Machine Translated by Google

SPRINGER
REFERENCE

Yadolah Dodge

The Concise Encyclopedia of Statistics

Springer

Machine Translated by Google

Краткая энциклопедия статистики

Machine Translated by Google

Ядола Додж

Краткая энциклопедия статистики

С 247 столами

123

Machine Translated by Google

Автор

Ядола Додж

Почетный профессор

Невшательский университет

Швейцария

yadolah.dodge@uninc.ch

Запись каталога CiP для этой книги доступна в Библиотеке Конгресса США.

ISBN: 978-0-387-32833-1

Эта публикация доступна также как:

Печатная публикация под ISBN 978-0-387-31742-7и

Печатная и электронная пачка под ISBN 978-0-387-33828-6

Эта работа является объектом авторского права. Все права защищены, будь то весь материал или его часть, в частности права на перевод, перепечатку, повторное использование иллюстраций, декламацию, трансляцию, воспроизведение на микрофильмах или другими способами, а также хранение в банках данных. Копирование этой публикации или ее частей разрешено только в соответствии с положениями Закона об авторском праве Германии от 9 сентября 1965 г. в его действующей редакции, и разрешение на использование всегда должно быть получено от Springer-Verlag. Нарушения подлежат судебному преследованию в соответствии с Законом об авторском праве Германии.

Springer является частью Springer Science+Business Media.

Springer.com

© 2008 Springer Science + Business Media, LLC.

Использование в этой публикации зарегистрированных наименований, товарных знаков и т. д. не означает, даже при отсутствии конкретного заявления, что такие наименования не подпадают под действие соответствующих охранных законов и нормативных актов и, следовательно, бесплатны для общего использования.

Отпечатано на бескислотной бумаге.

СПИН: 10944523 2109–543210

Machine Translated by Google

В память о моей любимой жене К.,
заботливой маме,
трудолюбивом отце и

моим двум добрым и сердечным сыновьям, Али и Арашу

Machine Translated by Google

Предисловие

Мы надеемся, что этим кратким томом удовлетворим потребности большого научного сообщества, которое прежде обслуживалось в основном огромными энциклопедическими справочниками. Вместо того, чтобы стремиться к всестороннему освещению нашего предмета, мы сосредоточились на наиболее важных темах, но объяснили их настолько глубоко, насколько позволяло место. В результате получилась компактная работа, которая, как мы верим, не упускает из виду ни одной центральной темы.
Записи имеют жесткую структуру для облегчения поиска информации. Каждый представленный здесь термин включает в себя определение, историю, математические детали, ограничения в использовании терминов, за которыми следуют примеры, ссылки и соответствующая литература для дальнейшего чтения. Справочник расположен в алфавитном порядке, чтобы обеспечить быстрый доступ к основным инструментам статистической методологии и биографиям известных статистиков, в том числе некоторых современных, которые продолжают вносить свой вклад в статистику, таких как сэр Дэвид Кокс, Брэдли Эфрон и Т.В. Андерсон. немного. Критерии отбора этих статистиков, живущих или отсутствующих, конечно, довольно личные, и вполне возможно, что некоторые из тех известных личностей, заслуживающих внимания, отсутствуют. Я искренне извиняюсь за любые такие непреднамеренные упущения.

Кроме того, была предпринята попытка представить основную информацию о статистических тестах, концепциях и аналитических методах на языке, доступном для практиков и студентов, а также для широкого сообщества, использующего статистику в медицине, технике, физике, науках о жизни, социальных науках. и бизнес/экономика.

Первые шаги по написанию этой книги были предприняты в 1983 году. В 1993 году первая французская версия была опубликована издательством Dunod в Париже. Позже, в 2004 г., обновленная и более длинная версия на французском языке была опубликована Springer France, а в 2007 г. в Springer было опубликовано студенческое издание французского издания.
В этой энциклопедии, как и в Оксфордском словаре статистических терминов, изданном для Международного статистического института в 2003 г., для каждого термина дается одна или несколько ссылок, в некоторых случаях на более ранний источник, а в других — на более позднюю публикацию. Несмотря на то, что при выборе ссылок были приняты определенные меры предосторожности, общеизвестно, что установление исторических приоритетов представляет собой трудную задачу, и исторические отнесения не следует считать авторитетными.
Дополнительную информацию о терминах, отсутствующих в этой энциклопедии, можно найти в коротких статьях в следующих энциклопедиях и словарях:

Machine Translated by Google

Международная энциклопедия статистики, ред. Уильям Крускал и Джудит М. Танур (The Свободная пресса, 1978).

Энциклопедия статистических наук, ред. Сэмюэл Котц, Норман Л. Джонсон и Кэмбелл Рид (Джон Уайли и сыновья, 1982).

Энциклопедия биостатистики, ред. Питер Армитидж и Тед Колтон (Чичестер: Джон Уайли и сыновья, 1998).

Энциклопедия Environmetrics, ред. А. Х. Эль-Шараави и В. В. Парегорик (Джон Уайли и сыновья, 2001).

Энциклопедия статистики качества и надежности, ред. Ф. Руджери, Р.С. Кенетт и Ф.В. Фалтин (John Wiley and Sons, 2008).

Dictionnaire - Encylopédique en Statistique, Yadolah Dodge, Springer 2004

В период между публикацией первой версии данной книги на французском языке в 1993 г. и более поздним изданием в 2004 г. рукопись претерпела множество исправлений. Особое внимание было уделено выбору подходящих переводов терминов, чтобы обеспечить их правильное значение как на английском, так и на французском языках. Если в некоторых случаях этого не произошло, приношу свои извинения. Я был бы очень признателен читателям за любые комментарии, касающиеся неточностей, исправлений и предложений по включению новых терминов или любых вопросов, которые могли бы улучшить следующее издание. Пожалуйста, присылайте свои комментарии в Springer-Verlag.

Я хочу поблагодарить многих людей, которые помогали мне на протяжении многих лет доводить эту рукопись до ее нынешнего вида. Начиная с моих бывших помощников с 1983 по 2004 год, Николь Ребетез, Сильви Гонано-Вебер, Марии Зегами, Юрга Шмида, Северин Пфафф, Джимми Бриньони, Элизабет Пастер, Валентина Руссона, Александры Франьер и Ихи Мюрье.

Моим коллегам Джо Уиттакеру из Ланкастерского университета, Людевику Лебару из France Tele com и Бернарду Фишеру из Марсельского университета за чтение частей рукописи. Russian.Работа с Таносом, моим бывшим кандидатом наук. студент, был замечательный опыт. Коллега Томy Шахриар Худа, чьи полезные комментарии, критические замечания и исправления внесли большой вклад в эту книгу.

Наконец, я благодарю издательство Springer-Verlag, особенно Джона Киммела, Эндрю Спенсера и Уну Шмид за их кропотливую работу по созданию этой энциклопедии.

январь 2008 г.

Ядола Додж
Почетный профессор
Невшательский университет
Швейцария

Machine Translated by Google

об авторе

Основатель программы магистра статистики в 1989 г. для Невшательского университета в Швейцарии, профессор Ядола Додж получил степень магистра прикладной статистики в Университете штата Юта в 1970 г. и степень доктора философии в области статистики со степенью по биометрии в Университете штата Орегон. в 1973 году. Он опубликовал множество статей, а также является автором, соавтором и редактором нескольких книг на английском и французском языках, в том числе « Математическое программирование в статистике » (John Wiley, 1981, Classic Edition, 1993), « Анализ экспериментов с отсутствующими данными» (John Wiley). 1985), Альтернативные методы регрессии (John Wiley, 1993), Premier Pas en Statistique (Springer, 1999), Adaptive Regression (Springer, 2000), Оксфордский словарь статистических терминов (2003), Statistique: Dictionnaire encyclopédique (Springer, 2004) и Optimization. аппликация (Springer 2005). Профессор Додж является избранным членом Международного статистического института (1976 г.) и членом Королевского статистического общества

Machine Translated by Google

Область приема

Область приемлемости — это интервал внутри выборочного распределения тестовой статистики , который согласуется с нулевой гипотезой H0 из проверки гипотезы.
Это комплементарная область области отклонения.

Область приемлемости связана с вероятностью 1
α, где α — уровень значимости теста.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

См. область отклонения.

ПРИМЕРЫ

См. область отклонения.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Критическое значение

- Проверка гипотезы
- Область отторжения
- Уровень значимости

Точность

Общее значение точности — это близость значения или статистики к эталонному значению. Более конкретно, она измеряет близость оценщика T неизвестного параметра θ к истинному значению θ.

Точность оценки можно измерить ожидаемым значением квадрата отклонения между T и θ, другими словами:

$$E\,(T - \theta)2\,.$$

Точность не следует путать с термином точность, который указывает на степень точности измерения и обычно обозначается числом десятичных знаков после запятой.
ма.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Предвзятость

- Оценщик
- Параметр
- Статистика

Алгоритм

Алгоритм — это процесс, состоящий из последовательности четко определенных шагов, ведущих к решению проблемы определенного типа. Этот процесс может быть итеративным, то есть повторяться несколько раз. Обычно это числовой процесс.

ИСТОРИЯ

Термин «алгоритм» происходит от латинского произношения имени математика девятого века аль-Хорезми, который жил в Багдаде и был отцом алгебры.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Слово «алгоритм» в последние годы приобрело другое значение в связи с появлением компьютеров. В области вычислений это относится к процессу, который описан таким образом, что его можно использовать в компьютерной программе. Основной целью статистического программного обеспечения является разработка языка программирования, способного включать статистические алгоритмы, чтобы эти алгоритмы можно было затем представить в форме, понятной пользователю. Преимущество такого подхода состоит в том, что пользователь понимает результаты, выдаваемые алгоритмом, и доверяет точности решений. Среди различных статистических обзоров, в которых обсуждаются алгоритмы, можно выделить Journal of Algorithms of the Academic Press (Нью-Йорк), часть Journal of the Royal Statistical Society Series C (Applied Statistics) , посвященную алгоритмам, Computational Statistics от Physica-Verlag. (Гейдельберг) и «Случайные структуры и алгоритмы» под редакцией Уайли (Нью-Йорк) заслуживают особого упоминания.

ПРИМЕРЫ

Мы представляем здесь алгоритм, вычисляющий абсолютное значение ненулевого числа; другими словами $|x|$.
Процесс:

Шаг 1. Определите алгебраический знак данный номер.

Шаг 2. Если знак отрицательный, перейти к шагу 3. Если знак положительный, укажите абсолютное значение числа в качестве сам номер:

$$|x| = x$$

и остановить процесс.

Шаг 3. Укажите абсолютное значение данного числа как его противоположное число:

$$|x|=-x$$

и остановить процесс.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Статистическое программное обеспечение

Алгоритм Йейтса

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Чемберс, Дж. М.: Вычислительные методы анализа данных. Уайли, Нью-Йорк (1977)

Хорезми, Муса ибн Меусба (9 век). Джабр ва-аль-мукабала. Алгебра Мохаммеда бен Мусы, Розен, Ф. (изд. и перевод). Георг Олмс Верлаг, Хильдесхайм (1986)

Рашед, Р .: La naissance de l'algèbre. В: Ноэль, Э. (ред.) Le Matin des Mathémati ciens. Белин-Радио Франция, Париж (1985)

Альтернативная гипотеза

Альтернативная гипотеза – это гипотеза, которая отличается от проверяемой гипотезы.

Альтернативная гипотеза обычно обозначается H1.

ИСТОРИЯ

См. гипотезы и проверка гипотез.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

При проверке гипотезы о параметре совокупности нулевая гипотеза представляется в следующем виде :

$$H0 : \theta = \theta0 ,$$

Machine Translated by Google

где θ — параметр популяции
которую необходимо оценить, а θ0 –
предполагаемое значение этого параметра. Тогда
альтернативная гипотеза может принимать три различных
формы:

1. H1 : θ>θ0
2. H1 : θ<θ0
3. H1 : θ = θ0

В первых двух случаях проверка гипотезы
называется односторонним, тогда как в третьем
случае его называют двусторонним.

Альтернативная гипотеза также может принимать три
различные формы при проверке гипотез о
параметрах двух популяций. Если
нулевая гипотеза рассматривает два параметра θ1
и θ2 поровну, то:

H0 : θ1 = θ2 или

H0 : θ1 - θ2 знак равно 0 .

