424	0,048
			5.3455	0,050
			4.5333	0,097
			4.4121	0,109
5	4	1	6.9545	0,008
			6,8400	0,011
			4.9855	0,044
			4,8600	0,056
			3,9873	0,098
			3,9600	0,102
5	4	2	7.2045	0,009
			7.1182	0,010
			5.2727	0,049
			5.2682	0,050
			4.5409	0,098
			4.5182	0,101

N_1	N_2	N_3	Kritischer Wert	ÿ
5	4	3	7.4449	0,010
			7.3949	0,011
			5.6564	0,049
			5.6308	0,051
			4.5487	0,099
			4.5231	0,103
5	4	4	7.7604	0,009
			7.7440	0,011
			5.6571	0,049
			5;6116	0,050
			4.6187	0,100
			4.5527	0,102
5	5	1	7.3091	0,009
			6.8364	0,011
			5.1273	0,049
			4.9091	0,053
			4.1091	0,086
			4.0364	0,105
5	5	2	7.3385	0,010
			7.5429	0,010
			5.7055	0,046
			5.6264	0,051
			4.5451	0,100
			4.5363	0,102
5	5	4	7.8229	0,010
			7.1914	0,010
			5.6657	0,049
			5.6429	0,050
			4.5229	0,099
			4,6200	0,101
5	5	5	8.0000	0,09
			7,9800	0,010
			5,7800	0,049
			5,6600	0,051
			4,5600	0,100
			4.5000	0,102

Machine Translated by Google

Anhang E: Schülertisch

	ÿ					
ÿ	0,100	0,050	0,025	0,010	0,005	0,001
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.315	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.212	2.821	3.250	4.297
10	1.312	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.182	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.133	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.519	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.103	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.046	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.123	2.704	3.301
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.571	2.000	2.300	2.660	3.232
70	1.294	1.667	1.994	2.381	2.648	3.211
80	1.292	1.664	1.990	2.374	2.639	3.195
90	1.291	1.662	1.987	2.368	2.632	3.183
100	1.290	1.660	1.984	2.364	2.626	3.174
ÿ 1,282		1.645	1.960	2.326	2.576	3.090

Machine Translated by Google

Anhang F: Chi-Quadrat-Tabelle

	Grad der Freiheit					
ȳ	0,990	0,950	0,10	0,05	0,025	0,01
1	0,000	0,004	2,71	3,84	5.020	6.63
2	0.020	0,103	4.61	5,99	7.380	9.21
3	0,115	0,352	6.25	7,81	9.350	11.34
4	0,297	0,711	7,78	9.49	11.140	13.23
5	0,554	1.145	9.24	11.07	12.830	15.09
6	0,872	1,635	10,64	12.53	14.450	16.81
7	1.239	2.107	12.02	14.07	16.010	18.48
8	1.646	2.733	13.36	15.51	17.530	20.09
9	2.088	3,325	14,68	16.92	19.020	21.67
10	2.558	3.940	15,99	18.31	20.480	23.21
11	3.050	4.580	17.29	19.68	21.920	21.72
12	3.570	5.230	18.55	21.03	23.340	26.22
13	4.110	5.890	19.81	21.36	24.740	27.69
14	4.660	6.570	21.06	23.68	26.120	29.14
15	5.230	7.260	22.31	25.00	27.100	30.58
16	5.810	7.960	23.54	26.30	28.850	32.00
17	6.410	8.670	24.77	27.59	30,190	33,41
18	7.020	9.390	25,99	28.87	31,530	34,81
19	7.630	10.120	27.20	30.14	32.850	36.19
20	8.260	10.850	28.41	31.41	34,170	37,57
21	8.900	11.590	29,62	39,67	35,480	38,93
22	9,540	12,340	30,81	33,92	36.780	40.29
23	10.200	13.090	32.01	35.17	38.080	41.64
24	10.860	13.850	33.20	36,42	39.360	42,98
25	11.520	14.610	34.38	37,65	40.650	44.31
26	12.200	15.380	35.56	38,89	41.920	45.64
27	12.880	16.150	36.74	40.11	43,190	46,96
28	13.570	16.930	37.92	41,34	44,160	48,28
29	14.260	17.710	39.09	42,56	45.720	49.59
30	14.950	18.490	40.26	43,77	46,980	50,89

Machine Translated by Google

Anhang G: Standard-Normaltabelle

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0 5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6501	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8119	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Machine Translated by Google

Literaturverzeichnis

Abdi, H.: Binomialverteilung: Binomial- und Vorzeichentests. In: Salkind, NJ (Hrsg.) Encyclopedia of Measurement and Statistics. Salbei, tausend Eichen (2007) Binomialtest

Abdi, H.: Distanz. In: Salkind, NJ (Hrsg.) Enzyklopädie der Messung und Statistik. Salbei, tausend Eichen (2007) Distanz

Agresti, A.: Kategorische Datenanalyse. Wiley, Neue Analyse York (1990) kategorialer Daten

Airy, GB: Brief von Professor Airy, Astronomer Royal, an den Herausgeber. Astronom. J. 4, 137–138 (1856) Ausreißer

Allan, FE, Wishart, J.: Eine Methode zur Schätzung des Ertrags einer fehlenden Parzelle in experimentellen Feldarbeiten. J. Landwirtschaft. Wissenschaft. **20,** 399–406 (1930) Fehlende Daten

Altman, DG: Praktische Statistik für die medizinische Forschung. Chapman & Hall, London (1991) Logistische Regression

Anderson, TW, Darling, DA: Asymptotische Theorie bestimmter Anpassungsgütekriterien basierend auf stochastischen Prozessen. Ann. Mathematik. Stat. **23,** 193–212 (1952) Anderson-Darling-Test

Anderson, TW, Darling, DA: Ein Test der Anpassungsgüte. Marmelade. Stat. Assoc. **49,** 765–769 (1954) Anderson-Darling-Test

Anderson, TW, Sclove, SL: Eine Einführung in die Statistische Analyse von Daten. Houghton Mifflin, Boston (1978) Range

Anderson, VL, McLean, RA: Design von Experimente: ein realistischer Ansatz. Marcel Dekker, New York (1974) Modell

Andrews DF, Hertzberg, AM: Daten: Eine Sammlung von Problemen aus vielen Bereichen für Studenten und Forscher. Springer, Berlin Heidelberg New York (1985) *Cp*- Kriterium

Anscombe, FJ, Tukey, JW: Analyse von Residuen. Technometrics 5, 141–160 (1963) Analyse von Reste

Anscombe, FJ: Untersuchung von Residuen. Proz. 4 Berkeley Symp. Mathematik. Statist. Wahrscheinlich. 1, 1–36 (1961) Analyse von Residuen

Anscombe, FJ: Diagramme in der statistischen Analyse. Bin. Stat. **27,** 17–21 (1973) Analyse von Residuen

Antille, G., Ujvari, A.: Pratique de la Statistique Inférentielle. PAN, Neuchâtel, Schweiz (1991) Datensammlung

Arbuthnott, J.: Ein Argument für die göttliche Vorsehung, basierend auf der konstanten Regelmäßigkeit, die bei den Geburten beider Geschlechter beobachtet wird. Philos. Trans. **27,** 186–190 (1710) Hypothesentest, nichtparametrischer Test, Vorzeichentest

Armatte, M.: La construction des notions Schätzung und Bewertung durch Ronald A. Fischer. In: Mairesse, J. (Hrsg.) Estimation et sondages. Economica, Paris (1988) Einschätzung

Armitage, P., Colton, T.: Encyclopedia of Biostatistics. Wiley, New York (1998) Bio-Statistiken

Arthanari, TS, Dodge, Y.: Mathematik Programmierung in der Statistik. Wiley, New York (1981) Lineare Programmierung, Optimierung

Asimov, D.: Die große Tour: ein Werkzeug zum Betrachten mehrdimensionale Daten. SIAM J. Sci. Stat. Berechnen. 6(1), 128–143 (1985) Statistiksoftware

Atteia, O., Dubois, J.-P., Webster, R.: Geostatistical Analyse der Bodenbelastung im Schweizer Jura. Umgebung. Umweltverschmutzung. **86,** 315–327 (1994) Geostatistik, Geodaten

Bailey, DE: Wahrscheinlichkeit und Statistik: Modelle für die Forschung. Wiley, New York (1971) Probenraum

Machine Translated by Google

Balestra, P.: Calcul matriciel pour économistes.
Editions Castella, Albeuve, Schweiz (1972)
Matrix

Barlett, MS: Eine Einführung in die Stochastik
Prozesse. Cambridge University Press, Cambridge (1960)
Stochastischer Prozess

Barnard, GA, Plackett, RL, Pearson, S.: „Student“: eine statistische
Biographie von William Sealy Gosset: basierend auf Schriften
von ES Pearson. Oxford University Press, Oxford
(1990) Standard
Abweichung

Barnett, V., Lewis, T.: Ausreißer in statistischen Daten.
Wiley, New York (1984) Ausreißer

Bartlett, MS: Zur theoretischen Spezifikation von
Stichprobeneigenschaften autokorrelierter Zeitreihen.
J. Roy. Stat. Soc. Ser. B 8, 27–41 (1946) Serielle
Korrelation

Bartlett, MS: Eigenschaften von Suffizienz und statistische Tests.
Proz. Roy. Soc. London. Ser. A **160,** 268–282 (1937)
Analyse kategorialer Daten

Bartlett, MS: Einige Beispiele statistischer Forschungsmethoden in
der Landwirtschaft und angewandten Biologie.
J. Roy. Stat. Soc. (Suppl.) 4, 137–183 (1937)
Chi-Quadrat-Test, Kovarianzanalyse, fehlende Daten

Basset, G., Koenker, R.: Asymptotische Theorie der Regression
des kleinsten absoluten Fehlers. Marmelade. Stat. Assoc. **73,**
618–622 (1978) L1- Schätzung

Bates, DM, Watts DG: Nichtlineare Regression
Analyse und ihre Anwendungen. Wiley, New York (1988)
Nichtlineare Regression

Bayes, T.: Ein Aufsatz zur Lösung eines Problems in der Zufallslehre.
Philos. Trans. Roy. Soc. London.
53, 370–418 (1763). Veröffentlicht auf Betreiben von Price, R.,
zwei Jahre nach seinem Tod. Neuveröffentlicht mit einer
Biographie von Barnard, George A. im Jahr 1958 und in
Pearson, ES, Kendall, MG: Studies in the
Geschichte der Statistik und Wahrscheinlichkeit. Griffin,
London, S. 131–153 (1970) Satz von Bayes,
Inferenzstatistik

Bealle, EML: Einführung in die Optimierung. Wiley, Operations
New York (1988) Research Becker, RA,
Chambers, JM, Wilks, AR: Die neue S-Sprache. Waldsworth und
Brooks/Cole, Pacific Grove, CA (1988)
Statistische Software

Belsley, D., Kuh, E., Welsch, RE: Regression
Diagnostik Identifizieren einflussreicher Daten und Quellen der
Kollinearität. Wiley, New York (1976)
Multikollinearität

Belsley, DA, Kuh, E., Welsch, RE: Regressionsdiagnostik.
Wiley, New York, S. 16–19 (1980)
Hutmatrix

Beniger, JR, Robyn, DL: Quantitative Grafiken in
Statistik: eine kurze Geschichte. Bin. Stat. **32,** 1–11 (1978)
Quantitativer Graph

Benzécri, JP: Histoire et préhistoire de l'analyse des données, les
cahiers de l'analyse des données 1, no. 1–4. Dunod, Paris
(1976) Korrespondenz
Analyse, Datenanalyse

Benzécri, JP: Geschichte und Vorgeschichte der Données-Analyse.
Dunod, Paris (1982) Probability

Benzécri, JP: L'Analyse des données. Bd. 1: La Taxinomie.
Bd. 2: L'Analyse Factorielle des Correspondances.
Dunod, Paris (1976)
Korrespondenzanalyse, Datenanalyse Bernardo, JM,
Smith, AFM: Bayesianische Theorie.
Wiley, Chichester (1994) Bayesianische Statistik

Bernoulli, D.: Quelle ist die Ursache für die Neigung der
Planeten (. . .). Empf. Stückbericht. Prix Acad. Roy.
Wissenschaft. 3, 95–122 (1734)
Hypothesentest Bernoulli,
D.: Die wahrscheinlichste Wahl zwischen mehreren
widersprüchlichen Beobachtungen, Übersetzungen und
Kommentaren (1778). In: Pearson, ES und Kendall,
MG (Hrsg.) Studies in the History of Statistics and
Probability. Griffin, London (1970)
Einschätzung

Bernoulli, J.: Ars Conjectandi, Opus Posthumum.
Accedit Tractatus de Seriebus infinitis und Epistola Gallice
scripta de ludo Pilae recticularis. Impensis Thurnisiorum,
Fratrum, Basel (1713) Bernoulli
Versuch, Satz von Bernoulli, Binomial
Verteilung, Schätzung, Gesetz der Großen
Zahlen, Normalverteilung

Berry, DA: Statistik, eine bayesianische Perspektive.
Wadsworth, Belmont, Kalifornien (1996)
Bayesianische Statistik