Тогда альтернативная гипотеза может быть

- H1 : θ1 > θ2 или H1 : θ1 θ2 > 0
- H1 : θ1 < θ2 или H1 : θ1 θ2 < 0
- H1 : θ1 = θ2 или H1 : θ1 θ2 = 0

При сравнении более двух
популяции, нулевая гипотеза предполагает
что значения всех параметров _
идентичный. Если мы хотим сравнить k популяций,
нулевая гипотеза будет следующей:

H0 : θ1 = θ2 = ⋯ = θк .

Тогда альтернативная гипотеза будет сформулирована
следующим образом:

H1: значения θi(i = 1,...,k) не все
идентичный.

Это означает, что требуется только один параметр
иметь значение, отличное от других
параметры, чтобы отвергнуть нулевую гипотезу
и принять альтернативную гипотезу.

ПРИМЕРЫ

Мы собираемся рассмотреть альтернативу
гипотезы три примера гипотезы
тестирование:

1. Проверка гипотезы о проценте
   население
   Кандидат на выборах хочет знать,
   получит более 50% голосов.

   Нулевая гипотеза для этой задачи может
   быть записано следующим образом:

   H0 : π = 0,5 ,

   где π – процент населения, подлежащий
   оценке.

   Мы проводим односторонний тест на правой
   стороне, что позволяет нам ответить на вопрос
   вопрос Дидата. Таким образом, альтернативная
   гипотеза будет следующей:

   H1 : π > 0,5 .

2. Проверка гипотезы о среднем значении популярности
   образование

   Производитель болтов хочет проверить точность
   новой машины, которая должна делать болты
   8 мм в диаметре.

   Мы можем использовать следующую нулевую
   гипотезу:

   H0 : μ = 8 ,

   где μ - среднее значение населения
   это предстоит оценить.

   Проводим двухсторонний тест для проверки
   диаметр болта слишком мал или
   слишком большой.

   Альтернативную гипотезу можно
   сформулировать следующим образом:

   H1 : μ = 8 .

3. Проверка гипотез о сравнении
   средства двух популяций
   Страховая компания решила оборудовать
   офисы с микрокомпьютерами. Он хочет

Machine Translated by Google

покупать эти компьютеры у двух разных компаний до тех пор, пока нет существенной разницы в долговечности между ними.
два бренда. Поэтому он проверяет время который проходит до первого пробоя на образец микрокомпьютеров от каждого бренд.

Согласно нулевой гипотезе, среднее время, прошедшее до первого разбивка одинакова для каждой марки:

$$H0 : \mu1 \quad \mu2 \text{ знак равно } 0 .$$

Здесь μ1 и μ2 – соответствующие средние из двух популяций.
Поскольку мы не знаем, какое среднее будет быть самым высоким, мы проводим двухсторонний тест. Поэтому альтернативная гипотеза будет:

$$H1 : \mu1 \quad \mu2 \text{ знак равно } 0 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Гипотеза
- Проверка гипотезы
- Нулевая гипотеза

ССЫЛКА

Леманн, Э.И., Романн, С.П.: Проверка статистической гипотезы, 3-е изд. Спрингер, Нью Йорк (2005)

Анализ бинарных данных

Изучение того, как вероятность успеха зависит от объяснительных переменных и группировки материалов.
Анализ бинарных данных также включает проверки согласия выборки бинарных переменных к теоретическому распределению, а также как изучение таблиц непредвиденных обстоятельств 2 × 2 и их последующий анализ. В последнем case отметим особенно тесты на независимость между атрибутами и тестами на однородность.

ИСТОРИЯ

См. анализ данных.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть Y — двоичная случайная величина и X1, X2,...,Xk — дополнительные бинарные переменные. Итак, зависимость Y от переменных X1, X2 ,...,Xk представлена следующими моделями (коэффициенты которых равны оценивается по максимальному правдоподобию):

1. Линейная модель: P(Y = 1) выражается как линейная функция (по параметрам) от Xi.
2. Логлинейная модель: log P(Y = 1) есть выражается в виде линейной функции (в параметры) Си.
3. Логистическая модель: log $\frac{П(У=1)}{П(У=0)}$ в виде линейной функции (в параметры) Си.

Модели 1 и 2 легче интерпретировать. Еще Преимущество последнего состоит в том, что объясняемая величина принимает все возможные значения. линейных моделей. Также важно обратите внимание на экстраполяцию модели за пределы области, в которой она применяется. Возможно, среди независимых переменные (X1, X2,...,Xk), есть категориальные переменные (например, бинарные). В этом случае необходимо рассматривать небинарный категориальные переменные следующим образом: пусть Z — случайная величина с m cate кровь. Перечислим категории от 1 к m , и мы определяем m 1 случайную переменную возможности Z1, Z2,...,Zm 1. Таким образом, Zi принимает значение 1, если Z принадлежит категории, представленной этим индексом. Таким образом , переменная Z заменяется этими m 1 переменными, т.е. коэффициенты которой выражают влияние

Machine Translated by Google

рассматриваемая категория. Ссылка (используется для того, чтобы избежать ситуации коллинеарности ity) будет иметь (для целей сравнения с другими категориями) параметр нуль.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Двоичные данные
- Анализ данных

ССЫЛКИ Кокс, Д.Р., Снелл, Э.Дж.: Анализ бинарных данных. Чепмен и Холл (1989)

Анализ категорийных данных

Анализ категорийных данных включает следующие методы:

а) исследование критерия согласия;

(б) изучение таблицы сопряженности и ее последующий анализ, который состоит в обнаружении и изучении взаимосвязей между атрибутами (если они существуют);

в) критерий однородности некоторых популяций, связанный с распределением бинарной качественной категориальной переменной;

(d) Проверка гипотезы независимости .

ИСТОРИЯ

Термин «непредвиденные обстоятельства», используемый в отношении кросс-таблиц категориальных данных , вероятно, впервые был использован Пирсоном, Карлом (1904). Тест хи-квадрат был предложен Барлеттом в 1937 году.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

См. таблицу соответствия и непредвиденных обстоятельств.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Данные
- Анализ данных
- Категориальные данные
- Хи-квадрат критерия согласия
- Таблица сопряженности
- Анализ корреспонденции
- Тест на пригодность
- Тест на однородность
- Тест на независимость

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Агрести, А.: Категориальный анализ данных. Уайли, Нью-Йорк (1990)

Бартлетт, М.С.: Свойства достаточности и статистических критериев. проц. Рой. соц. Лонд. сер. А 160, 268–282 (1937)

Кокс, Д.Р., Снелл, Э.Дж.: Анализ двоичных данных, 2-е изд. Чепмен и Холл, Лондон (1990)

Хаберман, С.Дж.: Анализ качественных данных. Том. I: Вводные темы. Академик, Нью-Йорк (1978)

Пирсон, К.: О теории случайности и ее связи с ассоциацией и нормальным явлением. корреляция. Мемуары исследований компании Drapers ', биометрическая сер. I., стр. 1–35 (1904 г.)

Анализ остатков

Анализ остатков используется для проверки достоверности статистической модели и контроля допущений, сделанных в отношении ошибки. Его также можно использовать для обнаружения выбросов.

ИСТОРИЯ

Анализ остатков восходит к Эйлеру (1749 г.) и Майеру (1750 г.) в середине

Machine Translated by Google

восемнадцатого века, которые столкнулись с проблемой оценки параметров по наблюдениям в области астрономии . Большинство методов, используемых для анализа остатков, основаны на работах Anscombe (1961) и Anscombe and Tukey (1963). В 1973 году Анскомб также представил интересное обсуждение причин использования графических методов анализа. Кук и Вайсберг (1982) посвятили целую книгу анализу остатков. Дрейпер и Смит (1981) также рассмотрели эту проблему в главе своей работы « Прикладной регрессионный анализ».

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим общую модель множественной линейной регрессии:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \varepsilon_i, \quad i = 1,\ldots,n,$$

где εi — ненаблюдаемая случайная ошибка срок.

Гипотезы для ошибок εi обычно следующие: • ошибки независимы; • Они нормально распределены (они следуют нормальному распределению); • их среднее значение равно нулю; • Их дисперсия постоянна и равна σ2.

Регрессионный анализ дает оценку для Yi, обозначаемого Yˆi. Если выбранная модель адекватна, то распределение невязок или «наблюдаемых ошибок» ei = Yi Yˆi должно подтверждать эти гипотезы.

Методы, используемые для анализа остатков, в основном графические. К таким методам относятся: 1. Представление остатков частотной диаграммой (например, точечной диаграммой).

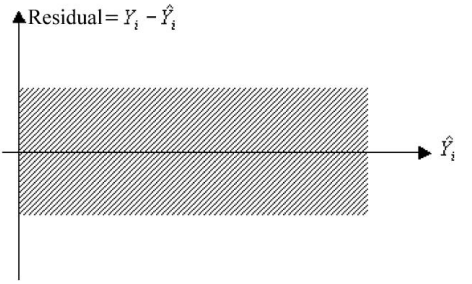
2. Построение графика зависимости остатков от времени (если известен хронологический порядок).
3. Построение графика невязок в зависимости от оценочных значений Yˆi.
4. Построение графика невязок в зависимости от независимых переменных Xij.
5. Создание графика Q–Q остатков.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Чтобы подтвердить анализ, некоторые гипотезы должны выполняться (например, нормальность остатков в оценках, основанных на среднем квадрате).

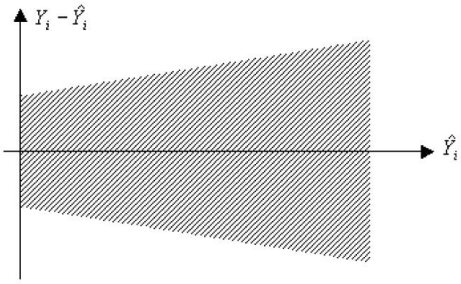
Рассмотрим график остатков как функцию оценочных значений Yˆi. Это один из наиболее часто используемых графических подходов к проверке достоверности модели . Он состоит в размещении: • остатков ei = Yi Yˆi по возрастанию заказ; • Оценочные значения Yˆi по оси абсцисс.

Если выбранная модель адекватна, остатки равномерно распределяются по горизонтальной полосе точек.

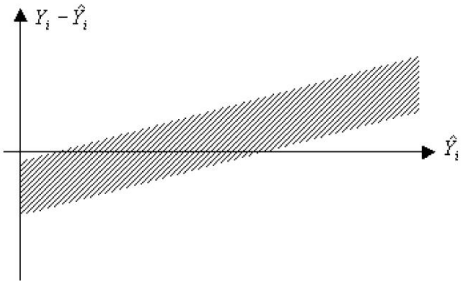


Однако, если гипотезы об остатках не проверяются, форма графика может быть иной. На трех рисунках ниже показаны формы, полученные в следующих случаях: 1. Дисперсия σ2 непостоянна. В этом случае необходимо выполнить преобразование данных Yi , прежде чем приступать к регрессионному анализу.

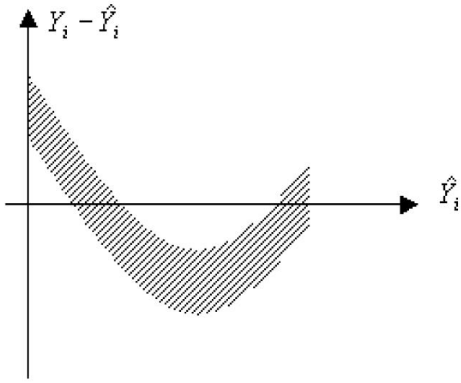
Machine Translated by Google



2. Выбранная модель неадекватна (для
Например, модель является линейной, но
постоянный член был опущен, когда это было необходимо
эссари).



3. Выбранная модель не подходит
(наблюдается параболическая тенденция).



Различные статистические данные были предложены в
для того, чтобы разрешить численные измерения, которые
дополняют визуальные техники
представленные выше, в том числе данные
Anscombe (1961) и Anscombe и
Тьюки (1963).

ПРИМЕРЫ

В девятнадцатом веке шотландский физик
имени Forbe, Джеймс Д. хотел оценить
высоту над уровнем моря, измеряя
температура кипения воды. Он знал, что высоту
можно определить по атмосфере.
сферическое давление; затем он изучал отношение
между давлением и температурой кипения
вода. Forbe предположил, что за интервал
наблюдаемых значений, график логарифма
давление как функция температуры кипения
воды должны давать прямую линию. С
логарифм этих давлений мал и
мало меняется, мы перемножили эти значения
на 100 ниже.

Икс точка кипения	Д 100 · log (давление)
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

Простая модель линейной регрессии для этого
проблема в следующем:

$$Yi = \beta 0 + \beta 1Xi + i , \qquad я = 1,..., 17 \quad .$$

Machine Translated by Google

Используя метод наименьших квадратов, мы можем найти следующую функцию оценки:

$$\hat{Y}_i = 42{,}131 + 0{,}895X_i$$

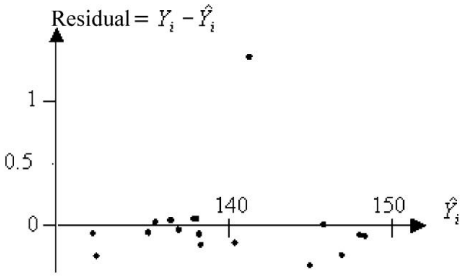
где $\hat{Y}_i$ – оценочное значение переменной Y для заданного X.

Для каждого из этих 17 значений Xi имеем оценочное значение $\hat{Y}_i$. Мы можем рассчитать остатки:

ei знак равно Yi - $\hat{Y}_i$.

Эти результаты представлены в следующем стол:

я	Си	И	Y я	эй И	"=" $\hat{Y}_я$
1	194,5	131,79	132,037		0,247
2	194,3	131,79	131,857		0,067
3	197,9	135,02	135,081		0,061
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		0,042
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		0,155
10	201,3	138,05	138,126		0,076
11	203,6	140,04	140,185		0,145
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		0,323
15	210,7	146,30	146,543		0,243
16	211,9	147,54	147,618		0,078
17	212,2	147,80	147,886		0,086



Построение остатков как функции оценочные значения $\hat{Y}_i$ дает предыдущую график.

Из этого графика видно, что, за исключением одно наблюдение (12-е), где значение остатка, кажется , указывает на выход э., остатки распределяются очень тонким горизонтальная полоса. В этом случае остатки не давать оснований сомневаться в действительности выбранной модели. Анализируя стандартизированные остатки, мы можем определить, 12-е наблюдение является выбросом или нет.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Тест Андерсона-Дарлинга

Наименьших квадратов

Множественная линейная регрессия

Выброс

Регрессивный анализ

Остаток

диаграмма рассеяния

Простая линейная регрессия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Anscombe, FJ: Исследование остатков. проц. 4-й симпозиум Беркли. Мат. Статист. Проб. 1, 1–36 (1961)

Anscombe, FJ: Графики в статистическом анализе. Являюсь. Стат. 27, 17–21 (1973)

Анскомб, Ф.Дж., Тьюки, Дж.В.: Анализ остатки. Технометрика 5, 141–160 (1963)

Кук, Р.Д., Вайсберг, С.: Остатки и Влияние в регрессии. Чепмен и Холл, Лондон (1982)

Кук, Р.Д., Вайсберг, С.: Введение в Regression Graphics.Wiley, Нью-Йорк (1994)

Кук, Р.Д., Вайсберг, С.: Прикладная регрессия, включая вычисления и графику. Уайли, Нью-Йорк (1999)

Machine Translated by Google

Дрейпер, Н.Р., Смит, Х.: Прикладной регрессионный анализ, 3-е изд. Уайли, Нью-Йорк (1998)

Эйлер, Л.: Исследования по вопросу о desiné galités du Mouvement de Saturne et de Юпитер, получивший премию l'Année 1748, par l'Académie royale des наук Парижа. Republié en 1960, данс Леонарди Эйлери, Opera Omnia, 2ème серия. Туричи, Базе, 25, стр. 47–157 (1749 г.)

Mayer, T.: Abhandlung über dieUmwälzung desMonds um seineAchse und die schein голый Bewegung der Mondfleken. Kos mographische Nachrichten und Samm lungen auf das Jahr 17481, 52–183 (1750)

Дисперсионный анализ

Дисперсионный анализ – это метод, состоит из разделения общей вариации набор данных в логические компоненты, связанные с конкретными источниками изменчивости, чтобы для сравнения среднего значения нескольких популяций. Этот анализ также помогает нам проверить определенных гипотез , касающихся параметров модели, или для оценки компонентов дисперсии . Источники вари суммируются в компоненте, называемом дисперсией ошибки, иногда называемой средний квадрат внутри лечения и другое компонент, который называется «эффектом» или лечением, иногда называемым межлечебным средний квадрат.

ИСТОРИЯ

Дисперсионный анализ восходит к Fisher, RA (1925). Он установил первые фундаментальные принципы в этой области. Анализ дисперсия была впервые применена в области биологии и сельского хозяйства.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Дисперсионный анализ сравнивает среднее значение трех или более случайных выборок и определяет, есть ли существенная разница между популяциями из которого берутся пробы. Этот метод может быть применен только в том случае, если случайный выборки независимы, если совокупность распределение примерно нормальное и все имеют одинаковую дисперсию σ2.

Установив, что нулевая гипотеза, предполагает, что средние равны, а альтернативная гипотеза утверждает, что по крайней мере один из них отличается, фиксируем существенное уровень. Затем делаем две оценки неизвестная дисперсия σ2:

- Первый, обозначаемый $s^2_{\bar{y}}$, соответствует среднее значение дисперсии каждой выборки;
- Второй, s^2_{Tr}, основан на вариации между средствами образцов.

В идеале, если нулевая гипотеза подтверждается, эти две оценки будут равны, и F отношение ($F = s^2_{Tr} / s^2_E$, определяемое как отношение второй оценки σ2 к первой) будет равно 1. как используется в тесте Фишера Значение коэффициента F , которое обычно больше 1 из-за отклонения от выборки, необходимо сравнить со значением в таблица Фишера , соответствующая фиксированному значительный уровень. Правило принятия решения состоит либо отвергнуть нулевую гипотезу, если расчетное значение больше или равно табличное значение, иначе средние равны, что показывает, что образцы взяты из такое же население.

Рассмотрим следующую модель:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

$$i = 1, 2, \ldots, t, \; j = 1, 2, \ldots, n_i .$$

Здесь

Y_{ij} представляет наблюдение j, получающее лечение я ,

Machine Translated by Google

μ — общее среднее, общее для всех обработок,

τi — фактический эффект лечения i на наблюдение,

εij — экспериментальная ошибка наблюдения Yij.

В этом случае нулевая гипотеза выражается следующим образом:

$$H0 : \tau1 = \tau2 = \quad \cdots \quad = \tau\tau ,$$

что означает, что t обработок идентичны тикальный.

Альтернативная гипотеза сформулирована в следующим образом:

H1 : значения τi (i = 1, 2,...,t)
не все одинаковы .

Используются следующие формулы:

$$SSTr = \sum_{я=1}^{т} ni(\bar{Y}i. \quad \bar{Y}..)^2 , \qquad c^2_{Тр} \text{ "=" } \frac{ССТр}{т - 1} ,$$

$$ССЭ = \sum_{я=1}^{т} \sum_{j=1}^{ни} (Yij \quad \bar{Y}i.)^2 , \qquad c^2_{Е} \text{ "=" } \frac{ССЭ}{Н - т} ,$$

и

$$ТСТ = \sum_{я=1}^{т} \sum_{j=1}^{ни} (Yij \quad \bar{Y}..)^2$$

или

$$SST = SSTr + SSE .$$

где

$\bar{Y}i.$ "=" $\frac{\sum_{j=1}^{ни} Йидж}{ни}$ является средним значением i -й набор

$\bar{Y}.. = \frac{1}{N} \sum_{я=1}^{т} \sum_{j=1}^{ни}$ Yij - глобальное среднее взято на все наблюдения и

$H = \sum_{я=1}^{т} ни$ это общее количество наблюдений.

и, наконец, значение коэффициента F

$$Ф = \frac{c^2_{Тр}}{c^2_{Е}} .$$

принято обобщать информацию из дисперсионного анализа в анализе таблицы дисперсии:

Источник разнообразия ция	Степени из свобода	Сумма квадраты	Иметь в виду из квадраты	Ф
Среди обращаться менты	т - 1	ССТр	$c^2_{Тр}$	$\frac{c^2_{Тр}}{c^2_{Е}}$
В пределах обращаться менты	Н - т	ССЭ	$c^2_{Е}$	
Всего N - 1		SST		

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Дисперсионный анализ всегда связан с моделью. Следовательно, существует разный анализ дисперсии в каждом отдельном случае. случай. Например, рассмотрим случай, когда дисперсионный анализ применяется к факторным экспериментам с одним или несколькими факторами, и эти факторные эксперименты связаны с несколько планов эксперимента.
Мы можем различать не только количество факторы эксперимента , но и тип гипотез , связанных с влиянием лечение. Затем у нас есть модель с фиксированным эффекты, модель с переменными эффектами и модель со смешанными эффектами. Каждый из них требует специального анализа, но какая бы модель ни использовалась, основные допущения аддитивность, нормальность, гомоскедастичность и независимость должна уважаться. Это означает что:

1. Экспериментальные ошибки модели _ случайные величины , которые независимы друг друга;

Machine Translated by Google

2. Все ошибки следуют нормальному дистрибутиву.
со средним значением , равным нулю, и
неизвестная дисперсия σ2.

Все планы эксперимента могут быть проанализированы с помощью дисперсионного анализа. Наиболее распространенные дизайны полностью рандомизированы дизайны, рандомизированные дизайны блоков и Латинские квадратные узоры.

Дисперсионный анализ также может быть выполнен с помощью простых или множественных линейных регресс.

Если при дисперсионном анализе гипотеза (случай равенства средних) является отклонено, тест на наименьшую значимую разницу используется для выявления популяций , значительно отличающиеся средства, чего дисперсионный анализ сделать не может.

ПРИМЕРЫ

См. двусторонний дисперсионный анализ, однофакторный дисперсионный анализ, линейно-кратный регрессия и простая линейная регрессия.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

План экспериментов
Фактор
Распределение Фишера
Таблица Фишера
Тест Фишера
Тест на наименьшее существенное различие
Множественная линейная регрессия
Односторонний дисперсионный анализ
Регрессивный анализ
Простая линейная регрессия
Двусторонний дисперсионный анализ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Статистические методы для
Исследовательские работники. Оливер и Бойд,
Эдин Бург (1925)

Рао, ЧР: Расширенные статистические методы
в биометрических исследованиях. Уайли, Нью-Йорк
(1952)

Шеффе, Х.: Дисперсионный анализ.
Уайли, Нью-Йорк (1959)

Андерсон, Оскар

Андерсон, Оскар (1887–1960) был важный член Континентальной школы статистики; его вклад коснулся широкий круг тем, включая корреляцию, анализ временных рядов, непараметрические методы и выборочное обследование, а также эконометрики и статистических приложений в социальные науки.

Андерсон, Оскар получил степень бакалавра с отличием Казанской гимназии а потом изучал математику и физику на год в Казанском университете. Он затем поступил на экономический факультет в Политехнический институт Санкт-Петербурга, где он изучал математику, статистику и экономика.

Публикации Андерсона, Оскара сочетают в себе традиции Континентальной школы.

статистики с понятиями англ.

Биометрическая школа, особенно в двух из его работы: «Einführung in die mathema tische Statistik» и «Probleme der Statis».

tischen Methodenlehre in den Sozialwis

чувства».

В 1949 году он основал журнал Mitteilungs blatt für Mathematische Statistik с Келлерер, Ганс и Мюнцнер, Ганс.

Некоторые основные работы Андерсона, Оскара:

1935 Einführung в математике
Статистика. Юлиус Спрингер, Вена

1954 Проблема статистического метода
denlehre in den Sozialwissenschaf
десять. Physica-Verlag, Вюрцберг

Machine Translated by Google

Андерсон, Теодор В.