Bickel, RD: Auswahl eines optimalen Ablehnungsbereichs für
Mehrfachtests. Arbeitspapier, Büro für Biostatistik und
Bioinformation, Medical College of Georgia (2002)
Ablehnungsregion

Bienaymé, IJ: Überlegungen zum Antrag
Entdecken Sie Laplace auf der Grundlage der
Wahrscheinlichkeitsrechnung anhand der Methode der
geringsten Erfolge. Comptes Rendus de l'Académie des Sciences,
Paris **37,** 5–13 (1853); 1867 im Journal de Liouville vor
dem Preuve des Inégalités de Bienaymé-
Chebyshev, Journal de Mathématiques, veröffentlicht

Machine Translated by Google

Pures et Appliquées **12,** 158–176 Law of Large Zahlen

Billingsley, P.: Statistische Inferenz für Markov Prozesse. University of Chicago Press, Chicago (1961) Stochastischer Prozess

Birkes, D., Dodge, Y., Seely, J.: Spanning-Sets für schätzbare Kontraste in Klassifikationsmodellen. Ann. Stat. 4, 86–107 (1976) Fehlende Daten

Birkes, D., Dodge, Y.: Alternative Methoden der Regression. Wiley, New York (1993) Kollinearität, Regression der kleinsten absoluten Abweichung, standardisierte Daten Bishop,

YMM, Fienberg, SE, Holland, PW: Diskrete multivariate Analyse: Theorie und Praxis. MIT Press, Cambridge, MA (1975) Binärdaten, verallgemeinerte lineare Regression, Logistische Regression

Bjerhammar, A.: Rechteckige reziproke Matrizen unter besonderer Berücksichtigung geodätischer Berechnungen. Stier. Geod. **20,** 188–220 (1951) Verallgemeinerte Umkehrung

Bock, RD: Multivariate statistische Methoden in Verhaltensforschung. McGraw-Hill, New York (1975) Kollinearität

Boscovich, RJ und Maire, C.: De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus. Palladis, Rom (1755) Model Boscovich, RJ: De Litteraria Expeditione

per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex exemplaria etiam sensorum impressa. Bononiensi Scientiarium et Artium Instituto Atque Academia Commentarii, vol. IV, S. 353–396 (1757)

L1- Schätzung, Median

Bourbonnais, R.: Econométrie, manuel et exercices corrigés, 2. Aufl. Dunod, Paris (1998) Autokorrelation, Durbin-Watson-Test

Boursin, J.-L.: Les structure du hasard. Seuil, Pariser (1986) Datensammlung

Bowerman, BL, O'Connel, RT: Zeitreihen und Prognosen: Ein angewandter Ansatz. Duxbury, Belmont, Kalifornien (1979) Unregelmäßige Variation

Bowley, AL: Ansprache des Präsidenten an die Wirtschaft Sektion der British Association. J. Roy. Stat. Soc. **69,** 540–558 (1906) Konfidenzintervall, Stichprobe

Box, GEP, Draper, NR: Eine Grundlage für die Auswahl des Reaktionsflächendesigns. Marmelade. Stat. Assoc. **54,** 622–654 (1959) Kriterium des Gesamtmittelwerts Quadratische

Fehlerbox , GEP, Hunter, WG, Hunter, JS: Statistiken für Experimentatoren. Eine Einführung in Design, Datenanalyse und Modellbildung. Wiley, New York (1978) Experiment, Haupteffekt, Standard Fehler

Box, GEP, Jenkins, GM: Zeitreihenanalyse: Prognose und Kontrolle (Reihen in der Zeitreihenanalyse). Holden Day, San Francisco (1970) ARMA-Modelle, Autokorrelation, zyklisch Fluktuation, serielle Korrelation

Box, GEP, Tiao, GP: Bayesianische Inferenz in Statistische Analyse. Addison-Wesley, Reading, MA (1973) Bayesian Statistics Box,

GEP: Nichtnormalität und Varianztests. Biometrie **40,** 318–335 (1953) Robust Einschätzung

Brown, LD, Olkin, I., Sacks, J., Wynn, HP: Jack Carl Kiefer, Collected Papers III: Design of Experiments. Springer, Berlin Heidelberg New York (1985) Optimales Design

Carli, GR: Del valore etc. In: Opere Scelte di Carli, vol. Ich, S. 299. Custodi, Mailand (1764) Index Nummer

Carlin, B., Louis, TA: Bayes & Empirical Bayes Methoden. Chapman & Hall /CRC, London (2000) Bayesianische Statistik

Casella, G., Berger, RL: Statistische Inferenz. Duxbury, Pacific Grove (2001) Gemeinsam Wahrscheinlichkeitsverteilungsfunktion

Cauchy, AL: Über die Ergebnisse früherer Beobachtungen aus meiner Natur und über die Ergebnisse mit mehr Wahrscheinlichkeiten. CR Acad. Wissenschaft. **37,** 198–206 (1853) Cauchy-Verteilung Celeux,

G., Diday, E., Govaert, G., Lechevallier, Y., Ralambondrainy, H.: Classification automatique des données – Aspekte der Statistik und Informatik. Dunod, Paris (1989) Cluster Analysis, Measure of Dissimilarity

Chambers, JM und Hastie, TK (Hrsg.): Statistical Modelle in S. Waldsworth und Brooks/Cole, Pazifik Grove, Kalifornien (1992) Statistiksoftware

Chambers, JM, Cleveland, WS, Kleiner, B., Tukey, PA: Grafische Methoden zur Datenanalyse. Wadsworth, Belmont, CA (1983) QQ-Diagramm (Quantil-zu-Quantil-Diagramm)

Chambers, JM: Computergestützte Methoden zur Datenanalyse. Wiley, New York (1977) Algorithmus

Machine Translated by Google

Chatfield, C.: Die Analyse von Zeitreihen: An
Einleitung, 4. Aufl. Chapman & Hall (1989)
Autokorrelation

Chatfield, C.: Die Analyse von Zeitreihen. Eine Einleitung.
Chapman'n'Hall (2003) Säkularen
Trend

Chatterjee, P., Hadi, AS: Regressionsanalyse anhand von
Beispielen. Wiley, New York (2006) Multiple
lineare Regression

Cochran, WG: Sampling Techniques, 2. Aufl.
Wiley, New York (1963) Stichprobengröße,
einfache Zufallsstichprobe, Standardfehler

Collatz, L., Wetterling, W.: Optimierungsprobleme.
Springer, Berlin Heidelberg New York (1975)
Unternehmensforschung

Cook, RD, Weisberg, S.: Eine Einführung in
Regressionsgrafiken. Wiley, New York (1994)
Analyse von Residuen

Cook, RD, Weisberg, S.: Angewandte Regression
Einschließlich Computer und Grafik. Wiley, New York (1999)
Analyse von Residuen

Cook, RD, Weisberg, S.: Residuen und Einfluss auf die Regression.
Chapman & Hall, London (1982)
Analyse von Residuen

Copeland, MT: Statistische Indizes der Geschäftsbedingungen.
QJ Econ. **29,** 522–562 (1915)
Zeitfolgen

Cornfield, J.: Eine Methode zur vergleichenden Schätzung
Raten aus klinischen Daten. Anwendungen bei Lungen-, Brust-
und Gebärmutterhalskrebs. J. Natl. Krebsinst. **11,** 1269–75
(1951) Zurechenbares Risiko,
Vermeidbares Risiko, Prävalenzrate, relativ
Risiko, Risiko

Cornfield, J.: Methoden zur Stichprobenziehung menschlicher
Populationen. Die Bestimmung der Stichprobengröße.
Amer. J. Publ. Gesundheit **41,** 654–661 (1951)
Probengröße

Cotes, R.: Aestimatio Errorum in Mixta Mathesi, per Variationes
partium Trianguli plani et sphaerici. In: Smith, R. (Hrsg.) Opera
Miscellania, Cambridge (1722)
Gewichtetes arithmetisches Mittel

Cox, DR, Box, GEP: Eine Analyse von
Transformationen (mit Diskussion). J. Roy. Stat. Soc.
Ser. B **26,** 211–243 (1964) Transformation

Cox, DR, Hinkley, DV: Theoretische Statistik.
Chapman & Hall, London (1973) Wahrscheinlichkeit
Verhältnistest

Cox, DR, Miller, HD: Theorie der Stochastik
Prozesse. Wiley, New York; Methuen, London (1965)
Stochastischer Prozess

Cox, DR, Reid, N.: Theorie des Designs von
Experimente. Chapman & Hall, London (2000)
Einseitige Varianzanalyse, zweiseitig
Varianzanalyse Cox, DR,

Snell, EJ: Analysis of Binary Data, 2. Aufl. Chapman & Hall,
London (1990)
Analyse kategorialer Daten Cox, DR,

Snell, EJ: Die Analyse binärer Daten.
Chapman & Hall (1989) Analyse von Binärdaten,
Binärdaten, logistische Regression

Cox, DR: Schätzung durch Doppelstichprobe.
Biometrie **39,** 217–227 (1952) Stichprobengröße

Cox, DR: Interaktion. Int. Stat. Rev. **52,** 1–25 (1984)
Interaktion

Cox, PR: Demographie, 5. Aufl. Cambridge University
Press, Cambridge (1976)
Demographie

Cressie, NAC: Statistik für räumliche Daten. Wiley, Geostatistik,
New York (1991) Räumlichkeit
Statistiken

Crome, AFW: Geographisch-statistische
Darstellung der Staatskräfte. Weygand, Leipzig (1820)
Grafische Darstellung Crome,

AFW: Über die Grösse und Bevölkerung der sämtlichen
Europäischen Staaten. Weygand, Leipzig (1785)
Grafische Darstellung

Dénes, J., Keedwell, AD: Lateinische Quadrate und ihre
Anwendungen. Akademisch, New York (1974)
Graeco-Latin Square Design Dantzig,

GB: Anwendungen und Erweiterungen der linearen
Programmierung. Dunod, Paris (1966)
Lineares Programmieren

David, FN: Würfeln und Spielen (eine Anmerkung zur Geschichte
der Wahrscheinlichkeit). Biometrie **42,** 1–15 (1955)

Wahrscheinlichkeit Davison, AC, Hinkley, DV: Bootstrap-Methoden
und ihre Anwendung. Cambridge University Press, Cambridge
(1997) Bootstrap

DeLury, DB: Die Analyse der Kovarianz. Biometrische
4, 153–170 (1948) Kovarianzanalyse

Delanius, T.: Elements of Survey Sampling. SAREC, Sampling
Stockholm (1985) Deming,

WE: Sample Design in Business Research.
Wiley, New York (1960) Systematic Sampling

Desrosières, A. Der Teil für alles: Allgemeiner Kommentar?
Die Vorgeschichte der Kontraindikation

Machine Translated by Google

repräsentativität. Schätzung und Sondagen. Fünf Beiträge zur Statistikgeschichte. Economica, Paris (1988) Konfidenzintervall, Stichprobe

Deuchler, G.: Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. Zeitung für Pädagogische Psychologie und Experimentelle Pädagogik, **15,** 114–131, 145–159, 229–242 (1914) Kendall Rangkorrelationskoeffizient

Diggle, PJ, P. Heagerty, K.-Y. Liang und SL Zeger.: Analyse von Längsschnittdaten, 2. Aufl. Oxford University Press, Oxford (2002) Panel

Dixon, WJ, Chefredakteur: BMDP Statistical Software Communications. Die University of California Press. Berkeley, Kalifornien (1981) Statistiksoftware

Dodge, Y. (2004) Optimierungsapplikation. Springer, Paris Lagrange Multiplier Dodge, Y.: Ein natürlicher Zufallszahlengenerator, Int. Stat. Review, 64, 329–344 (1996) von Random Numbers Generation

Dodge, Y.: Premiers pas en statistique. Springer, Paris (1999) Maß der zentralen Tendenz, Maß der Streuung

Dodge, Y. (Hrsg.): Statistische Datenanalyse basierend auf der L1-Norm und verwandten Methoden. Elsevier, Amsterdam (1987) L1- Schätzung

Dodge, Y., Hand, D.: Wie sollte zukünftige Statistiksoftware aussehen? Berechnen. Stat. Daten Anal. (1991) Statistiksoftware

Dodge, Y., Kondylis, A., Whittaker, J.: Erweiterung der PLS1- auf die PLAD-Regression und die Verwendung der L1-Norm bei der Softmodellierung. COMPSTAT 2004, S. 935–942, Physica/Verlag-Springer (2004) Partielle Regression der kleinsten absoluten Abweichung

Dodge, Y., Majumdar, D.: Ein Algorithmus zum Finden kleinster quadratischer verallgemeinerter Inversen für Klassifizierungsmodelle mit beliebigen Mustern. J. Stat. Berechnen. Simul. 9, 1–17 (1979) Verallgemeinerte Inverse, fehlende Daten

Dodge, Y., Rousson, V.: Multivariater L1-Mittelwert. Metrika **49,** 127–134 (1999) Norm eines Vektors

Dodge, Y., Shah, KR: Schätzung von Parametern in lateinischen Quadraten und griechisch-lateinischen Quadraten mit fehlenden Beobachtungen. Komm. Stat. Theorie A6(15), 1465–1472 (1977) Griechisch-lateinisches quadratisches Design