Андерсон, Теодор Уилбур родился 5 июня 1918 года в Миннеаполисе, в г. штат Миннесота в США. Он стал доктор математических наук в 1945 г. Принстонского университета, а в 1946 г. стал сотрудником отдела Математическая статистика в Университете г. Колумбия, где он был назван профессором в 1956 г. В 1967 г. ему было присвоено звание профессора. Статистики и экономики в Стэнфордском университете версия. Он был последовательно: членом Фонд Гуггенхайма между 1947 и 1948 год; редактор «Анналов математики». Статистика с 1950 по 1952 год; Президент Институт математической статистики в 1963 г.; и вице-президент американского Statis ассоциации с 1971 по 1973 год.

член Американской академии искусств и наук Национальной академии наук Института математики Статистика и Королевского статистического общества. Самый важный вклад Андерсона к статистике, безусловно, относится к области многомерного анализа. В 1958 году он опубликовал книга под названием «Введение в мультивари». ели Статистический анализ. Эта книга была справочная работа в этой области более сорока годы. Его даже перевели на русский язык.

Некоторые из основных работ и статей Теодор Уилбур Андерсон:

1952 г. (совместно с Дарлингом, Д.А.) Асимптотическая теория некоторых критериев согласия. на основе стохастических процессов. Анна. Мат. Стат. 23, 193–212.

1958 г. Введение в многомерную статистику тистический анализ. Уайли, Нью-Йорк.

1971 Статистический анализ времени Ряд. Уайли, Нью-Йорк.

1989 Линейные модели скрытых переменных и ковариационные структуры. Дж. Экономет рис., 41, 91–119.

1992 (совместно с Кунитомой, Н.) Асимптотика распределения коэффициентов регрессии и авторегрессии с разностными возмущениями Мартина-Гейла. J. Multivariate Anal., 40, 221–243.

1993 Критерии согласия для спектральных распределений. Анна. Стат. 21, 830–847.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Тест Андерсона-Дарлинга

Тест Андерсона-Дарлинга

Тест Андерсона–Дарлинга — это критерий согласия, который позволяет контролировать гипотезу о том, что распределение случайной величины наблюдаемое в образце следует определенному теоретическому распределению. В частности, он позволяет нам проверить, является ли эмпирическое распределение полученное соответствует нормальному распределению.

ИСТОРИЯ

Андерсон, Теодор В. и Дарлинг Д.А. первоначально использовал статистику Андерсона-Дарлинга, обозначенный A2, для проверки соответствия дистрибутива точно заданным параметрам (1952 и 1954). Позднее, в 1960-х гг. и особенно 1970-х гг., некоторые другие авторы (в основном Стивенс) адаптировали тест к более широкому кругу диапазон дистрибутивов, где некоторые из параметры могут быть неизвестны.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим случайную величину X, которое следует нормальному распределению с математическое ожидание μ и дисперсию σ2, и имеет функцию распределения FX(x; θ), где θ параметр (или набор параметров), который

Machine Translated by Google

определить, FX. Кроме того, мы предполагаем, что θ равно быть известным.

Выпущено наблюдение за выборкой размера n по переменной X дает функцию распределения Fn(x). Статистика Андерсона-Дарлинга, обозначается A2, тогда задается взвешенной суммой квадратов отклонений FX(x; θ) Φн(x):

$$A2 = \frac{1}{н} \sum_{я=1}^{н} (FX(x;\theta) - Fn(x))2 .$$

Исходя из того, что A2 является случайным переменная, которая следует определенному распределению на интервале [0; + [, можно проверить для уровня значимости, который априори фиксируется, является ли Fn(x) реализацией случайная величинаFX(X; θ); то есть X следует распределению вероятностей с функция распределения FX(x; θ).

2 Вычисление статистики

Расположите наблюдения x1, x2,...,xn в выборка выдана из X в порядке возрастания, т.е. x1 < x2 < ... < хн. Обратите внимание, что zi = FX(xi; θ), (i = 1, 2,...,n). Затем вычислите, A2 от:

$$A2 = \frac{1}{н} \sum_{я=1}^{н} (2i - 1)(\ln(zi) + \ln(1 - zn+1-i)) - n .$$

Для предпочтительной здесь ситуации (X следует нормальное распределение с математическим ожиданием μ и дисперсии σ2), можно перечислить четыре случая в зависимости от известных параметров μ и σ2 (F – функция распределения стандартное нормальное распределение):

1. μ и σ2 известны, поэтому FX(x;(μ, σ2)) отлично указано. Естественно, мы тогда имеют $zi = F(wi)$, где $wi = 2. \sigma 2 \frac{xi - \mu}{о}$. известно, но μ неизвестно и оценивается с помощью $x = \frac{1}{н} \sum_{я=1}^{н} кси$, среднее значение образец. Тогда пусть zi = F(wi), где ви = $\frac{xi-x}{о}$.
3. μ известно, но σ2 неизвестно и оценивается с помощью $s2 = \frac{1}{н} \sum_{я=1}^{н} (xi - u)2$ В этом случае пусть zi = F(wi), где $wi = \frac{x(i) - \mu}{c}$.
4. μ и σ2 неизвестны и оцениваются соответственно с помощью x и $\bar{s}2 = \frac{1}{n-1} \sum_{я=1}^{н} (xi - x)2)$. Тогда пусть zi = F(wi), где $wi = \frac{xi-x}{c}$.

Асимптотические распределения были найдены для A2 Андерсон и Дарлинг для первого случая, и Стивенсом для следующих двух случаев. Для В последнем случае Стивенс определил асимптотическое распределение преобразования: A = A2(1,0+). $\frac{0,75}{н} + \frac{2,25}{п2}$ Поэтому, как показано ниже, мы можем построить таблица, дающая в зависимости от случая и уровень значимости (10%, 5%, 2,5% или 1% ниже), предельные значения A2 (и A для случая 4), за пределами которого нормальность гипотеза отвергается:

	Уровень значимости			
Случай:	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: A* =	0,631	0,752	0,873	1,035

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Поскольку распределение A2 выражается асимптотически, тест требует, чтобы размер выборки n был большой. Если это не так, то для первого два случая, распределение A2 неизвестно и необходимо произвести преобразование типа A2 A , из которого A можно определить. Когда n > 20, мы можем избежать такого преобразования, поэтому данные в приведенной выше таблице действительны.

Преимущество теста Андерсона-Дарлинга в том, что его можно применять к широкому диапазону

Machine Translated by Google

распределений (не только нормального, но и экспоненциального, логистического и гамма-распределения и др.). Что позволяет использовать широкий спектр альтернативных распределения, если первоначальный тест отклоняет нуль гипотеза о распределении случайного переменная.

ПРИМЕРЫ

Следующие данные иллюстрируют применение теста Андерсона–Дарлинга для гипотезы нормальности:

Рассмотрим образец высоты (в см) 25 студентов мужского пола. В следующей таблице показано наблюдения в выборке, а также wi и зи. Мы также можем вычислить $\bar{x}$ и $\bar{s}$ из эти данные: $\bar{x} = 177{,}36$ и $s = 4{,}98$.

Предполагая, что F — стандартная функция нормального распределения, имеем:

Наблюдения:	xi	$c = \frac{xi - \bar{x}}{s}$	зи = Ж с
1	169	1,678	0,047
2	169	1,678	0,047
3	170	1,477	0,070
4	171	1,277	0,100
5	173	0,875	0,191
6	173	0,875	0,191
7	174	0,674	0,250
8	175	0,474	0,318
9	175	0,474	0,318
10	175	0,474	0,318
11	176	0,273	0,392
12	176	0,273	0,392
13	176	0,273	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

Тогда мы получаем A2 = 0,436, что дает

$$A^* = A2 \cdot 1{,}0 + \frac{0{,}75}{25} + \frac{0{,}25}{625}$$

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

Поскольку у нас есть случай 4 и уровень значимости зафиксирован на уровне 1%, расчетное значение A равно намного меньше, чем значение, указанное в таблице (1,035). Поэтому гипотеза нормальности нельзя отбрасывать на уровне значимости 1%.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Тест на пригодность
- Гистограмма
- Непараметрическая статистика
- Нормальное распределение
- Статистика

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Андерсон, Т.В., Дарлинг, Д.А.: Асимптотическая теория некоторых критериев согласия на основе стохастических процессов. Энн.Математика. Стат. 23, 193–212 (1952)

Андерсон, Т.В., Дарлинг, Д.А.: Тест хорошая посадка. Варенье. Стат. доц. 49, 765–769 (1954)

Дурбин Дж., Нотт М., Тейлор С.К. Компоненты статистики Крамера-Фон Мизеса. II. Дж. Рой. Стат. соц. сер. Б 37, 216–237 (1975)

Стивенс, Массачусетс: статистика EDF для добра подгонки и некоторых сравнений. Варенье. Стат. доц. 69, 730–737 (1974)

Machine Translated by Google

Среднее арифметическое

Среднее арифметическое является мерой центральная тенденция. Это позволяет нам охарактеризовать центр частотного распределения количественной переменной , рассматривая все наблюдения с одинаковым весом, присвоенным каждому (в отличие от взвешенного среднего арифметического).

Он рассчитывается путем суммирования наблюдений и последующего деления на количество наблюдений.

ИСТОРИЯ

Среднее арифметическое является одним из древнейших методы, используемые для объединения наблюдений в чтобы дать уникальное приблизительное значение. По-видимому, впервые его использовали вавилонские астрономы в третьем веке до нашей эры. Среднее арифметическое использовалось астрономами для определения положения Солнца, Луны и планет. Согласно Plackett (1958), концепция среднего арифметического возникла у греческого астронома Гиппарха.

В 1755 году Томас Симпсон официально предложил использовать среднее арифметическое в письме президенту Королевского общества.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x1, x2,...,xn — набор из n величин или n наблюдений, относящихся к количественной переменной X.

Среднее арифметическое $\bar{x}$ x1, x2,...,xn равно сумме этих наблюдений, деленной на число n наблюдений:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Когда наблюдения упорядочены в форма частотного распределения, арифметическая среднее значение рассчитывается следующим образом способ:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i},$$

где xi — различные значения переменной , fi — частоты, связанные с этими значениями, k — количество различных значений, а сумма частот равна количеству наблюдений:

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

Чтобы вычислить среднее значение частотного распределения, в котором значения количественной переменной X сгруппированы по классам, мы считаем, что все наблюдения, принадлежащие определенному классу, принимают центральное значение класса, предполагая, что наблюдения равномерно распределены внутри классов (если эта гипотеза неверна, полученное среднее арифметическое будет лишь приближением).

Следовательно, в данном случае имеем:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i},$$

где xi — центры классов, fi — частоты, связанные с каждым классом, а k — количество классов.

Свойства среднего арифметического •

Алгебраическая сумма отклонений между каждым значением набора и средним арифметическим этого набора равна 0:

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

Machine Translated by Google

• Сумма квадратов отклонений от каждого значение заданного числа «a» наименьшее когда «a» — среднее арифметическое:

$$\sum_{i=1}^{n} 2\,(x - a) \qquad \sum_{i=1}^{n} 2\,(x_i - \bar{x}) \cdot$$

Доказательство:

Мы можем написать:

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

Нахождение квадратов обоих членов равенство, суммируя их, а затем упрощение дает:

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

Поскольку $n \cdot (\bar{x} - a)^2$ неотрицательно, имеем доказал, что:

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

• Среднее арифметическое $\bar{x}$ выборки $(x_1,\ldots,x_n)$ обычно считается быть оценкой среднего μ _ население , из которого была отобрана выборка взятый.

• Предполагая, что x_i — независимые случайные величины с одинаковым распределением функции для среднего μ и дисперсии σ^2, можно показать, что

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

если эти моменты существуют.

Поскольку математическое ожидание $\bar{x}$ равно μ, среднее арифметическое является оценкой без смещения среднего значения генеральной совокупности.

• Если хрезультат случайной выборки без замены конечной популяции со средним μ тождество

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

по-прежнему действует, но дисперсия $\bar{x}$ должна быть регулируется фактором, зависящим от размер N населения и размер n образец:

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

где σ^2 — дисперсия населения.

Связь между средним арифметическим и другие меры центральной тенденции

• Среднее арифметическое связано с двумя основными мерами центральной тенденции: мода Mo и медиана Md.

Если распределение симметрично и унимодально:

$\bar{x}$ знак равно Md знак равно Mo .

Если распределение унимодальное, обычно верно, что:

$\bar{x}$ Md Mo , если распределение вытянут вправо,

$\bar{x}$ Md Mo , если распределение растянут влево.

Для унимодального, слегка асимметричного распределение, эти три меры центральная тенденция часто примерно удовлетворяют следующему соотношению:

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

• Точно так же для унимодального распределения, если мы рассматриваем набор положительных чисел, среднее геометрическое G равно

Machine Translated by Google

всегда меньше или равно среднему арифметическому $\bar{x}$ и всегда больше или равно среднему гармоническому H. Таким образом, мы имеем:

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

Эти три средства идентичны, только если все числа равны.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Среднее арифметическое – это простая мера центрального значения набора количественных наблюдения. Поиск среднего иногда может привести к плохой интерпретации данных:

Если месячная заработная плата (в евро) 5 человек 3000, 3200, 2900, 3500 и 6500, среднее арифметическое заработная $\frac{19100}{5} = 3820$. Это означает плата дает нам некоторое представление о размерах заработная плата выбрана, так как она расположена между самым большим и самым маленьким один. При этом 80% зарплаты меньше среднего, поэтому в этом случае это не особенно хорошее представление типичной заработной платы.

Этот случай показывает, что нам нужно обратить внимание к форме распределения и релиа возможность наблюдений до того, как мы воспользуемся среднее арифметическое как мера центрального склонность к определенному набору ценностей. Если абсурдное наблюдение происходит в распределении, среднее арифметическое может дать нерепрезентативное значение для центральной тенденции. Если некоторые наблюдения считать менее надежен, чем другие, это может быть полезно сделать их менее важными. Это может быть делается путем вычисления взвешенной арифметики среднее значение или с помощью медианы, которая не сильное влияние любых абсурдных наблюдений.

ПРИМЕРЫ

В компании A девять сотрудников имеют следующие месячные оклады (в евро):

3000 3200 2900 3440 5050

4150 3150 3300 5200

Среднее арифметическое этих месячных заработная плата составляет:

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

$$\text{"="} \frac{33390}{9} = 3710 \text{ евро.}$$

Теперь рассмотрим случай, когда данные представлена в виде частотного распределения.

Следующая таблица частот дает количество дней, в течение которых 50 сотрудников отсутствие на больничном в течение одного год:

xi : Дни болезни	$_{фи}$: Количество сотрудники
0	7
1	12
2	19
3	8
4	4
Общий	50

Попробуем вычислить среднее количество дней, когда сотрудники отсутствовали по причине болезнь.
Общее количество дней болезни за 50 сотрудников равняется сумме продукта каждого xi на его соответствующую частоту 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

Общее количество сотрудников равно:

$$\sum_{i=1}^{5} \text{фи знак равно } 7 + 12 + 19 + 8 + 4 = 50 .$$

L Среднее арифметическое числа больных дней на одного работника тогда:

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

а это означает, что в среднем 50 сотрудников взяли 1,8 выходных по болезни несс в год.

В следующем примере данные сгруппированы по классам.

Мы хотим вычислить среднее арифметическое ежедневная прибыль от продажи 50 видов бакалея. Распределение частот для продукты представлены в следующей таблице:

Классы (прибыль в евро)	Середина точки xi	Частоты фи_i (число продуктов)	$xi \cdot \text{фи}_i$
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
Общий		50	31950

Среднее арифметическое прибыли равно:

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

это означает, что в среднем каждый из 50 продуктовых магазинов обеспечивают ежедневную прибыль 639 евро.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Среднее геометрическое
- Гармоническое среднее

Иметь в виду

- Мера центральной тенденции
- Средневзвешенное арифметическое

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Плакетт, Р.Л.: Исследования по истории вероятность и статистика. VII. Принцип среднего арифметического. Биометрика 45, 130–135 (1958)

Симпсон, Т.: Письмо достопочтенному Джордж Эрл Маклсфилд, президент Королевское общество, опираясь на ряд наблюдений в практической астрономии. Филос.Транс.Рой. соц. Лонд. 49, 82–93 (1755)

Симпсон, Т.: Попытка показать преимущество, возникающее при использовании среднего значения ряда наблюдений в практической астрономии. В: Разные трактаты о некоторых Любопытные и очень интересные темы в механике, физической астрономии и Спекулятивная математика. Норс, Лондон (1757 г.). стр. 64–75

Арифметический треугольник

Арифметический треугольник используется для определения биномиальные коэффициенты (a + b)n при подсчете числа возможных комбинаций k объектов из общего числа n объектов (Kk н).

ИСТОРИЯ

Понятие нахождения количества комбинаций k объектов из n объектов в сумме исследуется в Индии с девятого век. Действительно, его следы есть в

Machine Translated by Google

Меру Прастара, написанная Пингалой примерно 200 г. до н.э.

Между четырнадцатым и пятнадцатым веками Туриес, аль-Каши, математик из Иранский город Кашан, написал «Ключ к Арифметика. В этой работе он называет биномиальным коэффициенты «элементы экспоненты».

В своей работе Traité du Triangle Arithmétique, опубликовано в 1665 г., Паскаль, Блез (1654 г.) определял числа в «арифметическом тройственном угол», и поэтому этот треугольник также известен как Треугольник Паскаля.

Следует также отметить, что треугольник был сделал популярным Тарталья, Никколо Фонтана в 1556 году, поэтому итальянцы часто называют его Треугольник Тартальи, хотя Тарталья на самом деле не изучал арифметическийтреугольник.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Арифметический треугольник имеет следующее форма:

```
                    1
                1       1
            1 2 1
        1 3 3 1
    1 4 6 4 1
1   5   10  10  5   1
1   6   15  20  15  6  1
            ...
```

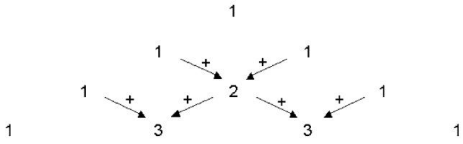
Каждый элемент представляет собой биномиальный коэффициент

$$C_{н}^{к} = \frac{н!}{к!(п - к)!}$$

$$= \frac{п \cdot (п - 1) \cdot \quad \dots \quad \cdot (п - к + 1)}{1 \cdot 2 \cdot \dots \cdot к} .$$

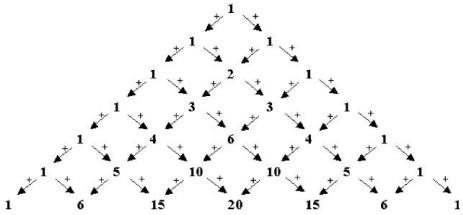
Этот коэффициент соответствует элементу k строки n + 1, k = 0,...,n.

Любое конкретное число получается сложением вместе соседние с ним числа в предыдущей строке.



Например:

$$C4_{6} = C3_{5} + C4_{5} = 10 + 5 = 15 .$$



В более общем виде имеем соотношение:

$$C_{н}^{к} + K_{н}^{к+1} = C_{п+1}^{к+1} ,$$

потому что:

$$C_{н}^{к} + K_{н}^{к+1} = \frac{н!}{(п - к)! \cdot к!} + \frac{н!}{(п - к - 1)! \cdot (к + 1)!}$$

$$= \frac{н! \cdot [(к + 1) + (п - к)]}{(п - к)! \cdot (к + 1)!}$$

$$= \frac{(п + 1)!}{(п - к)! \cdot (к + 1)!}$$

$$= C_{п+1}^{к+1} .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Биномиальный
- Биномиальное распределение
- Комбинация
- Комбинационный анализ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Паскаль, Б.: Трактат по арифметике треугольников (опубликовано посмертно в 1665 г.), Париж (1654 г.)

Паскаль, Б .: Œuvres, тт. 1–14. Брун Швик, Л., Бутру, П., Газье, Ф. (ред.)

Machine Translated by Google

LesGrandsEcrivainsdeFrance.Hachette,
Париж (1904–1925)

Паскаль, Б .: Меснар, Дж. (ред.) Œuvres com plètes.
Том. 2. Дескле де Брауэр, Париж
(1970)

Рашед, Р .: La naissance de l'algèbre. В:
Ноэль, Э. (ред.) Le Matin des Mathémati ciens.
Белин-Радио Франция, Париж (1985)
Глава. 12).

Юшкевич, А.П.: Математика.
арабы (VIIIème-XVème siecles). Частичный
перевод Казенав М., Жауиш К.
Врин, Париж (1976)

АРМА Модели

Модели ARMA (иногда называемые моделями Бокса Дженкинса) являются авторегрессивными движущимися средние модели, используемые при анализе временных рядов. Авторегрессионная часть, обозначенная AR, состоит из конечной линейной комбинации предыдущие наблюдения. Часть скользящей средней, MA, состоит из конечного линейного сочетание в t предыдущих значений для белый шум (последовательность взаимно независимых и одинаково распределенных случайных переменные).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

1. Модель дополненной реальности (авторегрессионная)

В авторегрессионном процессе порядка p настоящее наблюдение yt генерируется средневзвешенное значение прошлых наблюдений до p-го периода. Это принимает следующий вид:

$$\text{AR(1): } y_t = \theta_1 y_{t\ \ 1} + \varepsilon_t ,$$

$$\text{AR(2): } y_t = \theta_1 y_{t\ \ 1} + \theta_2 y_{t\ \ 2} + \varepsilon_t ,$$

$$\vdots$$

$$\text{AR(p): } y_t = \theta_1 y_{t\ \ 1} + \theta_2 y_{t\ \ 2} + \ldots + \theta_p y_{t\ \ p} + \varepsilon_t ,$$

где θ1, θ2,..., θp – положительные или отрицательные параметры, подлежащие оценке и εt — коэффициент ошибки, который следует нормальному распределению.

2. Модель MA (скользящее среднее)

В процессе скользящего среднего порядка q каждое наблюдение yt генерируется случайным образом взвешенным средним арифметическим до тех пор, пока q -й период:

$$\text{MA(1): } y_t = \varepsilon_t \quad \alpha_1 \varepsilon_{t\ \ 1}$$

$$\text{MA(2): } y_t = \varepsilon_t \quad \alpha_1 \varepsilon_{t\ \ 1} \quad \alpha_2 \varepsilon_{t\ \ 2}$$

$$\cdots$$

$$\text{MA(p): } y_t = \varepsilon_t \quad \alpha_1 \varepsilon_{t\ \ 1} \quad \alpha_2 \varepsilon_{t\ \ 2} \quad \cdots \quad \alpha_q \varepsilon_{t\ \ q} ,$$

где α1, α2,..., αq — положительные или отрицательные параметры, εt — гауссова случайная ошибка.

Модель MA представляет собой временной ряд колеблется вокруг своего среднего значения случайным образом образом, что приводит к термину «скользящее среднее», потому что оно сглаживает ряд, вычитая белый шум, генерируемый случайностью элемента.

3. Модель ARMA (авторегрессионная скользящая средняя модель)

Модели ARMA представляют собой процессы, созданные из комбинации прошлых значений. и прошлые ошибки. Они определяются следующее уравнение:

$$\text{APMA(p, q): } y_t = \theta_1 y_{t\ \ 1} + \theta_2 y_{t\ \ 2} + \ldots + \theta_p y_{t\ \ p} + \varepsilon_t \quad \alpha_1 \varepsilon_{t\ \ 1} \quad \alpha_2 \varepsilon_{t\ \ 2} \quad \cdots \quad \alpha_q \varepsilon_{t\ \ q} ,$$

при θp = 0, αq = 0 и (εt, t Z) есть слабый белый шум.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Временная последовательность

Средневзвешенное арифметическое

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Коробка, GEP, Дженкинс, GM: временные ряды Анализ: прогнозирование и управление (серия inTimeSeriesAnalysis).HoldenDay,San Франциско (1970)

Договоренность

Аранжировки — это понятие, встречающееся в комбинаторном анализе.
Количество аранжировок – это количество способов рисования k объектов из n объектов где порядок расположения объектов учитывается нарисованное (в отличие от комбинации).