Dodge, Y., Thomas, DR: Zur Leistung nichtparametrischer und normaler Theoriemultiplikatoren Vergleichsverfahren. Sankhya **B42,** 11–27 (1980) Test auf kleinste signifikante Unterschiede

Dodge, Y., Whittaker, J. (Hrsg.): Computational Statistiken, Bde. 1 und 2. COMPSTAT, Neuchâtel, Schweiz, 24.–28. August 1992 (Englisch). Physica-Verlag, Heidelberg (1992) Statistisch Software

Dodge, Y., Whittaker, J., Kondylis, A.: Fortschritte bei PLAD und PQR. In: Proceedings 55th Session of the ISI 2005, S. 495–503 (2005) Teilweise zumindest Absolute Abweichungsregression

Dodge, Y.: Analyse der angewandten Regression. Dunod, Paris (1999) Kriterium des Gesamtmittelwerts im Quadrat Fehler

Dodge, Y.: Analyse von Experimenten mit fehlenden Daten. Wiley, New York (1985) Fehlende Daten

Dodge, Y.: Basismathematik für Wirtschaftswissenschaftler. Springer, Berlin Heidelberg New York (2002) Eigenvektor, Norm eines Vektors

Dodge, Y.: Wichtigste Erfahrungspläne. In: Aeschlimann, J., Bonjour, C., Stocker, E. (Hrsg.) Methoden und Techniken von Erfahrungsplänen: 28. Kurs der Perfektionierung durch die Vaudoise des Chercheurs en Association Körperbau, Saas-Fee, 2.–8. März 1986 (1986) Faktorielles Experiment

Dodge, Y.: Einige Schwierigkeiten bei der nichtparametrischen Schätzung einer Dichtefunktion. J. Offic. Stat. 2(2), 193–202 (1986) Histogramm

Doob, JL: Stochastische Prozesse. Wiley, New York (1953) Stochastischer Prozess

Doob, JL: Was ist ein stochastischer Prozess? Bin. Mathematik. Monatlich **49,** 648–653 (1942) Stochastisch Verfahren

Draper, NR, Smith, H.: Angewandte Regression Analyse, 3. Aufl. Wiley, New York (1998) Residuenanalyse, nichtlineare Regression

Drobisch, MW: Über Mittelgrössen und die Anwendbarkeit derselben auf die Berechnung des Steigens und Sinkens des Geldwerts. Berichte über die Verhandlungen des Königs. Sachs. Ges. Wiss. Leipzig. Mathe-Phy. Klasse, 23, 25 (1871) Indexnummer

Droesbeke, JJ, Tassi, PH: Histoire de la Statistique. Editionen „Que sais-je?“ Presses universitaires de France, Paris (1990) Geometrisches Mittel

Dufour, IM: Histoire de l'analyse de Séries Chronologues, Université de Montréal, Kanada (2006) Zeitreihen

Machine Translated by Google

Duncan, GJ: Haushaltspanelstudien: Perspektiven und Probleme. Sozialwissenschaftliche Methodologie, Trient, Italien (1992) Panel

Duncan, JW, Gross, AC: Statistiken für das 21. Jahrhundert: Vorschläge zur Verbesserung der Statistiken für eine bessere Entscheidungsfindung. Irwin Professional Publishing, Chicago (1995) Offizielle Statistik

Dupâquier, J. und Dupâquier, M.: Histoire de la demographie. Académie Perrin, Paris (1985) Offizielle Statistik

Durbin, J., Knott, M., Taylor, CC: Komponenten der Cramer-Von-Mises-Statistik, II. J. Roy. Stat. Soc. Ser. B **37,** 216–237 (1975) Anderson–Liebling Prüfen

Durbin, J., Watson, GS: Testen auf serielle Korrelation in der Regression der kleinsten Quadrate, I. Biometrika **37,** 409–428 (1950) Durbin-Watson-Test

Durbin, J., Watson, GS: Testen auf serielle Korrelation in der Regression der kleinsten Quadrate, II. Biometrie **38,** 159–177 (1951) Durbin-Watson-Test, Seriell Korrelation

Durbin, J.: Alternative Methode zum D-Test. Biometrischer **56,** 1–15 (1969) Durbin-Watson-Test

Dutot, C.: Politische Überlegungen zu Finanzen und Handel. Vaillant und Nicolas Prevost, The Den Haag (1738) Indexnummer

Edgeworth FY: Eine neue Methode zur Reduzierung Beobachtungen, die sich auf mehrere Größen beziehen. Philos. Mag. (5. Ser) **24,** 222–223 (1887) L1 -Schätzung

Edgeworth, FY: Methoden der Statistik. Jubiläum Band der Royal Statistical Society, London (1885) Hypothesentests

Edington, ES: Randomisierungstests. Marcel Dekker, New York (1980) Resampling

Edwards, AWF: Wahrscheinlichkeit. Ein Bericht über die statistisches Konzept der Wahrscheinlichkeit und seine Anwendung auf wissenschaftliche Schlussfolgerungen. Cambridge University Press, Cambridge (1972) Likelihood-Ratio-Test

Edwards, D.: Einführung in die grafische Modellierung. Springer, Berlin Heidelberg New York (1995) Statistiksoftware

Efron, B., Tibshirani, RJ: Eine Einführung in die Bootstrap. Chapman & Hall, New York (1993) Bootstrap

Efron, B.: Bootstrap-Methoden: Ein weiterer Blick auf das Klappmesser. Ann. Stat. 7, 1–26 (1979) Bootstrap

Efron, B.: Das Jackknife, der Bootstrap und die anderen Resampling-Pläne. Gesellschaft für Industrie und Angewandte Mathematik, Philadelphia (1982) Jackknife-Methode

Efron, B.: Das Klappmesser, der Bootstrap und andere Resampling-Pläne, CBMS Monograph. Nr. 38. SIAM, Philadelphia (1982) Resampling

Eggenberger, F. und Polya, G.: Über die Statistik verketteter Vorgänge. Zeitschrift für angewandte Mathematische Mechanik 3, 279–289 (1923) Negative Binomialverteilung

Eisenhart, C.: Boscovich und die Kombination von Beobachtungen. In: Kendall, M., Plackett, RL (Hrsg.) Studien zur Geschichte der Statistik und Wahrscheinlichkeit, vol. II. Griffin, London (1977) Regressionsanalyse

Elderton, WP: Tabellen zum Testen der Anpassungsfähigkeit der Theorie an die Beobachtung. Biometrie 1, 155–163 (1902) Chi-Square Table

Euler, L.: Recherches sur la questions of inégalités du movement of Saturne and de Jupiter, Teil des Jahres 1748, von der Académie Royale des Sciences de Paris. Neuauflage 1960, von Leonhardi Euleri, Opera Omnia, 2. Serie. Turici, Basel, 25, S. 47–157 (1749) Analyse von Residuen, Median, Modell

Euler, L.: Recherches sur une nouvelle espèce de carrés magiques. Verh. Zeeuw. General Weten. Vlissengen 9, 85–239 (1782) Griechisch-lateinisches quadratisches Design

Evelyn, Sir GS: Ein Bericht über einige Bemühungen, einen Standard für Gewicht und Maß festzulegen. Philos. Trans. 113 (1798) Indexnummer

Everitt, BS: Clusteranalyse. Halstead, London -Klassifikation, (1974) Clusteranalyse, Datenanalyse

Fechner, GT: Kollektivmasslehre. W. Engelmann, Leipzig (1897) Kendall-Rangkorrelation Koeffizient

Federer, WT, Balaam, LN: Bibliographie zu Experiment- und Behandlungsdesign vor 1968. Hafner, New York (1973) Versuchsplanung

Federer, WT: Datenerfassung. In: Kotz, S., Johnson, NL (Hrsg.) Encyclopedia of Statistical Sciences, vol. 2. Wiley, New York (1982) Daten

Fedorov, VV: Theorie optimaler Experimente. Akademisch, New York (1972) Optimales Design

Festinger, L.: Die Bedeutung der Mittelwertdifferenz ohne Bezug zur Häufigkeit

Machine Translated by Google

Verteilungsfunktion. Psychometrika **11,** 97–105 (1946) Mann-Whitney-Test Fienberg, SE:
Grafische Methode in der Statistik. Bin. Stat. **33,** 165–178 (1979) Grafisch Darstellung
Fienberg, SE: Die Analyse von Cross-Classified Kategoriale Daten, 2. Aufl. MIT Press, Cambridge, MA (1980) Kontingenztabelle
Finetti, B. de: Wahrscheinlichkeitstheorie; Eine kritische Einführungsbehandlung. Trans. Machi, A., Smith, A. Wiley, New York (1975) Bayesianische Statistik
Fisher, I.: Die Entstehung von Indexzahlen. Houghton Mifflin, Boston (1922) Fisher-Index
Fisher, RA, Mackenzie, WA: Das Manurial Reaktion verschiedener Kartoffelsorten. J. Landwirtschaft. Wissenschaft. **13,** 311–320 (1923) Nichtlineare Regression
Fisher, RA, Yates, F.: Statistische Tabellen für Biologische, landwirtschaftliche und medizinische Forschung, 6. Aufl. Hafner (Macmillan), New York (1963) Griechisch-lateinisches Quadratdesign Fisher, RA, Yates, F.: Statistische Tabellen für Biologische, landwirtschaftliche und medizinische Forschung. Oliver und Boyd, Edinburgh und London (1963) Fisher-Tisch
Fisher, RA: Anwendungen der „Student"-Verteilung. Metron 5, 90–104 (1925) Freiheitsgrad, Studentenverteilung, Studententabelle, Studententest
Fisher, RA: Momente und Produktmomente von Stichprobenverteilungen. Proz. London. Mathematik. Soc. **30,** 199–238 (1929) Methode der Momente
Fisher, RA: Über ein absolutes Anpassungskriterium Frequenzkurven. Durcheinander. Mathematik. **41,** 155–160 (1912) Schätzung, maximale Wahrscheinlichkeit
Fisher, RA: Zur Interpretation von ÿ2 aus Notfalltabellen und die Berechnung von PJ Roy. Stat. Soc. Ser. A **85,** 87–94 (1922) Chi-Quadrat-Test der Unabhängigkeit
Fisher, RA: Auf den mathematischen Grundlagen von Theoretische Statistik. Philos. Trans. Roy. Soc. London. Ser. A **222,** 309–368 (1922) Schätzung, nichtlineare Regression
Fisher, RA: Statistische Methoden und wissenschaftliche Schlussfolgerungen. Oliver & Boyd, Edinburgh (1956) Inferenzstatistik
Fisher, RA: Statistische Methoden für Forscher. Oliver & Boyd, Edinburgh (1925) Analyse von Varianz, Block, Kovarianz Analyse, Fisher-Verteilung, Interaktion, lateinische Quadratdesigns, Randomisierung, bidirektionale Varianzanalyse, Varianz Fisher, RA: Statistische
Methoden, experimentelles Design und wissenschaftliche Schlussfolgerung. Eine Neuauflage von Statistical Methods for Research Workers, „The Design of Experiments" und „Statistical Methods and Scientific Inference". Bennett, JH, Yates, F. (Hrsg.) Oxford University Press, Oxford (1990) Randomisierung
Fisher, RA: Das Design von Experimenten. Oliver & Boyd, Edinburgh (1935) Hypothese, Interaktion, Test auf kleinste signifikante Differenz, Randomisierung, Resampling, Varianz Fisher, RA:
Die Anordnung von Feldexperimenten. J. Landwirtschaftsministerium. **33,** 503–513 (1926) Versuchsplanung, Randomisierung
Fisher, RA: Die Korrelation zwischen Verwandten unter der Annahme einer Mendelschen Vererbung. Trans. Roy. Soc. Edinb. **52,** 399–433 (1918) Varianz
Fisher, RA: Die Momente der Verteilung für Normalstichproben von Maßen für die Abweichung von der Normalität. Proz. Roy. Soc. London. Ser. A **130,** 16–28 (1930) Methode der Momente Fisher, RA: Die Präzision
von Diskriminanzfunktionen. Ann. Eugen. (London) **10,** 422–429 (1940) Korrespondenzanalyse, Datenanalyse
Fisher, RA: Theorie der statistischen Schätzung. Proz. Camb. Philos. Soc. **22,** 700–725 (1925) Einschätzung
Fisher-Box, J.: Guinness, Gosset, Fisher und klein Proben. Stat. Wissenschaft. 2, 45–52 (1987) Gosset, William Sealy
Fisher-Box, J.: RA Fisher, das Leben eines Wissenschaftlers. Wiley, New York (1978) Fischer, Ronald Aylmer
Fitzmaurice, G., Laird, N., Ware, J.: Applied Längsschnittanalyse. Wiley, New York (2004) Längsschnittdaten
Flückiger, Y.: Analyse social-économique des Unterschiede in den Schweizer Kantonen. Le chômage en Suisse: bilan et perspectives. Forum Helvet. 6, 91–113 (1995) Gini-Index
Fleetwood, W.: Chronicon Preciosum, 2. Aufl. Fleetwood, London (1707) Indexnummer
Fletcher, R.: Praktische Methoden der Optimierung. Wiley, New York (1997) Nichtlineare Regression
Foody, WN, Hedagat, AS: Zur Theorie und Anwendung von BIB-Designs mit wiederholten Blöcken.