ИСТОРИЯ

См. комбинаторный анализ.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

1. Аранжировки без повторов

Аранжировка без повторения относится ситуации, когда нарисованные объекты не помещаются обратно для следующего рисунка. Каждый объект может быть нарисован только один раз во время k розыгрышей.
Количество размещений k объектов среди n без повторения равно:

$$A_n^k = \frac{n!}{(n-k)!}.$$

2. Аранжировки с повторениями

Аранжировки с повторением возникают, когда каждый вытащенный предмет помещается обратно для следующего рисунка. Тогда каждый объект может быть нарисовано r раз из k рисунков, r = 0, 1,...,k.

Количество размещений k объектов среди n с повторениями равно от n до мощность к:

$$A_n^k = n^k .$$

ПРИМЕРЫ

1. Аранжировки без повторов

Рассмотрим урну с шестью шарами, пронумерованными от 1 до 6. Вытащим четыре шара. от урнин преемственности, и мы хотим узнать, сколько чисел можно формируются из числа выпавших шаров.
Тогда нас интересует количество договоренности (поскольку мы учитываем порядок шаров) без повторения (поскольку каждый шарик можно вытащить только один раз) четырех объектов из шести. Мы получать:

$$A_n^k = \frac{n!}{(n-k)!} = \frac{6!}{(6-4)!} = 360$$

возможные договоренности. Следовательно, это можно составить 360 различных чисел вытягивая четыре числа из чисел 1, 2, 3, 4, 5, 6, когда каждое число может появляются только один раз в четырехзначном числе сформировался.

В качестве второго примера рассмотрим аранжировки без повторений двух буквы из букв A, B и C. C
n = 3 и k = 2 имеем:

$$A_n^k = \frac{n!}{(n-k)!} = \frac{3!}{(3-2)!} = 6 .$$

Затем мы получаем:

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

2. Аранжировки с повторениями

Рассмотрим ту же урну, что описана ранее. Мы делаем четыре последовательных розыгрыша, но на этот раз мы кладем каждый вытащенный шар обратно в урну.

Machine Translated by Google

Мы хотим знать, сколько четырехзначных чисел (или расположения) возможны, если выпадают четыре номера.

В этом случае мы исследуем количество аранжировок с повторением (поскольку каждый шар помещается обратно в урну до следующий рисунок). Мы получаем

$$A^{k}_{n} = n^{k} = 64 = 1296$$

разные аранжировки. можно составить 1296 четырехзначных чисел из числа 1,2,3,4,5,6, если каждое число может встречаются более одного раза в четырехзначном число.

В качестве второго примера снова возьмем три буквы A, B и C и образуют расположение двух букв с повторами. При n = 3 и k = 2 имеем:

$$A^{k}_{n} = n^{k} = 32 = 9 .$$

Затем мы получаем:

AA, AB, AC, БА, ВВ, ВС, СА, СВ, СС.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Комбинация
- Комбинационный анализ
- Перестановка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

См. комбинаторный анализ.

Свойственный риск

Приписываемый риск – это разница между риском , с которым сталкиваются лица, подвергающиеся воздействию определенного фактора, и риску, с которым сталкиваются лица, не подвергается этому. Это наоборот, чтобы избежать риска. Он измеряет абсолютный эффект причина (то есть чрезмерный риск или случаи болезнь).

ИСТОРИЯ

См. риск.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

По определению имеем:

относимый риск = риск для тех, кто подвергается

риск для тех, кто не подвергается воздействию .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Доверительный интервал атрибутивного риск эквивалентен доверительному интервалу разницы между пропорциями pE и pNE, где pE и pNE представляют риски, с которыми сталкиваются лица, подвергающиеся и не подверженные действию изучаемого фактора соответственно. Возьмем nE и nNE соответственно размеру экспонированного и неэкспонированного населения. Тогда для уровня достоверности (1 α), определяется как:

$$(pE - pNE) \pm z\alpha \sqrt{\frac{pE\cdot(1-pE)}{nE} + \frac{pNE\cdot(1-pNE)}{nCB}},$$

где zα — значение, полученное из нормального таблица (например, для доверительного интервала 95 %, α = 0,05 и zα = 1,96). Доверительный интервал для (1 α) предотвратимой риск имеет границы, заданные:

$$(pNE - pE) \pm z\alpha \sqrt{\frac{pE\cdot(1-pE)}{nE} \cdot \frac{pNE\cdot(1-pNE)}{nNE}} .$$

Здесь nE и nNE должны быть большими. Если доверительный интервал включает ноль, мы не можем из-за отсутствия атрибутивного риска.

ПРИМЕРЫ

В качестве примера рассмотрим исследование риск рака молочной железы у женщин из-за курения инг:

Machine Translated by Google

Группа	Заболеваемость ставка (/100000 /год)	Относится к риску от курения (/100000 /год)
Нонекс поставленный	57,0	57,0 - 57,0 = 0
Пассивный курильщики	126,2	126,2 - 57,0 = 69,2
Активный курильщики	138,1	138,1 - 57,0 = 81,1
Общий	114,7	114,7 - 57,0 = 57,7

Риски, связанные с пассивным и активное курение соответственно 69 и 81 (/100000 лет). Другими словами, если воздействие табака было устранено, уровень заболеваемости среди активных курильщиков (138/100 000 в год) может быть уменьшена на 81/100000 в год, а для пассивных курильщиков (126/100000 в год) на 69/100000 в год. Показатели заболеваемости в обе категории курильщиков станут равен показателю для не подвергавшихся воздействию женщин (57/100000 в год). Обратите внимание, что уровень заболеваемости для не подвергавшихся воздействию женщин не ноль, из-за влияния других факторов помимо курения.

№ группы	индив. наблюдаемый более двух годы	Случаи атрибут к курение (для два года период)	Случаи атрибут к курение (за год)
Нонекс поставленный	70160	0,0	0,0
Пассивный дым эры	110860	76,7	38,4
Активный дым эры	118636	96,2	48.1
Общий	299656	172,9	86,5

Мы можем подсчитать количество случаев рак молочной железы, связанный с воздействием табака, путем умножения числа лиц, наблюдаемых в год, на атрибутивную риск. Разделив количество инцидентов приходящихся на курение в двухлетнем периоде на два, получаем число случаев приходится на курение в год, и мы можем затем определите риск, связанный с дымом в популяции, обозначенный PAR, как показано в следующем примере. Предыдущая таблица показывает детали расчета.

Здесь мы описываем исчисление для пассивных курильщиков. В двухлетнем исследовании, Под наблюдением находилось 110860 пассивных курильщиков. Риск, связанный с пассивным курением был 69,2/100000 в год. Это значит, что количество случаев, связанных с курением ing за двухлетний период составляет (110860 · 69,2)/100000 = 76,7. Если мы хотим подсчитать количество случаев пассивного курения в год, мы должны разделить последнее значение на 2, получив 38,4. Более того, мы можем рассчитать риск, связанный с к курению в год, просто разделив число случаев, связанных с курением за двухлетнего периода (172,9) по количеству лиц, изученных за эти два года (299656 человек). Тогда мы получаем риск в связи с курением – 57,7/100 000 чел. год. Заметим, что мы можем получить тот же результат взяв разницу между общим заболеваемость (114,7/100000 в год, см. примеры под записями для совпадения скорость, уровень распространенности) и заболеваемость показатель неэкспонированной группы (57,0/100000 в год).

Риск рака молочной железы, связанный с курение среди населения (PAR) – отношение числа случаев рака молочной железы cer, связанный с воздействием табака и число случаев диагностированного рака молочной железы.

Machine Translated by Google

нос среди населения (см. вышеприведенную таблицу). Атрибутивный риск в популяции 22,3% (38,4/172) за пассивное курение и 28% (48,1/172) за активное курение. Для обоих форм облучения, она составляет 50,3 % (22,3 % + 28%). Так, половина случаев рака груди диагностируется каждый год в этой популяции курение (активное или пассивное).

Групповой случай	атрибут к курение (на два года)	Случай атрибут к курение (в год)	PAR
Нонекс поставленный	0,0	20	0,0
Пассивный дым эры	38,4	70	22,3
Активный дым эры	48.1	82	28,0
Общий	86,5	172	50,3

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Предотвратимый риск
- Причина и следствие в эпидемиологии
- Уровень заболеваемости
- Шансы и отношение шансов
- Уровень распространенности
- Относительный риск
- Риск

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Корнфилд, Дж.: Метод оценки сравнительных показателей на основе клинических данных. Применение при раке легкого, молочной железы и шейка матки. Дж. Натл. Рак инст. 11, 1269–1275 гг. (1951)

Лилиенфельд, А.М., Лилиенфельд, Д.Э.: Основы эпидемиологии, 2-е изд. Кларен Дон, Оксфорд (1980)

MacMahon, B., Pugh, TF: Эпидемиология: Принципы и методы. Маленький Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Морабия, А .: L'Epidémiologie Clinique. Издания «Que sais-je?». Presses Universitaires de France, Париж (1996)

Автокорреляция

Автокорреляция, обозначаемая ρk, является мерой корреляция определенного временного ряда с тем же временным рядом, задержанным на k лагов (расстояние между наблюдениями, которые так соотносятся). Его получают путем деления ковариация между двумя наблюдениями, разделенные k лагами , временного ряда (автоковариация) на стандартное отклонение yt и yt-k. Если вычислить автокорреляцию для всех значений k получаем автокорреляцию функция. Для временного ряда, который не при изменении во времени автокорреляционная функция экспоненциально убывает до 0.

ИСТОРИЯ

Первое исследование автокорреляции, Частичная автокорреляция и коррелограмма исполнялась в 1920-х и 1930-х годах Юл, Джордж, разработавший теорию авторегрессивные процессы.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Определим автокорреляцию временных рядов Yt от:

$$\rho_k = \frac{\text{ков}\,(y_t, y_{t-k})}{\sigma_{y_t}\sigma_{y_t}\quad k}$$

Machine Translated by Google

$$\text{"="} \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=k+1}^{T} (y_t - \bar{y})^2 \sum_{t=k+1}^{T} (y_{t-k} - \bar{y})^2}.$$

Здесь $\bar{y}$ — среднее значение ряда, рассчитанного по $T-k$ лагов, где T – число наблюдений.

Выясняем, что:

$$\rho_0 = 1 \text{ и}$$

$$\rho_k = \rho_{-k}.$$

Можно оценить автокорреляцию (обозначается ρk) при достаточно большом количестве наблюдений (T > 30) с использованием следующая формула:

$$\rho_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}.$$

Частичная автокорреляционная функция для задержка k лагов определяется как автокорреляция между yt и yt k, влияние других переменных смещается на k лагов

(yt 1, yt 2,...,yt k+1).

Проверка гипотезы

При анализе автокорреляционной функции временного ряда может быть полезно знать члены ρk , существенно отличающиеся от 0. Затем проверка гипотезы продолжается как следует:

$$H_0 : \rho_k = 0$$

$$H_1 : \rho_k = 0 .$$

Для большой выборки (T > 30) коэффициент ρk асимптотически стремится к нормальному распределению со средним значением 0 и стандартным отклонением $\frac{1}{\sqrt{T}}$ ация . Студенческое яичко основано на сравнение эмпирического t и теоретического t.

Доверительный интервал для коэффициента ρk дан кем-то:

$$\rho_k = 0 \pm t_{\alpha/2} \frac{1}{\sqrt{T}}.$$

Если рассчитанный коэффициент ρk не падает в пределах этого доверительного интервала она достоверно отличается от 0 на уровне α (в общем случае α = 0,05 и tα/2 = 1,96).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

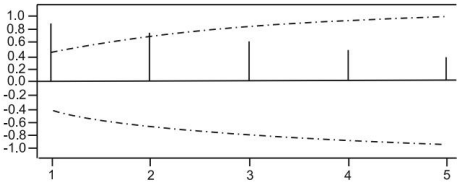
Частичная автокорреляционная функция в основном используется при изучении временных рядов и, точнее, когда мы хотим настроить АРМАмодель. Эти функции также используются в пространственной статистике, хотя в контексте пространственной автокорреляции, где мы исследуем корреляцию переменной с самой собой в космосе. Если наличие феномена конкретный регион пространства влияет на вероятность того, что явление присутствует в соседних регионах явление имеет пространственную автокорреляцию. В этом случае, Положительная автокорреляция имеет место, когда соседние регионы, как правило, имеют идентичные свойства или близкие значения (примеры включают однородные области и регулярные градиенты). Возникает отрицательная автокорреляция когда соседние регионы имеют разные качества или чередуются между сильными и слабые значения явления. Показатели автокорреляции зависят от масштабирования переменные, которые используются в анализе а также на сетке, регистрирующей наблюдения.