Machine Translated by Google

Annals of Statistics, V=S 932–945 (1977)
Sampling

Francis, I.: Statistische Software: eine vergleichende Übersicht.
Nordholland, New York (1981)
Statistiksoftware

Frank, I., Friedman, J.: Eine statistische Sicht auf einige
Chemometrie-Regressionstools. Technometrics **35,** 109–135
(1993) Partielle Regression der
kleinsten Quadrate

Fredholm, J.: Über eine Klasse von Gleichungen
Funktionsmerkmale. Acta Math. **27,** 365–390 (1903)
Verallgemeinerte Umkehrung

Freedman, D., Pisani, R., Purves, R.: Statistik.
Norton, New York (1978) Histogramm

Fuller, AW: Einführung in statistische Zeitreihen.
John Wiley'n'Sas (1976) Saisonale Unterschiede

Funkhauser, HG: Eine Anmerkung zu einer Grafik aus dem 10. Jahrhundert.
Osiris 1, 260–262 (1936) Zeitreihe

Galton, F.: Zusammenhänge und ihre Messung,
hauptsächlich aus anthropologischen Daten. Proz. Roy. Soc.
London. **45,** 135–145 (1888) Korrelation
Koeffizient, genetische Statistik

Galton, F.: Verwandtschaft und Korrelation. North Am. Rev. **150,**
419–431 (1890) Korrelationskoeffizient

Galton, F.: Erinnerungen an mein Leben. Methuen, London
(1908) Korrelationskoeffizient

Galton, F.: Natürliche Vererbung. Macmillan, London
(1889) Korrelationskoeffizient, Inferenz
Statistik, Regressionsanalyse

Galton, F.: Das geometrische Mittel in der Vital- und
Sozialstatistik. Proz. Roy. Soc. London. **29,** 365–367
(1879) Lognormalverteilung Galton,

F.: Typische Gesetze der Vererbung. Proz. Roy. Inst.
Great Britain 8, 282–301 (1877) (nachgedruckt in: Stigler,
SM (1986) The History of Statistics: The Measurement of
Uncertainty Before 1900. Belknap Press, Cambridge, MA, S. 280)
Genetisch
Statistiken

Gauss, CF: Bestimmung der Genauigkeit der
Beobachtungen, Bd. 4, S. 109–117 (1816). In: Gauss, CF
Werke (erschienen 1880).
Dieterichsche Universitäts-Druckerei, Göttingen
Normalverteilung, Standardfehler

Gauss, CF: Gauß' Werk (1803–1826) über die Theorie der kleinsten
Quadrate. Trotter, HF, trans. Statistical Techniques Research
Group, Technical Report, Nr. 5, Princeton University Press,
Princeton, NJ (1957)
Standart Fehler

Gauss, CF: In: Gauss' Werk (1803–1826) über die Theorie der
kleinsten Quadrate. Trans. Trotter, HF
Statistical Techniques Research Group, Technischer Bericht
Nr. 5. Princeton University Press, Princeton, NJ
(1821) Mean Squared Error

Gauss, CF: Méthode des Moindres Carrés.
Erinnerungen an die Kombination von Beobachtungen.
Französische Übersetzung von J. Bertrand.
Mallet-Bachelier, Paris (1855)

Gauss, CF: Theoria Combinationis Observationum
Erroribus Minimis Obnoxiae, Teile 1, 2 und Ergänzungen.
Werke 4, 1–108 (1821, 1823, 1826)
Gauß-Markov-Theorem

Gauss, CF: Theoria Motus Corporum Coelestium.
Werke, 7 (1809) Fehler, Schätzung,
Inferenzstatistik, Normalverteilung,
Regressionsanalyse

Gavarret, J.: Allgemeine Statistikprinzipien
medizinisch. Beché Jeune & Labé, Paris (1840)
Hypothesentest Gelman,

A., Carlin, JB, Stern HS, Rubin, DB:
Bayesianische Datenanalyse. Chapman & Hall /CRC, London
(2003) Bayesianische Statistik

Gibbens, RJ, Pratt, JW: Interpretation und Methodik der P-Werte.
Bin. Stat. 29(1), 20–25 (1975) p-Value Gibbons, JD,

Chakraborti, S.: Nonparametric Statistical Inference, 4.
Auflage. CRC (2003)
Nichtparametrische Statistik-

M-Inferenz. McGraw-Hill, New York (1971)
Wilcoxon-Test

Glasser, GJ, Winter, RF: Kritische Werte der
Rangkorrelationskoeffizient zum Testen der
Unabhängigkeitshypothese. Biometrika **48,** 444–448
(1961) Spearman-Rangkorrelationskoeffizient

Goovaerts, P.: Geostatistik für natürliche Ressourcen
Auswertung. Oxford University Press, Oxford (1997)
Geostatistik, Geodaten, Raumstatistik

Gordon, AD: Klassifizierung. Methoden zur explorativen
Analyse multivariater Daten.
Chapman & Hall, London (1981)
Klassifikation, Clusteranalyse Gosset, SW

(„Student"): Zum Zählfehler mit einem Hämacytometer. Biometrie
5, 351–360 (1907)
Negative Binomialverteilung Gosset,

SW „Student": Neue Tabellen zum Testen der Signifikanz von
Beobachtungen. Metron 5, 105–120

Machine Translated by Google

(1925) Studentenverteilung, Studententisch, Schülertest

Gosset, SW „Student“: Der wahrscheinliche Fehler von ein Mittelwert. Biometrie 6, 1–25 (1908) Inferenz Statistik, Simulation, Standardabweichung, Studentenverteilung, Studententisch, Student Prüfen

Gourieroux, C.: Théorie des sondages. Economica, Paris (1981) Survey

Gower, JC: Klassifizierung, Geometrie und Daten Analyse. In: Bock, HH (Hrsg.) Klassifizierung und verwandte Methoden der Datenanalyse. Elsevier, Amsterdam (1988) Datenanalyse

Graunt, J.: Natürliche und politische Beobachtungen in einem folgenden Index erwähnt und auf den Sterbeurkunden erstellt: mit Bezug auf die Regierung, die Religion, den Handel, das Wachstum, das Land, die Krankheiten und die verschiedenen Veränderungen der besagten Stadt. Tho. Roycroft für John Martin, James Allestry und Tho. Dicas (1662) Bio-Statistiken

Graybill, FA: Theorie und Anwendungen des linearen Modells. Duxbury, North Scituate, MA (Waldsworth und Brooks/Cole, Pacific Grove, CA) (1976) Gauß-Markov-Theorem, verallgemeinerte Linearität Regression, Matrix

Greenacre, M.: Theorie und Anwendungen von Korrespondenzanalyse. Akademisch, London (1984) Korrespondenzanalyse

Greenwald, D. (Hrsg.): Encyclopédie économique. Economica, Paris, S. 164–165 (1984) Gini-Index

Greenwood, JA, Hartley, HO: Leitfaden zu Tabellen in der mathematischen Statistik. Princeton University Press, Princeton, NJ (1962) Normal Table Greenwood, M., Yule, GU: Eine Untersuchung der Art der Häufigkeitsverteilungen, die für mehrere Ereignisse repräsentativ sind, mit besonderem Bezug auf das Auftreten mehrerer Krankheitsanfälle oder wiederholter Unfälle. J. Roy. Stat. Soc. Ser. A **83,** 255–279 (1920) Negative Binomialverteilung

Greenwood, M.: Medizinische Statistik von Graunt bis Farr. Cambridge University Press, Cambridge (1948) Bio-Statistiken

Grosbras, JM: Méthodes statistiques des sondages. Economica, Paris (1987) Umfrage

Guerry, AM: Essai sur la statistique morale de la France. Crochard, Paris (1833) Grafische Darstellung

Haberman, SJ: Analyse qualitativer Daten. Bd. I: Einführungsthemen. Akademisch, New York (1978) Analyse kategorialer Daten Hadley, G.: Lineare Programmierung. Addison-Wesley, Reading, MA (1962) Linear Programming Hall, P.: Die Bootstrap- und Edgeworth-Erweiterung. Springer, Berlin Heidelberg New York (1992) Bootstrap

Hand, DJ: Diskriminierung und Klassifizierung. Wiley, New York (1981) Maß der Unähnlichkeit

Hansen, MH, Hurwitz, WN, Madow, MG: Methoden und Theorie der Stichprobenerhebung. Bd. I. Methoden und Anwendungen. Bd. II. Theorie. Chapman & Hall, London (1953) Cluster-Sampling

Hardy, GH: Mendelsche Proportionen in einer gemischten Population. Wissenschaft **28,** 49–50 (1908) Genetisch Statistiken

Harnett, DL, Murphy, JL: Einführende statistische Analyse. Addison-Wesley, Reading, MA (1975) Unregelmäßige Variation, Stichprobengröße

Harter, HL, Owen, DB: Ausgewählte Tabellen in Mathematische Statistik, Bd. 1. Markham, Chicago (5.7 Anhang) (1970) Von Wilcoxon signierter Tisch

Harvard University: Tabellen des Kumulativs Binomiale Wahrscheinlichkeitsverteilung, Bd. 35. Annals of the Computation Laboratory, Harvard University Press, Cambridge, MA (1955) Binomialverteilung, Binomialtabelle

Harvey, AC: Die ökonometrische Analyse von Zeitreihen. Philip Allan, Oxford (Wiley, New York) Durbin-Watson-Test (1981)

Hauser, PM, Duncan, OD: Das Studium der Bevölkerung: Eine Bestandsaufnahme und Bewertung. University of Chicago Press, Chicago, IL (1959) Demographie

Hayes, AR: Statistische Software: Eine Übersicht und Kritik ihrer Entwicklung. Büro für Marineforschung, Arlington, VA (1982) Statistisch Software

Hecht, J.: L'idée du dénombrement jusqu'à la Revolution. Für eine Geschichte der Statistik, Band 1, S. 21–82. Economica/INSEE (1978) Volkszählung

Hedayat, AS, Sinha, BK: Design und Inferenz bei der Stichprobenziehung endlicher Populationen. Wiley, New York (1991) Probenahme

Helland: Zur Struktur der PLS-Regression, Comm. in Stat-Simul. und Comp., 17, 581–607 (1988) Partielle Regression der kleinsten Quadrate

Machine Translated by Google

Heyde, CE, Seneta, E.: IJ Bienaymé. Statistische Theorie erwartet. Springer, Berlin Heidelberg New York (1977) Tschebyschew, Pafnutii Lwowitsch

Hirschfeld, HO: Ein Zusammenhang zwischen Korrelation und Kontingenz. Proz. Camb. Philos. Soc. **31,** 520–524 (1935) Korrespondenzanalyse, Datenanalyse

Hoaglin, DC, Mosteller, F., Tukey, JW (Hrsg.): Robuste und explorative Datenanalyse verstehen. Wiley, New York (1983) QQ-Diagramm (Quantil-zu-Quantil-Diagramm)

Hoaglin, DC, Welsch, RE: Die Hutmatrix in Regression und ANOVA. Bin. Stat. **32,** 17–22 (und Korrektur bei **32,** 146) (1978) Hutmatrix

Hoerl AE, Kennard RW, Baldwin KF: Ridge-Regression: einige Simulationen. Komm. Stat., Bd. 4, S. 105–123 (1975) Ridge-Regression

Hoerl AE, Kennard RW: Ridge-Regression: voreingenommene Schätzung für nichtorthogonale Probleme. Technometrics **12,** 55–67 (1970) Ridge-Regression

Hoerl AE: Anwendung der Ridge-Analyse auf Regressionsprobleme. Chem. Ing. Prog. 58(3), 54–59 (1962) Ridge-Regression

Hogg, RV, Craig, AT: Einführung in die Mathematische Statistik, 2. Aufl. Macmillan, Neu York (1965) Schätzer, gemeinsame Verteilung Funktion, Randdichtefunktion, Paar Zufällige Variablen

Holgate, P.: Der Einfluss von Huygens' Arbeit in Dynamik seines Beitrags zur Wahrscheinlichkeit. Int. Stat. Rev. **52,** 137–140 (1984) Wahrscheinlichkeit

Hosmer, DW, Lemeshow S.: Angewandte logistische Regression. Wiley-Reihe in Wahrscheinlichkeit und Statistik. Wiley, New York (1989) Logistische Regression

Huber, PJ: Robuste Statistik. Wiley, New York Leverage (1981) Point, Robuste Schätzung

Huber, PJ: Robuste Standortschätzung Parameter. Ann. Mathematik. Stat. **35,** 73–101 (1964) Robuste Schätzung

Huber, PJ: Robustheit: Wo stehen wir jetzt? Student, Bd. 1, nein. 2, S. 75–86 (1995) Robust Einschätzung

Huitema, BE: Die Analyse von Kovarianz und Alternativen. Wiley, New York (1980) Kovarianzanalyse

INSEE: Für eine Geschichte der Statistik. Bd. 1: Beiträge. INSEE, Economica, Paris (1977) Offizielle Statistik

INSEE: Für eine Geschichte der Statistik. Bd. 2: Material. J. Affichard (Hrsg.) INSEE, Economica, Paris (1987) Offizielle Statistik

ISI: Die Zukunft der Statistik: Eine Internationale Perspektive. Voorburg (1994) Offizielle Statistik

Internationale mathematische und statistische Bibliotheken.: Informationen zur IMSL-Bibliothek. IMSLK, Houston (1981) Statistiksoftware

Isaaks, EH, Srivastava, RM: Eine Einführung in die angewandte Geostatistik. Oxford University Press, Oxford (1989) Geostatistik, räumliche Statistiken

Jaffé, W.: Walras, Léon. In: Sills, DL (Hrsg.) Internationale Enzyklopädie der Sozialwissenschaften, Bd. 16. Macmillan und Free Press, New York, S. 447–453 (1968) Ökonometrie

Jambu, M., Lebeaux, MO: Klassifizierung Automatische Analyse der Données. Dunod, Paris (1978) Clusteranalyse

Jevons, WS: Ein gravierender Wertverlust von Gold wurde festgestellt und seine sozialen Auswirkungen dargelegt. Stanford, London (1863) Indexnummer

Jevons, WS: Die Prinzipien der Wissenschaft: Eine Abhandlung über Logik und wissenschaftliche Methoden (zwei Bände). Macmillan, London (1874) Geometrisches Mittel

Joanes, DN & Gill, CA: Vergleich von Messungen der Probenschiefe und Kurtosis, JRSS.D 47(1), 183–189 (1998) Maß der Kurtosis

Johnson, NL, Kotz, S., Balakrishnan, N.: Discrete Multivariate Distributions, John Wiley (1997) Wahrscheinlichkeitsverteilung

Johnson, NL, Kotz, S.: Verteilungen in der Statistik: Kontinuierliche univariate Verteilungen, Bde. 1 und 2. Wiley, New York (1970) Kontinuierlich Wahrscheinlichkeitsverteilung, Lognormal Verteilung, Wahrscheinlichkeitsverteilung

Johnson, NL, Kotz, S.: Verteilungen in der Statistik: Diskrete Verteilungen. Wiley, New York (1969) Diskrete Wahrscheinlichkeitsverteilung, Wahrscheinlichkeit Verteilung

Johnson, NL, Leone, FC: Statistiken und experimentelles Design in den Ingenieurwissenschaften und den Naturwissenschaften. Wiley, New York (1964) Variationskoeffizient

Machine Translated by Google

Journel, A., Huijbregts, CJ: Bergbaugeostatistik. Akademisch, New York (1981) Geostatistik, Raumstatistik

Kaarsemaker, L., van Wijngaarden, A.: Tabellen zur Verwendung in der Rangkorrelation. Stat. Neerland. 7, 53 (1953) Kendall-Rangkorrelationskoeffizient

Kapteyn, J., van Uven, MJ: Skew-Häufigkeitskurven in Biologie und Statistik. Gebrüder Hoitsema, Groningen (1916) Lognormalverteilung

Kapteyn, J.: Skew-Häufigkeitskurven in Biologie und Statistik. Astronomisches Labor Noordhoff, Groningen (1903) Lognormalverteilung

Kasprzyk, D., Duncan, G. et al. (Hrsg.): Panel-Umfragen. New York, Chichester, Brisbane, Toronto, Singapur. Wiley, New York (1989) Panel

Kaufman, L., Rousseeuw, PJ: Gruppen finden in Daten: Eine Einführung in die Clusteranalyse. Wiley, New York (1990) Klassifizierung, Cluster Analyse

Kendall, MG, Steward, A.: The Advanced Theory of Statistics, vol. 2. Griffin, London (1967) Likelihood-Ratio-Test

Kendall, MG, Stuard, A.: Die fortgeschrittene Theorie der Statistik. Bd. 1. Verteilungstheorie, 4. Aufl. Griffin, London (1977) Methode der Momente

Kendall, MG: Ein neues Maß für die Rangkorrelation. Biometrie **30,** 81–93 (1938) Kendall-Rang Korrelationskoeffizient

Kendall, MG: Rangkorrelationsmethoden. Greif, London (1948) Kendall-Rangkorrelation Koeffizient

Kendall, MG: Studien zur Geschichte der Wahrscheinlichkeit und Statistik: II. Die Anfänge einer Wahrscheinlichkeitsrechnung. Biometrie **43,** 1–14 (1956) Wahrscheinlichkeit

Kendall, MG: Die frühe Geschichte der Indexzahlen. In: Kendall, M., Plackett, RL (Hrsg.) Studies in the History of Statistics and Probability, vol. II. Griffin, London (1977) Indexnummer

Kendall, MG: Zeitreihen. Griffin, London (1973) Kovariation, Zeitreihen

Kendall, MG: Wo soll die Geschichte der Statistik beginnen? Biometrie **47,** 447–449 (1960) Statistiken

Khwarizmi, Musa ibn Meusba (9. Jh.). Jabr wa-al-muqeabalah. Die Algebra von Mohammed ben Musa, Rosen, F. (Hrsg. und Übersetzung). Georg Olms Verlag, Hildesheim (1986) Algorithmus

Kiefer, J.: Optimale experimentelle Designs. J. Roy. Stat. Soc. Ser. B **21,** 272–319 (1959) Kiefer, Jack Carl, Optimales Design

Kleijnen, JPC: Statistische Techniken in der Simulation (in zwei Teilen). Bd. 9 der Statistik: Lehrbücher und Monographien. Marcel Dekker, New York (1974) Simulation

Kleijnen, JPC: Statistische Werkzeuge für die Simulation Praktiker. Bd. 76 der Statistik: Lehrbücher und Monographien. Marcel Dekker, New York (1987) Simulation

Knight, K.: Mathematische Statistik. Chapman & Gelenkdichte Hall/CRC, London (2000) Funktion

Kokoska, S., Zwillinger, D.: CRC-Standard Wahrscheinlichkeits- und Statistiktabellen und Formeln. CRC Press (2000) Statistische Tabelle

Kolmogorov, AN: Grundlagen der Wahrscheinlichkeitstheorie. Chelsea Publishing Company, New York (1956) Bedingte Wahrscheinlichkeit

Kolmogorov, AN: Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin Heidelberg New York (1933) Bedingt

Wahrscheinlichkeit Kolmogorov, AN: Mathe. Sbornik. NS 1, 607–610 (1936) Stochastischer Prozess

Kolmogorov, AN: Die empirische Bestimmung einer Vertriebslinie. Giornale dell'Instituto Italiano degli Attuari 4, 83–91 (6.1) (1933) Kolmogorov-Smirnov-Test, nichtparametrisch Prüfen

Kondylis, A., Whittaker, J.: Verwendung des Bootstrap für die PLAD-Regression. PLS und verwandte Methoden. In: Proceedings of the PLS'05 International Symposium 2005, S. 395–402. Edition Aluja et al. (2005) Teilweise kleinste absolute Abweichung Rückschritt

Kotz, S., Johnson, NL: Encyclopedia of Statistical Wissenschaften, Bde. 1–9. Wiley, New York (1982–1988) Statistiken

Kruskal, WH, Tanur, JM: International Encyclopedia of Statistics, Bde. 1–2. Die freie Presse, New York (1978) Statistiken

Kruskal, WH, Wallis, WA: Verwendung von Dienstgraden in Varianzanalyse mit einem Kriterium. Marmelade. Stat. Assoc. **47,** 583–621 und Errata, ebenda. **48,** 907–911 (1952) Kruskal-Wallis-Tabelle, Kruskal-Wallis-Test, Nichtparametrischer Test

Machine Translated by Google

LD Brown, I. Olkin, J. Sacks und HP Wynn (Hrsg.): Jack Karl Kiefer, Collected Papers. I: Statistische Schlussfolgerung und Wahrscheinlichkeit (1951–1963). New York (1985) Kiefer, Jack Carl LD

Brown, I. Olkin, J. Sacks und HP Wynn (Hrsg.): Jack Karl Kiefer, Collected Papers. II: Statistische Schlussfolgerung und Wahrscheinlichkeit (1964–1984). New York (1985) Kiefer, Jack Carl LD

Brown, I. Olkin, J. Sacks und HP Wynn (Hrsg.): Jack Karl Kiefer, Collected Papers. III: Versuchsplanung. New York (1985) Kiefer, Jack Carl

Lagarde, J. de: Initiation à l'analyse des données. Dunod, Paris (1983) Trägheitsmatrix

Lagrange, JL: Erinnerungen an die Verwendung der Methode die Umgebung zwischen den Ergebnissen zusätzlicher Beobachtungen betrachten; Untersuchen Sie daher die Vorteile dieser Methode zur Wahrscheinlichkeitsberechnung. Und es gibt je nach Material verschiedene Probleme. Sonstiges Taurinensia 5, 167–232 (1776) Fehler

Lancaster, HO: Vorläufer des Pearson Chi-Quadrat. Aust. J. Stat. 8, 117–126 (1966) Chi-Quadrat-Verteilung, Gamma-Verteilung

Laplace, PS de: Zweite Ergänzung zur Theorie Wahrscheinlichkeitsanalyse. Courcier, Paris (1818). S. 531–580 im vollständigen Werk von Laplace, Bd. 7. Gauthier-Villars, Paris, 1886 Robuste Schätzung

Laplace, PS de: Erinnerung an die moyenne inclinaison des orbites des comètes. Mém. Acad. Roy. Wissenschaft. Paris 7, 503–524 (1773) Hypothesentest

Laplace, PS de: Erinnerung an die Wahrscheinlichkeit des Ursachen par les événements. Mem. Acad. Roy. Wissenschaft. (von verschiedenen Wissenschaftlern präsentiert) 6, 621–656 (1774) (oder Laplace, PS de (1891) Œuvres complètes, Bd. 8. Gauthier-Villars, Paris, S. 27–65) Fehler, Schätzung, Laplace-Verteilung, Normal Verteilung, Normaltabelle

Laplace, PS: Merken Sie sich die Annäherungsformeln, die für sehr große Zahlen nützlich sind, und ihre Anwendung auf Wahrscheinlichkeiten. Mémoires de l'Académie Royale des Sciences de Paris, 10. Abgebildet in: Œuvres de Laplace **12,** 301–347 (1810) Zentraler Grenzwertsatz, De Moivre-Laplace-Theorem

Laplace, PS de: Mémoire sur les probabilités. Mem. Acad. Roy. Wissenschaft. Paris, 227–332 (1781) (oder Laplace, PS de (1891) Œuvres complètes, Bd. 9. Gauthier-Villars, Paris, S. 385–485.) Fehler, Normaler Tisch

Laplace, PS de: Sur les degrés mesurés des Meridiens, und sur les longueurs observées sur pendule. Geschichte der Königlichen Akademie der Inschriften und Schönen Künste, einschließlich der Erinnerungen an die literarischen Werke der Register dieser Akademie. Paris (1789) Regressionsanalyse

Laplace, PS de: Théorie analytique des probabilités, 3. Aufl. Courcier, Paris (1820) Mittlerer quadratischer Fehler

Laplace, PS de: Théorie analytique des probabilités, suppl. bis 3. Aufl. Courcier, Paris (1836) Gammaverteilung

Laplace, PS de: Théorie analytique des probabilités. Courcier, Paris (1812) Einschätzung, Inferenzstatistik

Laspeyres, E.: Die Berechnung einer mittleren Waarenpreissteigerung. Jahrb. Natl. Stat. **16,** 296–314 (1871) Indexnummer, Laspeyres Index

1863 und die kalifornisch-australischen Geldentdeckung seit 1848. Jahrb. Natl. Stat. 3, 81–118, 209–236 (1864) Indexnummer, Laspeyres-Index Le

Cam, LM, Yang, CL: Asymptotik in der Statistik: Einige grundlegende Konzepte. Springer, Berlin Heidelberg New York (1990) Konvergenz

Le Cam, L.: Der zentrale Grenzwertsatz um 1935. Stat. Wissenschaft. 1, 78–96 (1986) Zentrale Grenze Satz

Lebart, L., Salem, A.: Analyze Statistique des Données Textuelles. Dunod, Paris (1988) Korrespondenzanalyse

Legendre, AM: Neue Methoden für la Bestimmung der Umlaufbahnen der Kometen. Courcier, Paris (1805) Schätzung, Methode der kleinsten Quadrate, Normalgleichungen, Regressionsanalyse

Lehmann, Erich L., Casella, G.: Theorie der Punktschätzung. Springer-Verlag, New York (1998) Punktschätzung

Lehmann, EI, Romann, SP: Testing Statistical Hypothese, 3. Aufl. Springer, New York (2005) Alternative Hypothese

Lehmann, EL: Testen statistischer Hypothesen, 2. Hypothese Hrsg. Wiley, New York (1986) Testen

Machine Translated by Google

Lehmann, EL: Theory of Point Estimation, 2. Aufl. Wiley, New York (1983) Schätzer, Parameter

Lerman, LC: Klassifizierung und Ordnungsanalyse des données. Dunod, Paris (1981) Clusteranalyse