ПРИМЕРЫ

Возьмем в качестве примера среднюю заработную плату по стране в Швейцарии с 1950 по 1994 г. измеряется каждые два года.

Machine Translated by Google

Рассчитываем автокорреляционную функцию между данными; мы хотели бы найти положительную автокорреляцию. Следующие цифры показывают наличие этой автокорреляции.



Отметим, что корреляционная значимость пики между наблюдением в момент времени t и наблюдение в момент времени 1, а также между наблюдение в момент времени t и наблюдение в момент времени t 2. Такая конфигурация данных типична для авторегрессионного процесса. Fortwo первый значений, мы можем видеть, что эта автокорреляция является значимым, поскольку статистика Стьюдента t для T = 23 наблюдений дает:

$$\text{pk} = 0 \pm 1{,}96 \quad \frac{1}{\sqrt{23}}.$$

Год	Средняя заработная плата по стране
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

Год	Средняя заработная плата по стране
82	42108
84	44095
86	48323
88	51584
90	55480
92	54348
94	54316

Источник: Федеральное статистическое управление Швейцарии.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Студенческий тест

Временная последовательность

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бурбонне, Р .: Econométrie, manuel et упражнения с исправлениями, 2-е изд. Дюно, Париж (1998)

Коробка, GEP, Дженкинс, GM: временные ряды Анализ: прогнозирование и управление (серия inTimeSeriesAnalysis).HoldenDay,San Франциско (1970)

Чатфилд, К.: Анализ временных рядов: Введение, 4-е изд. Чепмен и Холл (1989)

Предотвратимый риск

Предотвратимый риск (который, конечно, предотвратимого, если мы нейтрализуем эффект воздействия определенного явления) противоположно атрибутивному риску. Другими словами, это разница между риском столкновения неэкспонированными лицами, и что с которыми сталкиваются лица, подвергшиеся воздействию явление.

ИСТОРИЯ

См. риск.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

По определению имеем:

предотвратимый риск = риск, если не подвергаться

риск при воздействии .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Предотвратимый риск был введен для того, чтобы

чтобы избежать необходимости определения отрицательного атрибутивный риск. Он позволяет нам вычислить

количество пациентов, которое необходимо

лечится, потому что:

$$\text{Количество пациентов лечиться} = \frac{1}{\text{Предотвратимый риск}}.$$

См. также сопутствующий риск.

ПРИМЕРЫ

В качестве примера рассмотрим исследование эффективности лекарства, используемого для лечения болезни. 223 пациента, включенных в исследование, все рискуют заразиться, но еще не сделали этого. Мы разделяем их на две группы: пациенты первой группы (114 пациентов) получали препарат; те, кто в второй группе (109 больных) давали плацебо. Срок обучения составлял два года. Всего диагностировано 11 случаев заболевания. в первой группе и 27 в группе плацебо.

Группа	Случаи из болезнь	Количество пациентов в группа	Риск для в два года период
	(А)	(Б)	(А / Б в %)
1 группа 11		114	9,6%
2 группа 27		109	24,8%

Таким образом, предотвратимый риск, связанный с приемом препарата, составляет 24,8 .

9,6 = 15,2% за два года.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Атрибутивный риск

Причина и следствие в эпидемиологии

Уровень заболеваемости

Шансы и отношение шансов

Уровень распространенности

Относительный риск

Риск

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Корнфилд, Дж.: Метод оценки сравнительных показателей на основе клинических данных. Применение при раке легкого, молочной железы и шейка матки. Дж. Натл. Рак инст. 11, 1269–1275 гг. (1951)

Лилиенфельд, А.М., Лилиенфельд, Д.Э.: Основы эпидемиологии, 2-е изд. Кларен Дон, Оксфорд (1980)

MacMahon, B., Pugh, TF: Эпидемиология: Принципы и методы. Маленький Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Морабия, А .: L'Epidémiologie Clinique. Издания «Que sais-je?». Presses Universitaires de France, Париж (1996)

Machine Translated by Google

Гистограмма

Гистограмма — это тип количественного графика. Он состоит из ряда вертикальных или горизонтальных стержни одинаковой ширины, но с длиной относительно представленных количеств.

Гистограммы используются для сравнения категорий категориальной качественной переменной или для сравнения наборов данных за разные годы или из разных мест для конкретной переменной.

ИСТОРИЯ

См. графическое изображение.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Вертикальная ось и горизонтальная ось должны быть определено для построения вертикальной полосы диаграмма.

Горизонтальная ось разделена на разные категории; вертикальная ось показывает значение каждой категории.

Для построения горизонтальной гистограммы оси просто инвертируются.

Все полосы должны быть одинаковой ширины, т.к. сравниваются только их длины.

Для облегчения понимания графика можно использовать затенение, штриховку или цвет.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Столбчатая диаграмма также может использоваться для представления значений отрицательных категорий. Для этого шкала оси, показывающая значения категории, должна быть ниже нуля.

Существует несколько типов гистограмм. Тот, что описан выше, называется простой гистограммой.

Гистограмма с несколькими столбцами используется для сравнения нескольких переменных.

Составная столбчатая диаграмма — это составная столбчатая диаграмма, в которой различные наборы данных сложены друг с другом. друг над другом. Этот тип диаграммы используется, когда различные наборы данных могут быть объединены. объединены в общую популяцию, и мы хотели бы сравнить изменения в наборах данных и общей популяции с течением времени.

Существует еще один способ представления подмножеств всего населения. В этом случае общая численность населения составляет 100%, а значение, данное для каждого подмножества, представляет собой процент от общего числа (см. также круговую диаграмму).

ПРИМЕРЫ

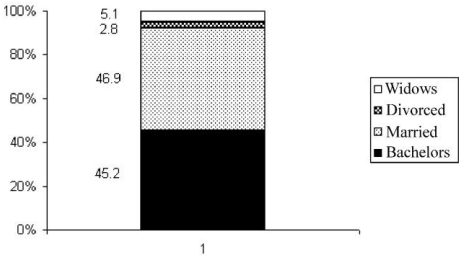
Давайте построим столбчатую диаграмму, разделенную на проценты для данных в следующей таблице частот:

Machine Translated by Google

Семейное положение в выборке австралийца женского населения на 30 июня 1981 г. (в тысячи)

Семейный положение дел	Фре качество	Родственник свободный качество	Процент возраст
Бакалавр	6587,3	0,452	45,2
Женатый	6836,8	0,469	46,9
Разведен	403,5	0,028	2,8
вдова	748,7	0,051	5.1
Общий	14576,3	1.000	100,0

Источник: ABS (1984), Австралийский карманный год. Книга. Австралийское статистическое бюро, Канбер, р. 11



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Графическое представление
- Количественный график

Барнард, Джордж А.

Барнард, Джордж Альфред родился в 1915 г. в Уолтемстоу, Эссекс, Англия. Он получил степень по математике в Кембридже Университет в 1936 г. С 1942 по 1945 г. он работал в Министерстве снабжения в качестве научного консультанта. Барнард работал на математическом факультете Имперского колледжа Лондона с 1945 по 1966 год. С 1966 по 1975 год он был профессором математики в универе университета Эссекса, а с 1975 г. до его на пенсии в 1981 году онбыл профессором Статистики тики в Университете Ватерлоо, Канада.

Барнард, Джордж Альфред получил многочисленные отличия, включая золотую медаль Королевского статистического общества и Института математики и ее приложений. В 1987 году он был назван почетным членом Международный статистический институт. Он умер в 2002 году в августе.

Некоторые статьи Барнарда, Джорджа Альфреда:

1954 г. Выборочный контроль и статистические решения. Дж. Рой. Стат. соц. сер. Б 16, 151–174.

1958 Томас Байес - Биографическая справка. Биометрика 45, 293–315.

1989 г. О предполагаемом приходе к власти низших р-значения. Стат. Мед., 8, 1469–1477.

1990 Должны ли клинические испытания быть большими? интерпретация р-значений и комбинирование результатов испытаний. Стат. мед., 9, 601–614.

Теорема Байеса

Если мы рассмотрим совокупность «причин», которые событие происходит, теорема Байеса дает формулу вероятности того , что событие является прямое следствие определенной причины. Поэтому теорему Байеса можно интерпретировать как формулу условной вероятности события.

ИСТОРИЯ

Теорема Байеса названа в честь Байеса, Томаса и был разработан в середине восемнадцатого века. Тем не менее, Байес не опубликовал теорему при жизни; вместо этого он был представлен Прайсом, Р. 23 декабря 1763 г., через два года после его смерти Лондонскому королевскому обществу,

Machine Translated by Google

членом которого Байес был во время последние двадцать последних лет своей жизни.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть {A1, A2,...,Ak} — разбиение пространства выборки . Мы предполагаем, что каждое событие A1,...,Ak имеет ненулевую вероятность. Пусть E событие такое, что P(E) > 0.
Итак, для каждого i (1 i k) теорема Байеса (для дискретного случая) дает:

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

В непрерывном случае, когда X — случайная величина с функцией плотности f(x), также называется априорной функцией плотности, Теорема Байеса дает плотность апостериорно в соответствии с

$$e(x \mid E) = \frac{f(x) \cdot P(E|X = x)}{\int f(t) \cdot P(E|X = t)dt}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Теорема Байеса была предметом многих споры, касающиеся возможности его использования когда значения вероятностей, используемые для определяют функцию вероятности апостериорно, как правило, не устанавливаются в точном способ.

ПРИМЕРЫ

В трех урнах находятся красные, белые и черные шары:

- Урна A содержит 5 красных шаров, 2 белых шара. и 3 черных шара;
- Урна B содержит 2 красных шара, 3 белых шара. и 1 черный шар;
- Урна C содержит 5 красных шаров, 2 белых шара. и 5 черных шаров.

Случайным образом выбрав урну, рисуем шар наугад: он белый. Мы хотим определить вероятность того , что оно было взято из урны A.
Пусть A1 соответствует событию, когда мы «выбрать урну A», A2 — событие, в котором мы «выберите урну B», а A3 — событие, когда мы «выбираем урну C». {A1, A2, A3} образует раздел выборочного пространства.
Пусть E будет событием, когда «взятый мяч белый», который имеет строго положительную вероятность.
У нас есть:

$$P(A1) = P(A2) = P(A3) = \frac{1}{3},$$

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

Формула Байеса позволяет определить вероятность того, что вытащенный белый шар выпадет из урны A:

$$P(A1|E) = \frac{P(A1) \cdot P(E|A1)}{\sum_{i=1}^{3} P(Ai) \cdot P(E|Ai)}$$

$$= \frac{\frac{1}{3} \cdot \frac{2}{10}}{\frac{1}{3} \cdot \frac{2}{10} + \frac{1}{3} \cdot \frac{3}{6} + \frac{1}{3} \cdot \frac{2}{12}} = \frac{3}{13}.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Условная возможность
- Вероятность

ССЫЛКА

Байес, Т.: Очерк решения проблемы в учении о шансах. Филос.
Транс. Рой. соц. Лонд. 53, 370–418 (1763 г.). Опубликовано по инициативе Прайс, Р., через 2 года после его смерти. Переиздано с биографией Барнарда, Джордж А. в 1958 году и в Пирсоне, ES, Кендалл, MG: Исследования по истории Статистика и вероятность. Гриффин, Лондон, стр. 131–153 (1970).

Machine Translated by Google

Байес, Томас

Байес, Томас (1702–1761) был старшим сыном Байеса, Джошуа, который был одним из первых шести священников-нонконформистов, которые были рукоположены в Англии и был членом Королевского общества. Он получил частное образование у профессоров, как это было принято в семьях нонконформистов. В 1731 году он стал преподобным

Пресвитерианская часовня в Танбридж-Уэллсе, городе, расположенном примерно в 150 км к юго-западу от Лондон. Благодаря некоторым религиозным публикациям он был избран членом Королевского общества в 1742 году.