Liapounov, AM: Über einen Vorschlag der Theorie der Wahrscheinlichkeiten. Bulletin de l'Academie Imperiale des Sciences de St.-Petersbourg 8, 1–24 (1900) Zentraler Grenzwertsatz

Lilienfeld, AM, Lilienfeld, DE: Grundlagen von Epidemiologie, 2. Aufl. Clarendon, Oxford (1980) Zurechenbares Risiko, vermeidbares Risiko, Ursache und Wirkung in der Epidemiologie, Epidemiologie, Inzidenzrate, Quoten und Quotenverhältnis, Prävalenzrate, relatives Risiko, Risiko

Lindley, DV: Die Philosophie der Statistik. Statistiker **49,** 293–337 (2000) Bayesianische Statistik

Lipps, GF: Die Psychischen Massmethoden. F. Vieweg und Sohn, Braunschweig, Deutschland (1906) Kendall-Rangkorrelationskoeffizient

Lowe, J.: Der gegenwärtige Zustand Englands im Hinblick auf Landwirtschaft, Handel und Finanzen. Kelley, London (1822) Indexnummer

MacMahon, B., Pugh, TF: Epidemiologie: Prinzipien und Methoden. Little Brown, Boston, MA (1970) Zurechenbares Risiko, vermeidbares Risiko, Ursache und Wirkung in der Epidemiologie, Epidemiologie, Inzidenzrate, Quoten und Quotenverhältnis, Prävalenzrate, relatives Risiko, Risiko

MacMahon, PA: Combinatory Analysis, Bde. ICH und II. Cambridge University Press, Cambridge (1915–1916) Kombinatorische Analyse

Macauley, FR: Die Glättung von Zeitreihen. National Bureau of Economic Research, 121–136 (1930) Zeitreihen

Mahalanobis, PC: Zur verallgemeinerten Distanz in der Statistik, Proc. natn. Inst. Wissenschaft. Indien 2, 49–55 (1936) Mahalanobis-Entfernung

Mahalanobis, PC: Über Tests und Messungen von Gruppendivergenz. Teil I. Theoretische Formeln. J. Asiatic Soc. Bengalen **26,** 541–588 (1930) Mahalanobis-Entfernung

Maistrov, LE: Wahrscheinlichkeitstheorie – Eine Geschichte Skizze (übersetzt und herausgegeben von Kotz, S.). Akademisch, New York (1974) dependence

Mallows, CL: Auswahl von Variablen in einer linearen Regression: Eine grafische Hilfe. Präsentiert auf dem Zentralen Regionaltreffen des Instituts für Mathematische Statistik, Manhattan, KS, 7.–9. Mai 1964 (1964) *Cp*- Kriterium

Mallows, CL: Einige Kommentare zu *Cp*. Technometrics **15,** 661–675 (1973) *Vgl* Kriterium

Mann, HB, Whitney, DR: Bei einem Test, ob eine von zwei Zufallsvariablen stochastisch größer ist als die andere. Ann. Mathematik. Stat. **18,** 50–60 (1947) Mann-Whitney-Test, nichtparametrischer Test

Markov, AA: Die Anwendung von Funktionen setzt sich in der Wahrscheinlichkeitsrechnung fort. Kasan. Stier. 9(2), 29–34 (1899) Markow, Andrei Andrejewitsch

Markov, AA: Erweiterung der theoretischen Grenzen der Wahrscheinlichkeitsberechnung entlang einiger Kettengrößen. Mem. Acad. Wissenschaft. St. Petersburg 8, 365–397 (1908) Zentraler Grenzwertsatz

Martens, H. und Naes, T.: Multivariate Kalibrierung. Wiley, New York (1989) Teilweise kleinste Quadrate Rückschritt

Massey, FJ: Verteilungstabelle für die Abweichung zwischen zwei Stichprobenkumulationen. Ann. Mathematik. Stat. **23,** 435–441 (1952) Kolmogorov-Smirnov-Test

Matheron, G.: Die Theorie regionaler Variablen und ihre Anwendungen. Masson, Paris (1965) Geostatistik, Geodaten, Raumstatistik

Mayer, T.: Abhandlung über die Umwälzung des Monds um seine Achse und die scheinbare Bewegung der Mondflecken. Kosmographische Nachrichten und Sammlungen auf das Jahr 1748 1, 52–183 (1750) Analyse von Residuen, Median, Modell

Mayr, E., Linsley, EG, Usinger, RL: Methoden und Prinzipien der systematischen Zoologie. McGraw-Hill, New York (1953) Dendrogramm

McAlister, D.: Das Gesetz des geometrischen Mittels. Proz. Roy. Soc. London. **29,** 367–376 (1879) Lognormalverteilung

McMullen, L., Pearson, ES: William Sealy Gosset, 1876–1937. In: Pearson, ES, Kendall, M. (Hrsg.) Studien zur Geschichte der Statistik und Wahrscheinlichkeit, vol. I. Griffin, London (1970) Gosset, William Sealy

Merrington, M., Thompson, CM: Tabellen mit Prozentpunkten der invertierten Beta- Verteilung *(F)* . Biometrie **33,** 73–88 (1943) Fisher-Tisch

Machine Translated by Google

Metropolis, N., Ulam, S.: Die Monte-Carlo-Methode. Marmelade. Stat. Assoc. **44,** 335–341 (1949) Monte Carlo-Methode

Meyer, CD: Matrixanalyse und angewandte lineare Algebra, SIAM (2000) Eigenvektor

Miller, LH: Tabelle der Prozentpunkte von Kolmogorov-Statistik. Marmelade. Stat. Assoc. **31,** 111–121 (1956) Kolmogorov-Smirnov-Test

Miller, RG, Jr.: Simultaneous Statistical Inference, 2. Aufl. Springer, Berlin Heidelberg New York (1981) Kleinstsignifikanter Differenztest, einseitiger Test

Mitrinovic, DS (mit Vasic, PM): Analytisch Ungleichheiten. Springer, Berlin Heidelberg New York (1970) Harmonische Mittel

Moivre, A. de: Approximatio ad summam terminorum binomii *(a + b)n,* in seriem expansi. Supplementum II zu Miscellanae Analytica, S. 1–7 (1733). Fotografischer Nachdruck in einer seltenen Broschüre über Moivre und einige seiner Entdeckungen. Herausgegeben von Archibald, RC Isis 8, 671–683 (1926) Zentraler Grenzwertsatz, Satz von De Moivre-Laplace, Normalverteilung

Moivre, A. de: Die Lehre vom Zufall: oder, Eine Methode zur Berechnung der Wahrscheinlichkeit von Ereignissen im Spiel. Pearson, London (1718) Inferenz Statistik, Normalverteilung

Molenberghs, G., Geert, V.: Modelle für diskrete Längsschnittdatenreihen. Springer-Reihe in Statistik. Springer, Berlin Heidelberg New York (2005) Längsschnittdaten

Montgomery, DC: Design und Analyse von Experimente, 4. Aufl. Wiley, Chichester (1997) One-Way Analysis of Variance

Montmort, PR de: Essai d'analyse sur les jeux de hasard, 2. Aufl. Quillau, Paris (1713) Negative Binomialverteilungsstimmung ,

AM: Einführung in die Theorie der Statistik. McGraw-Hill, New York (1950) Nichtparametrischer

Test Moore, EH: Allgemeine Analyse, Teil I. Mem. Bin. Philos. Soc., Philadelphia, PA, S. 147–209 (1935) Verallgemeinerte Umkehrung

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Genf (1996) Zurechenbares Risiko, vermeidbares Risiko, Ursache und Wirkung in der Epidemiologie, Epidemiologie, Inzidenzrate, Quoten und Quotenverhältnis, Prävalenzrate, relatives Risiko, Risiko

Morabia, A.: L'Épidémiologie Clinique. Editionen „Que sais-je?“ Presses Universitaires de France, Paris (1996) Zuordenbares Risiko, vermeidbares Risiko, Inzidenzrate, Chancen und Chancenverhältnis, Prävalenzrate, relatives Risiko, Risiko

Mosteller, F., Tukey, JW: Datenanalyse und Regression: Ein zweiter Kurs in Statistik. Addison-Wesley, Reading, MA (1977) Korrelationskoeffizient, Hebelpunkt Mudholkar,

GS: Mehrfacher Korrelationskoeffizient. In: Kotz, S., Johnson, NL (Hrsg.) Encyclopedia of Statistical Sciences, vol. 5. Wiley, New York (1982) Korrelationskoeffizient

National Bureau of Standards.: Ein Leitfaden zu Tabellen von das Normalwahrscheinlichkeitsintegral. US-Handelsministerium. Applied Mathematics Series, 21 (1952) Normaltabelle

National Bureau of Standards.: Tabellen des Binomials Wahrscheinlichkeitsverteilung. US-Handelsministerium. Angewandte Mathematik Reihe 6 (1950) Binomialverteilung, Binomialtabelle Naylor, TH,

Balintfy, JL, Burdick, DS, Chu, K.: Computersimulationstechniken. Wiley, New York (1967) Simulation

Neyman, J., Pearson, ES: Zum Problem der effizienteste Tests statistischer Hypothesen. Philos. Trans. Roy. Soc. London. Ser. A **231,** 289–337 (1933) Hypothesentests

Neyman, J., Pearson, ES: Zur Verwendung und Interpretation bestimmter Testkriterien für Zwecke der statistischen Inferenz Teil I (1928). Nachdruck von Cambridge University Press im Jahr 1967 Signifikanzniveau

Neyman, J., Pearson, ES: Zur Verwendung und Interpretation bestimmter Testkriterien für statistische Schlussfolgerungen, Teile I und II. Biometrie 20A, 175–240, 263–294 (1928) Kritischer Wert, Hypothesentest, Inferenzstatistik, Likelihood-Ratio-Test, Typ-I-Fehler

Nordin, JA: Bestimmung der Stichprobengröße. Marmelade. Stat. Assoc. **39,** 497–506 (1944) Probengröße

OCDE: OCDE und internationale Indikatoren l'enseignement. Centre de la Recherche et l'innovation dans l'enseignement, OCDE, Paris (1992) Indikator

OCDE: Hauptindikatoren für Wissenschaft und Forschung Technologie. OCDE, Paris (1988) Indikator

Oliver, MA, Webster, R.: Kriging: eine Methode von Interpolation für geografische Informationssysteme.

Machine Translated by Google

Int. J. Geogr. Inf. Syst. 4(3), 313–332 (1990)
Geostatistik, Raumstatistik Ostle, B.:
Statistik in der Forschung: Grundlegende Konzepte und Techniken für Forscher. Iowa State College Press, Ames, IA (1954) Chi-Quadrat-Test, Kontrast
Owen, DB: Handbuch der statistischen Tabellen. (3.3, 5.7, 5.8, 5.10). Addison-Wesley, Reading, MA (1962)
Von Wilcoxon signierter Tisch
Paasche, H.: Über die Preisentwicklung der letzten Jahre nach den Hamburger Börsennotierungen. Jahrb. Natl. Stat. **23,** 168–178 (1874)
Indexnummer, Paasche-Index
Palgrave, RHI: Währung und Wertstandard in England, Frankreich und Indien usw. (1886) In: IUP (1969) Britische Parlamentspapiere: Sitzung 1886: Dritter Bericht und Abschlussbericht der Königlichen Kommission zur Depression in Handel und Industrie, mit Beweisprotokoll und Anhängen, Anhang B. Irish University Press, Shannon Index Number
Pannatier, Y.: Variowin 2.2. Software zur räumlichen Datenanalyse in 2D. Springer, Berlin Heidelberg New York (1996) Geostatistik
Parzen, E.: Zur Schätzung einer Wahrscheinlichkeitsdichtefunktion und -modus. Ann. Mathematik. Stat. **33,** 1065–1076 (1962) Nichtparametrische Statistik
Pascal, B.: Mesnard, J. (Hrsg.) Œuvres complètes. Bd. 2. Desclée de Brouwer, Paris (1970)
Arithmetisches Dreieck
Pascal, B.: Traité du Triangle Arithmétique (Hrsg. posthum. 1665), Paris (1654) Arithmetik Dreieck
Pascal, B.: Varia Opera Mathematica. D. Petri de Fermat. Tolosae (1679) Negative Binomialverteilung
Pascal, B.: Œuvres, Bde. 1–14. Brunschvicg, L., Boutroux, P., Gazier, F. (Hrsg.) Les Grands Ecrivains de France. Hachette, Paris (1904–1925)
Arithmetisches Dreieck
Pearson, ES, Hartley, HO: Biometrika Tables for Statisticians, 1 (2. Aufl.). Cambridge University Press, Cambridge (1948)
Normal Table Pearson, ES, Hartley, HO: Biometrika Tables for Statisticians, vols. I und II. Cambridge University Press, Cambridge (1966, 1972)
Schiefekoeffizient
Pearson, ES: Die Neyman-Pearson-Geschichte: 1926–34. In: David, FN (Hrsg.) Forschungsarbeiten zur Statistik: Festschrift für J. Neyman. Wiley, New York (1966)
Hypothesentest
Pearson, K., Filon, LNG: Mathematisch Beiträge zur Evolutionstheorie. IV: Zu den wahrscheinlichen Fehlern von Frequenzkonstanten und zum Einfluss der Zufallsauswahl auf Variation und Korrelation. Philos. Trans. Roy. Soc. London. Ser. A **191,** 229–311 (1898) Einschätzung
Pearson, K.: Beiträge zur mathematischen Evolutionstheorie. I. In: Karl Pearsons frühe statistische Arbeiten. Cambridge University Press, Cambridge, S. 1–40 (1948). Erstveröffentlichung 1894 als: Zur Zerlegung asymmetrischer Frequenzkurven. Philos. Trans. Roy. Soc. London. Ser. A **185,** 71–110
Schiefekoeffizient, Schätzung, Formmaß, Momentenmethode, Standardabweichung
Pearson, K.: Beiträge zur mathematischen Evolutionstheorie. II: Skew-Variation in homogenem Material. In: Karl Pearsons frühe statistische Arbeiten. Cambridge University Press, Cambridge, S. 41–112 (1948). Erstveröffentlichung 1895 in Philos. Trans. Roy. Soc. London. Ser. A **186,** 343–414 Schiefekoeffizient, Frequenzkurve, Histogramm, Formmaß
Pearson, K.: Über das Kriterium, dass ein gegebenes System von Abweichungen vom Wahrscheinlichen im Fall eines korrelierten Variablensystems so beschaffen ist, dass vernünftigerweise angenommen werden kann, dass es aus Zufallsstichproben entstanden ist. In: Karl Pearsons frühe statistische Arbeiten. Cambridge University Press, S. 339–357. Erstveröffentlichung 1900 in Philos. Mag. (5. Ser) **50,** 157–175 (1948) Chi-Quadrat-Verteilung, Chi-Quadrat-Anpassungstest, Chi-Quadrat-Unabhängigkeitstest, Inferenz Statistiken
Pearson, K.: Zur Theorie der Kontingenz und ihrer Beziehung zu Assoziation und normaler Korrelation. Drapers' Company Research Memoirs, Biometric Ser. I., S. 1–35 (1904) Analyse kategorialer Daten, Kontingenztabelle
Pearson, K.: Studien zur Geschichte der Statistik und Wahrscheinlichkeit. Biometrie **13,** 25–45 (1920). Nachdruck in: Pearson, ES, Kendall, MG (Hrsg.) Studien zur Geschichte der Statistik und Wahrscheinlichkeit, vol. I. Griffin, London Korrelationskoeffizient
Pearson, K.: Tabellen der unvollständigen Funktion. HM Stationery Office (Universität Cambridge).

Machine Translated by Google

Press, Cambridge seit 1934), London (1922)
Chi-Quadrat-Tisch

Peirce, B.: Kriterium für die Ablehnung zweifelhafter Beobachtungen. Astron. J. 2, 161–163 (1852)
Ausreißer

Penrose, R.: Eine verallgemeinerte Umkehrung für Matrizen. Proz. Camb. Philos. Soc. **51,** 406–413 (1955)
Verallgemeinerte Umkehrung

Personen, WM: Ein Index des allgemeinen Geschäfts Bedingungen. Rev. Econom. Stat. 1, 111–205 (1919)
Zeitfolgen

Petty, W.: Ein weiterer Aufsatz in politischer Arithmetik über das Wachstum der Stadt London. Kelley, New York (1682) (2. Aufl. 1963)
Bio-Statistiken

Petty, W.: Beobachtungen zu den Dublin-Bills von Sterblichkeit, 1681, und der Zustand dieser Stadt. Kelley, New York (1683) (2. Aufl. 1963) Bio-Statistiken

Petty, W.: Politische Arithmetik. In: William Petty, The Economic Writings. . . , Bd. 1. Kelley, New York, S. 233–313 (1676)
Ökonometrie

Pitman, EJG: Signifikanztests, die auf Proben aus beliebigen Populationen angewendet werden können. Beilage, J. Roy. Stat. Soc. 4, 119–130 (1937)
Resampling

Plackett, RL, Barnard, GA (Hrsg.): Student: Eine statistische Biographie von William Sealy Gosset (Basierend auf Schriften von ES Pearson). Clarendon, Oxford (1990) Gosset, William Sealy

Plackett, RL: Studien zur Geschichte der Wahrscheinlichkeit und Statistiken. In: Kendall, M., Plackett, RL (Hrsg.) Die Entdeckung der Methode der kleinsten Quadrate. Bd. II. Griffin, London (1977) Legendre, Adrien Marie, Regressionsanalyse Plackett, RL: Studien zur Geschichte der Wahrscheinlichkeit und Statistiken. VII. Das Prinzip des arithmetischen Mittels. Biometrie **45,** 130–135 (1958)
Arithmetisches Mittel

Playfair, W.: Der kommerzielle und politische Atlas. Schauplatz, London (1786) Grafisch Darstellung

Poisson, SD: Recherche über die Wahrscheinlichkeit von Urteilen im kriminellen und im zivilen Bereich. In: Procédés des Règles Générales du Calcul des Probabilités. Bachelier, Imprimeur-Libraire pour les Mathématiques, Paris (1837) Poisson-Verteilung

Poisson, SD: Über die Wahrscheinlichkeit der Ergebnisse vor mehreren Monaten. Connaissance des temps pour l'an 1827, S. 273–302 (1824) Zentrale Grenze Satz

Polyà, G.: Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentproblem. Mathematische Zeitschrift 8, 171–181 (1920) Zentraler Grenzwertsatz

Porter, TM: Der Aufstieg des statistischen Denkens, 1820–1900. Princeton University Press, Princeton, NJ (1986)
Korrelationskoeffizient

Preece, DA: Lateinische Quadrate, lateinische Würfel, Latein Rechtecke usw. In: Kotz, S., Johnson, NL (Hrsg.) Encyclopedia of Statistical Sciences, Bd. 4. Wiley, New York (1983) Latin Square Designs Py, B.: Statistique Déscriptive. Economica, Paris (1987) Kovariation

Quenouille, MH: Näherungstests der Korrelation in Zeitreihen. J. Roy. Stat. Soc. Ser. B **11,** 68–84 Jackknife-Methode (1949)

Quenouille, MH: Hinweise zur Verzerrung bei Schätzungen. Biometrie **43,** 353–360 (1956) Klappmesser Methode

Quetelet, A.: Briefe an Seine Königliche Hoheit den Großen Herzog von Sachsen-Coburg und Gotha, über die Wahrscheinlichkeitstheorie in ihrer Anwendung auf die Moral- und Politikwissenschaften. (Französische Übersetzung von Downs, Olinthus Gregory). Charles und Edwin Layton, London (1849)Chi-Quadrat-Test der Unabhängigkeit

Rüegg, A.: Probabilités et statistique. Pressen Polytechniques Romandes, Lausanne, Schweiz (1985)
Experiment

Rényi, A.: Wahrscheinlichkeitstheorie. Nordholland, Amsterdam (1970) Satz von Bernoulli Raktoe, BL Hedayat, A., Federer, WT: Factorial Designs. Wiley, New York (1981) Faktor

Rao, CR, Mitra, SK, Matthai, A., Ramamurthy, KG: Formeln und Tabellen für statistische Arbeiten. Statistical Publishing Company, Kalkutta, S. 34–37 (1966) Binomial Verteilung

Rao, CR, Mitra, SK: Verallgemeinerte Umkehrungen von Matrizen und ihre Anwendungen. Wiley, New York (1971)
Verallgemeinerte Umkehrung

Rao, CR: Fortgeschrittene statistische Methoden in Biometrische Forschung. Wiley, New York (1952) Varianzanalyse Rao, CR: Eine Anmerkung zur verallgemeinerten Umkehrung einer Matrix mit Anwendungen auf Probleme in der mathematischen Statistik. J. Roy. Stat. Soc. Ser. B **24,** 152–158 (1962)
Verallgemeinerte Umkehrung

Machine Translated by Google

Rao, CR: Berechnung der verallgemeinerten Umkehrung von Matrizen. I. Allgemeine Theorie. Sankhya **A29,** 317–342 (1967) Verallgemeinerte Umkehrung

Rao, CR: Lineare statistische Inferenz und ihre Bewerbungen, 2. Aufl. Wiley, New York (1973) Gauß-Markov-Theorem, verallgemeinerte lineare Regression

Rashed, R.: La naissance de l'algèbre. In: Noël, E. (Hrsg.) Le Matin des Mathématiciens. Belin-Radio France, Paris (1985) Algorithmus, Arithmetisches Dreieck

Reid, C.: Neyman – Aus dem Leben. Springer, Berlin Heidelberg New York (1982) Hypothesentest

Ripley, BD, Venables, WN (2000) S-Programmierung. Springer, Berlin Heidelberg New York Statistiksoftware

Ripley, BD: Gedanken zu Pseudozufallszahlengeneratoren. J. Comput. Appl. Mathematik. **31,** 153–163 (1990) Erzeugung von Zufallszahlen

Robert, CP: Le choix Bayesien. Springer, Paris Bayesian (2006) Statistics Rose, D.: Haushaltspanelstudien: ein Überblick. Innovation 8(1), 7–24 (1995) Panel

Ross, SM: Einführung in Wahrscheinlichkeitsmodelle, 8. Aufl. John Wiley, New York (2006) Bernoulli-Studie, Gelenkverteilungsfunktion, Randdichtefunktion

Rothmann, JK: Epidemiologie. Eine Einleitung. Oxford University Press (2002) Ursache und Wirkung in der Epidemiologie, Epidemiologie

Rothschild, V., Logothetis, N.: Wahrscheinlichkeitsverteilungen. Wiley, New York (1986) Wahrscheinlichkeitsverteilung

Royston, E.: Eine Anmerkung zur Geschichte des Grafischen Präsentation von Daten. In: Pearson, ES, Kendall, M. (Hrsg.) Studies in the History of Statistics and Probability, vol. I. Griffin, London (1970) Grafische Darstellung Rubin, DB: Ein nicht-iterativer Algorithmus für die Kleinste-Quadrate-Schätzung fehlender Werte in jeder Varianzanalyse. Appl. Stat. **21,** 136–141 (1972) Fehlende Daten

Saporta, G.: Wahrscheinlichkeit, Analyse von Données und Statistiken. Technip, Paris (1990) Korrespondenzanalyse Savage, IR: Bibliographie der Nichtparametrie Statistiken. Harvard University Press, Cambridge, MA (1962) Nichtparametrischer Test

Savage, R.: Bibliographie der nichtparametrischen Statistik und verwandte Themen: Einführung. Marmelade. Stat. Assoc. **48,** 844–849 (1953) Nichtparametrisch Statistiken

Scheffé, H.: Eine Methode zur Beurteilung aller Kontraste in der Varianzanalyse. Biometrie **40,** 87–104 (1953) Kontrast

Scheffé, H.: Die Varianzanalyse. Wiley, New York (1959) Varianzanalyse Schmid, CF, Schmid, SE: Handbook of Graphic Presentation, 2. Aufl. Wiley, New York (1979) Quantitatives Diagramm

Schmid, CF: Handbuch der grafischen Präsentation. Ronald Press, New York (1954) Grafisch Darstellung

Searle, SR: Matrixalgebra nützlich für die Statistik. Wiley, New York (1982) Matrix

Seber, AF, Wilt, SC: Nichtlineare Regression. Wiley, Nichtlineare New York (2003) Regression Seber, GAF (1977) Lineare Regressionsanalyse. Wiley, New York Einfache lineare Regression, gewichtete Methode der kleinsten Quadrate

Senn, St.: Cross-over-Studien in der klinischen Forschung. Wiley (1993) Behandlung

Shao, J., Tu, D.: Das Jackknife und Bootstrap. Springer, Berlin Heidelberg New York (1995) Bootstrap

Sheffé, H.: Die Varianzanalyse. C'n'H (1959) Einfaktorielle Varianzanalyse

Sheppard, WF: Neue Tabellen des Wahrscheinlichkeitsintegrals. Biometrie 2, 174–190 (1903) Normaler Tisch

Sheppard, WF: Zur Berechnung des Meisten wahrscheinliche Werte von Frequenzkonstanten für Daten, die nach äquidistanten Unterteilungen einer Skala angeordnet sind. Londoner Mathematik. Soc. Proz. **29,** 353–380 (1898) Augenblick

Sheppard, WF: Tabelle der Abweichungen von der Normalkurve. Biometrie 5, 404–406 (1907) Normaler Tisch

Sheynin, OB: Zur Geschichte einiger statistischer Verteilungsgesetze. In: Kendall, M., Plackett, RL (Hrsg.) Studies in the History of Statistics and Probability, vol. II. Griffin, London (1977) Chi-Quadrat-Verteilung

Sheynin, OB: PS Laplaces Fehlertheorie, Archiv für Geschichte der exakten Wissenschaften **17,** 1–61 (1977) De Moivre-Laplace-Theorem

Machine Translated by Google

Sheynin, OB: PS Laplaces Arbeit über Wahrscheinlichkeit, Archiv für Geschichte der exakten Wissenschaften **16,** 137–187 (1976) De Moivre-Laplace-Theorem

Simpson, T.: Ein Versuch, den Vorteil aufzuzeigen, der sich aus der Mittelwertbildung aus einer Reihe von Beobachtungen in der praktischen Astronomie ergibt. In: Verschiedene Traktate zu einigen kuriosen und sehr interessanten Themen der Mechanik, der physikalischen Astronomie und der spekulativen Mathematik. Nourse, London (1757). S. 64–75 Arithmetisches Mittel

Simpson, T.: Ein Brief an den ehrenwerten George Earl of Macclesfield, Präsident der Royal Society, über den Vorteil der Mittelwertbildung aus einer Reihe von Beobachtungen in der praktischen Astronomie. Philos. Trans. Roy. Soc. London. **49,** 82–93 (1755) Arithmetisches Mittel, Fehler

Simpson, T.: Verschiedene Traktate zu einigen kuriosen und sehr interessanten Themen der Mechanik, der physikalischen Astronomie und der spekulativen Mathematik. Nourse, London (1757) Fehler

Sittig, J.: Die wirtschaftliche Wahl des Probenahmesystems bei der Annahmeprobenahme. Stier. Int. Stat. Inst. 33(V), 51–84 (1952) Stichprobengröße

Smirnov, NV: Schätzung der Abweichung zwischen empirische Verteilungsfunktionen in zwei unabhängigen Stichproben. (Russisch). Stier. Moskauer Universität. 2(2), 3–16 (6.1, 6.2) (1939) Kolmogorov-Smirnov-Test, nichtparametrischer Test

Smirnov, NV: Tabelle zur Schätzung der Anpassungsgüte empirischer Verteilungen. Ann. Mathematik. Stat. **19,** 279–281 (6.1) (1948) Kolmogorov–Smirnov Prüfen

Smith, CAB: Statistik in der Humangenetik. In: Kotz, S., Johnson, NL (Hrsg.) Encyclopedia of Statistical Sciences, vol. 3. Wiley, New York (1983) Genetische Statistik

Sneath, PHA, Sokal, RR: Numerische Taxonomie: Die Prinzipien und die Praxis der numerischen Klassifikation (Eine Reihe von Büchern zur Biologie). WH Freeman, San Francisco, CA (1973) Dendrogramm

Snedecor, GW: Berechnung und Interpretation von Analyse von Varianz und Kovarianz. Collegiate, Ames, IA (1934) Fisher-Verteilung

Sobol, IM: Die Monte-Carlo-Methode. Mir Verlag, Moskau (1975) Monte Carlo Methode

Sokal, RR, Sneath, PH: Prinzipien der numerischen Taxonomie. Freeman, San Francisco (1963) Dendrogram

Spearman, C.: Der Beweis und die Messung von Assoziation zwischen zwei Dingen. Bin. J. Psychol. **15,** 72–101 (1904) Spearman-Rangkorrelationskoeffizient

Statistische Wissenschaften: Statistische Analyse in S-Plus, Version 3.1. Seattle: StatSci, eine Abteilung von MathSoft (1993) Statistiksoftware

Staudte, RG, Sheeater, SJ: Robuste Schätzung und Konvergenz Testen. Wiley, New York (1990) Stein, C.: Ein Zwei-Stichproben-Test für eine lineare Hypothese, deren Aussagekraft unabhängig von der Varianz ist. Ann. Mathematik. Stat. **16,** 243–258 (1945) Probengröße

Stephens, MA: EDF-Statistiken zur Anpassungsgüte und einige Vergleiche. Marmelade. Stat. Assoc. **69,** 730–737 (1974) Anderson-Darling-Test

Stigler, S.: Francis Galtons Bericht über die Erfindung der Korrelation. Stat. Wissenschaft. 4, 73–79 (1989) Korrelationskoeffizient

Stigler, S.: Simon Newcomb, Percy Daniell und die Geschichte der robusten Schätzung 1885–1920. Marmelade. Stat. Assoc. **68,** 872–879 (1973) Ausreißer

Stigler, S.: Stiglers Eponymiegesetz, Transactions of the New York Academy of Sciences, 2. Reihe **39,** 147–157 (1980) Normalverteilung Stigler, S.: Die Geschichte der Statistik, die Messung der Unsicherheit vor 1900. Belknap , London (1986) Familie Bernoulli, Kombinatorische Analyse, Offizielle Statistik, Wahrscheinlichkeit, Regressionsanalyse, Statistik Stuart, A.,

Ord, JK: Kendalls fortgeschrittene Statistiktheorie. Bd. I. Verteilungstheorie. Wiley, New York (1943) Maß der zentralen Tendenz

Sukhatme, PV: Stichprobentheorie von Umfragen mit Anwendungen. Iowa State University Press, Ames, IA (1954) Quotenstichprobe

Takács, L.: Kombinatorik. In: Kotz, S., Johnson, NL (Hrsg.) Encyclopedia of Statistical Sciences, vol. 2. Wiley, New York (1982) Kombinatorische Analyse

Tassi, Ph.: Ausführlich zum Teil: eine Geschichte über die Entwicklung von Sonden. Schätzungen und Untersuchungen, fünf Beiträge zur Geschichte der Statistik. Economica, Paris (1988) Probenahme

Tchebychev, PL (1890–1891). Über zwei relative Theorien und Wahrscheinlichkeiten. Acta Math. **14,** 305–315 Zentraler Grenzwertsatz

Machine Translated by Google

Tschebychev, PL: Des valeurs moyennes. J. Mathe. Pures Appl. 12(2), 177–184 (1867). Gleichzeitig in Russland veröffentlicht, das Mat. Sbornik 2(2), 1–9 Gesetz der großen Zahlen Tenenhaus, M.: La régression PLS. Theorie und Praxis. Technip, Paris (1998) Teilweise Regression der kleinsten Quadrate

Thiele, TN: Theorie der Beobachtungen. C. und E. Layton, London; Ann. Mathematik. Stat. 2, 165–308 (1903) Methode der Momente

Thorndike, RL: Wer gehört in eine Familie? Psychometrika **18,** 267–276 (1953) Klassifikation, Datenanalyse Todhunter, I.: Eine Geschichte der mathematischen Wahrscheinlichkeitstheorie. Chelsea Publishing Company, New York (1949) Wahrscheinlichkeit

Tomasson, R., Daudin, JJ, Danzart, M., Masson, JP: Discrimination et classement. Masson, Paris (1988) Clusteranalyse

Tryon, RC: Clusteranalyse. McGraw-Hill, Neu York (1939) Einstufung

Tukey, JW: Voreingenommenheit und Vertrauen in nicht ganz große Stichproben. Ann. Mathematik. Stat. **29,** 614 (1958) Jackknife-Methode

Tukey, JW: Vergleich einzelner Mittel in der Varianzanalyse. Biometrie 5, 99–114 (1949) Kontrast

Tukey, JW: Explanatory Data Analysis, limitiert Vorläufige Ausgabe. Addison-Wesley, Reading, MA (1970–1971) Datenanalyse, Stamm-Blatt-Diagramm Tukey, JW: Explorative Datenanalyse. Addison-Wesley, Reading, MA (1977) Daten Analyse, explorative Datenanalyse Tukey, JW: Quick and Dirty Methods in Statistics. Teil II: Einfache Analysen für Standarddesigns. Konferenzbeiträge zur Qualitätskontrolle 1951. American Society for Quality Control, New York, S. 189–197 (1951) Kontrast

Tukey, JW: Einige grafische und halbgrafische zeigt an. In: Bancroft, TA (Hrsg.) Statistical Papers zu Ehren von George W. Snedecor. Iowa State University Press, Ames, IA, S. 293–316 (1972) Boxplot, explorative Datenanalyse, Hutmatrix

Twisk, JWR: Angewandte Längsschnittdatenanalyse für die Epidemiologie. Ein praktischer Leitfaden. Cambridge University Press, Cambridge (2003) Längsdaten

Grundlagen der Statistikkommission der Vereinten Nationen Grundsätze der amtlichen Statistik Offiziell Statistiken

Venables, WN, Ripley, BD: Moderne angewandte Statistik mit S-PLUS. Springer, Berlin Heidelberg New York (1994) Statistiksoftware

Von Neumann, J.: Verschiedene Techniken in Verbindung mit Zufallsziffern. Symposium des National Bureau of Standards, NBS, Applied Mathematics Series 12, S. 36–38, National Bureau of Standards, Washington, DC (1951) Monte-Carlo-Methode

Von Neumann, J.: Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. Ergebn. Mathematik. Kolloqu. Wien 8, 73–83 (1937) Linear Programmierung

Wald, A.: Asymptotisch leistungsfähigste Tests von Statistische Hypothesen. Ann. Mathematik. Stat. **12,** 1–19 (1941) Likelihood-Ratio-Test

Wald, A.: Einige Beispiele für asymptotisch leistungsfähigste Tests. Ann. Mathematik. Stat. **12,** 396–408 (1941) Likelihood-Ratio-Test

Wegman, EJ, Hayes, AR: Statistische Software. Enzykl. Stat. Wissenschaft. 8, 667–674 (1988) Statistiksoftware

Weiseberg, S.: Angewandte lineare Regression. Wiley, New York (2006) Multiple Linear Regression White, C.: Die Verwendung von Rängen in einem Signifikanztest zum Vergleich zweier Behandlungen. Biometrics 8, 33–41 (1952) Mann-Whitney-Test

Whyte, Lancelot Law: Roger Joseph Boscovich, Studien zu seinem Leben und Werk zum 250 Jahrestag seiner Geburt. Allen und Unwin, London (1961) Boscovich, Roger J.

Wilcoxon, F.: Einzelvergleiche nach Ranking-Methoden. Biometrie 1, 80–83 (1945) Mann-Whitney-Test, nichtparametrischer Test, Von Wilcoxon signierter Tisch, von Wilcoxon signiert Test, Wilcoxon-Test

Wilcoxon, F.: Wahrscheinlichkeitstabellen für Einzelpersonen Vergleiche nach Ranking-Methoden. Biometrie 3, 119–122 (1947) Wilcoxon-Tisch

Wilcoxon, F.: Einige schnelle ungefähre statistische Verfahren. Amerikanisches Cyanamid, Stamford Research Laboratories, Stamford, CT (1957) Nichtparametrischer Test, Wilcoxon-Signed-Test

Machine Translated by Google

Wildt, AR, Ahtola, O.: Analyse der Kovarianz (Sage University Papers Series on Quantitative Applications in the Social Sciences, Paper 12). Sage, Thousand Oaks, CA (1978)

Kovarianzanalyse Wilk, MB, Ganadesikan, R.: Wahrscheinlichkeitsdarstellungsmethoden für die Analyse von Daten. Biometrika **55,** 1–17 (1968) QQ-Diagramm (Quantil-zu-Quantil-Diagramm)

Wilks, SS: Mathematische Statistik. Wiley, Neu York (1962) Likelihood-Ratio-Test

Wold, HO (Hrsg.): Bibliographie zu Zeitreihen und stochastischen Prozessen. Oliver & Boyd, Edinburgh (1965) Zyklische Fluktuation

Wonnacott, RJ, Wonnacott, TH: Ökonometrie. Wiley, New York (1970) Ökonometrie

Yates, F.: Das Design und die Analyse der Fakultät Experimente, Techn. Komm. 35, Imperial Bureau of Soil Science, Harpenden (1937) Bruchteil Faktorielles Design

Yates, F.: Komplexe Experimente. Zus. J. Roy. Stat. Soc. 2, 181–247 (1935) Zwei-Wege-Varianzanalyse

Yates, F.: Stichprobenmethoden für Volkszählungen und Umfragen. Griffin, London (1949) Stratified Sampling

Yates, F.: Die Analyse wiederholter Experimente wenn die Feldergebnisse unvollständig sind. Empire J. Exp. Landwirtschaft. 1, 129–142 (1933) Fehlende Daten

Yates, F.: Das Design und die Analyse der Fakultät Experimente. Technische Mitteilung des Commonwealth Bureau of Soils **35,** Commonwealth Agricultural Bureaux, Farnham Royal (1937) Two-way Analysis of Variance, Yates-Algorithmus

Youschkevitch, AP: Les mathématiques arabes (VIIIème-XVème siècles). Teilübersetzung von Cazenave, M., Jaouiche, K. Vrin, Paris (1976) Arithmetisches Dreieck

Yule, GU (1926) Warum erhalten wir manchmal unsinnige Korrelationen zwischen Zeitreihen? Eine Studie über Stichproben und die Natur von Zeitreihen. J. Roy. Stat. Soc. **(2)** 89, 1–64 Korrelationskoeffizient, partielle Korrelation Yule, GU: Zur Assoziation von Attributen in Statistik: mit Illustration aus dem Material der Kindheitsgesellschaft. Philos. Trans. Roy. Soc. London. Ser. A **194,** 257–319 (1900) Chi-Quadrat-Test der Unabhängigkeit

Yule, GU: Zur Korrelationstheorie. J. Roy. Stat. Soc. **60,** 812–854 (1897) Korrelationskoeffizient

Yule, GV, Kendall, MG: An Introduction to the Theory of Statistics, 14. Aufl. Griffin, London (1968) Yule- und Kendall-Koeffizient

Zubin, J.: Eine Technik zur Messung von Gleichgesinntheit. J. Abnorm. Sozialpsychol. **33,** 508–516 (1938) Klassifizierung, Datenanalyse