Его интерес к математике был хорошо известен своим современникам, несмотря на то, что он не написал никаких технических публикаций, потому что его наставником был де Муавр А., один из основателей теории вероятностей. В 1763 году Прайс Р. разобрал документы, оставленные Байесом, и опубликовал свою основную работу:

1763 г. Очерк решения проблемы учения о шансах.

Филос. Транс. Королевский соц. Лондон, 53, стр. 370–418. Переиздано с биографией Барнарда, Джорджия (1958).
В: Пирсон, Э.С. и Кендалл, М. (1970). Исследования по истории статистики и вероятностей. Гриффин, Лондон, стр. 131–153.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

теорема Байеса

Байесовская статистика

Байесовская статистика - это большая область в области статистики, которая отличается из-за аксиомы

атизацию статистики, которая придает ей определенную внутреннюю согласованность.

Основная идея состоит в том, чтобы интерпретировать вероятность события так , как она обычно используется; другими словами, как связанная с ним неопределенность. Напротив, классический подход считает вероятность события пределом относительной частоты (более формальный подход см. в разделе Вероятность).

Наиболее известным аспектом байесовского вывода является вероятность вычисления совместного распределения вероятностей (или функции плотности) f(θ, X = x1,...,X = xn) одного или многих параметров θ (один параметр или вектор параметры), наблюдая за данными x1,...,xn , выбранными независимо от случайной величины X , от которой зависит θ . (Стоит отметить, что он также позволяет вычислить распределение вероятностей для нового наблюдения xn+1).

Байесовская статистика рассматривает неизвестные параметры как случайные величины не из-за возможная изменчивость (на самом деле неизвестные параметры считаются фиксированными), а из-за нашего незнания или неуверенности в них.

Апостериорное распределение f(θ |X = x1,..., X = xn) вычисляется напрямую, поскольку оно представляет собой априорное (f(θ)) умноженное на вероятность f(X = x1,...,X = xn |θ).

апостериорная предыдущая x вероятность

Второе выражение не вызывает проблем, потому что это функция, которую мы часто используем в классической статистике, известная как подобная вероятность (см. максимальное правдоподобие).

Напротив, первая часть предполагает априорное распределение для θ. Мы часто используем начальную распределение θ для включения возможной дополнительной информации об интересующих параметрах. При отсутствии этой информации

Для этого мы используем эталонную функцию, которая максимизирует недостаток информации (которая в таком случае является наиболее «объективной» или «неинформативной»).

Machine Translated by Google

функцию, следуя общему, но не точному использованию).

После вычисления распределения f(θ |x1,...,xn) становится доступной вся информация об интересующих параметрах. Таким образом, мы можем вычислить правдоподобные значения неизвестного параметра (среднего , медианы или какой-либо другой меры центральной десятки) . плотности), его стандартное отклонение, доверительные интервалы или выполнить проверку гипотез по его значению.

ИСТОРИЯ

См. Байеса, Томаса и теорему Байеса.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть D будет набором данных X = x1,...,X = xn , независимо выбранных из случайной величины X с неизвестным распределением. Мы будем рассмотрим простой случай, когда есть только один интересный параметр, θ, который зависит от X.

Тогда стандартная байесовская процедура может быть выражена следующим образом: 1. Определите известные величины x1,...,xn.

2. Укажите модель данных; другими словами, параметрическое семейство f (x | θ) распределений, описывающее генерацию данных.

3. Задайте неопределенность, касающуюся θ, с помощью функции начального распределения f (θ).

4. Затем мы можем рассчитать распределение f (θ |D) (называемое окончательным распределением) с использованием теоремы Байеса.

Первые два пункта являются общими для любого статистического вывода.

Третий пункт более проблематичен. При отсутствии дополнительной информации о θ идея состоит в том, чтобы вычислить эталонное распределение f (θ) путем максимизации функции, задающей недостающую информацию о параметр θ. Как только эта проблема решена, четвертый пункт легко решается с помощью теоремы Байеса.

Теорема Байеса может быть выражена в непрерывной форме следующим образом:

$$D) = f (D) f (D \mid \frac{f (D \mid \theta) \cdot f (\theta) f (\theta \mid}{\theta) \cdot f (\theta) f (D \mid} \quad "=" \quad \frac{\theta)f (\theta) d\theta}{} .$$

Поскольку xi независимы, мы можем написать:

$$f (\theta \mid D) = \frac{\prod_{i=1}^{n} f(xi \mid \theta) \cdot f (\theta)}{\prod_{i=1}^{n} f(xi \mid \theta)f (\theta) d\theta} .$$

Теперь у нас есть средства для расчета логова функция нового (независимого) наблюдения xn+1 при заданных x1,...,xn:

$$f(X = xn+1 \mid D) f (X = \quad "=" \quad \frac{x1,...,X = xn+1) f (X = x1,...,X = xn)}{}$$

$$"=" \quad \frac{f (X = x1,...,X = xn+1 \mid \theta) \cdot f (\theta) d\theta}{f (D \mid \theta) \cdot f (\theta) d\theta}$$

$$"=" \quad \frac{f (X = xi \mid \theta) \cdot f (\theta) d\theta}{\prod_{i=1}^{n} f (X = xi \mid \theta) \cdot f (\theta) d\theta}$$

$$= f (X = xn+1 \mid \theta) \cdot f (\theta \mid D) d\theta .$$

Сейчас мы кратко объясним методы, которые позволяют нам:

- найти значение оцениваемого параметра, более вероятное, чем остальные;
- Найдите доверительный интервал для θ, и;
- Выполните проверку гипотез.

Эти методы тесно связаны с теорией принятия решений, которая играет значительную роль в байесовской статистике.

Machine Translated by Google

Точечная оценка параметра

Чтобы получить точечную оценку параметра θ, зададим функцию потерь l(θˆ, θ), которая происходит от использования θˆ (оценочное значение) вместо истинного значения (неизвестного) θ. Затем минимизируем эту функцию от θˆ. Для Например, если θ действительно и «проигрыш квадратичен» (т. е. l(θˆ, θ) = (θˆ θ)2), то точка оценка θ будет средним значением калибровки кулированное распределение.

Доверительные интервалы

Понятие доверительного интервала заменен в байесовской статистике понятием α-достоверного региона (где α — «уровень достоверности»), который просто определяется как интервал I такой, что:

$$\int_{я} f(\theta \mid D)\, d\theta = \alpha .$$

Часто мы также требуем, чтобы ширина интервал сведен к минимуму.

Проверка гипотезы

Общий подход к проверке гипотез:

H0 : θ I по сравнению с H1 : θ / I ·

относится к теории принятия решений. Основываясь на этом, мы определяем функцию потерь l(a0, θ), которая принимает H0 (где истинное значение параметра равно θ) и функцию потерь l(a1, θ), отвергающую H0. Если значение для истинного значения, полученного принимая H0, т.е.

$$l(a0, \theta)d\theta ,$$

меньше, чем полученная при отказе от H0, то можно принять H0.
Используя это ограничение, мы отвергаем ограничения, налагаемые на θ нулевой гипотезой (θ I).

ПРИМЕРЫ

Следующий пример включает оценку параметр θ из бернуллиевского распределения X с помощью n независимых наблюдения x1,...,xn, принимающие значение 1 в в случае успеха и 0 в случае неудачи.
Пусть r — количество успехов, а n r – число отказов среди наблюдений.

Имеем тогда:

$$L(\theta) = P(X = x1,...,X = xn \mid \theta) \overset{н}{=} \prod_{я=1} P(X = xi \mid \theta) = \theta r \cdot (1 \theta)^{n-p} .$$

Оценка максимальной вероятности θ, обозначаемый θˆ mle, максимизирует эту функцию. Для этого рассмотрим логарифм этой функции, другими словами, логарифмическая вероятность:

$$\log L(\theta) = r\log(\theta) + (n r)\log(1 \theta).$$

Мы максимизируем это, устанавливая его производную θ равно нулю:

$$\frac{(\log L)}{\theta} \overset{"="}{} \frac{p}{\theta} \frac{n - p}{1 - \theta} = 0 .$$

Это эквивалентно r(1 θ) = (n r) θ, и упрощается до θˆ мл = $\frac{p}{н}$. Оценщик для максимальной вероятности θ тогда sim ply доля наблюдаемых успехов.
Теперь вернемся к байесовскому методу.
В этом случае распределение Бернулли, априорное эталонное распределение параметра θ, выражается как:

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

где c - соответствующая константа (такая как f(θ)dθ = 1) и где h(θ) называется

Machine Translated by Google

Информация Фишера X:

$$h(\theta) = E\left[\frac{2л}{\theta^2}\right]$$

$$= -E\left[\theta^2 \frac{Икс}{}\right] \frac{1 - X}{(1 - \theta)^2}$$

$$"=" \frac{Э[X]}{2} + \frac{1 - Э[X]}{(1 - \theta)^2}$$

$$"=" \frac{\theta}{\theta^2 (1 - \theta)} + \frac{1 - \theta}{(1 - \theta)^2}$$

$$"=" \frac{1}{\theta} + \frac{1}{1 - \theta} "=" \frac{1}{\theta (1 - \theta)} .$$

Тогда распределение будет:

$$f(\theta) = c \cdot \theta \quad 1\,2 \cdot (1 \quad \theta)^{\frac{1}{2}} .$$

Функция распределения θ, учитывая наблюдения x1,...,xn, можно выразить следующим образом:

$$f(\theta \mid X = x1,...,X = xn)$$

$$"=" \frac{1}{г} \prod_{я=1}^{н} P(X = xi \mid \theta) \cdot f(\theta)$$

$$"=" \frac{1}{г} \cdot \theta r \cdot (1 \quad \theta)^{п-р} \cdot c \cdot \theta \quad 1\,2 \cdot (1 \quad \theta) \quad 1\,2$$

$$"=" \frac{}{г} \cdot \theta r \quad 1\,2 \cdot (1 \quad \theta)^{п-к-1} \quad 2$$

$$"=" \frac{}{г} \cdot \theta(r+1\,2) \quad 1 \cdot (1 \quad \theta)(n \quad r+1\,2) \quad 1 .$$

что представляет собой бета- распределение с параметрами

$$\alpha = p + \frac{1}{2} \quad \text{и } \beta = n \quad r + \frac{1}{2} ,$$

и с константой

$$\frac{c}{г} "=" \frac{(\alpha + \beta)}{(\alpha)(\beta)} ,$$

где гамма-функция (см. гамма распределение).

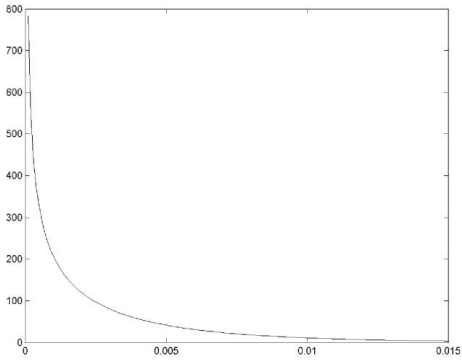
Теперь рассмотрим конкретный случай, когда мы хотим оценить долю ВИЧ-позитивных студентов. Мы тестируем 200 студентов и среди них нет ВИЧ-позитивных. Доля поэтому число ВИЧ-позитивных студентов оценивается как 0 по максимальной вероятности. Доверительные интервалы не очень полезны в этом случае, потому что (если следовать обычному подход) они рассчитываются по:

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

В этом случае, поскольку p – доля наблюдаемых успехов, доверительный интервал уменьшается до 0.

Следуя байесовской методологии, получаем распределение, основанное только на данных, который описывает неопределенность оцениваемых параметров. Чем больше n , тем более уверены, что мы около θ; последняя ссылка распределение для θ тогда более концентрировано вокруг истинного значения θ.

В этом случае в качестве окончательного эталонного распределения находим бета-распределение с параметрами α = 0,5 и β = 200,5, что суммирует информацию о θ (значения, соответствующие всплескам в распределении, наиболее вероятно). Часто бывает полезно графически представить такие результаты:



Мы можем видеть из этого, что:

- Вероятность того, что доля проценты меньше, чем 0,015, это почти 1;

Machine Translated by Google

• Вероятность того, что процентная доля меньше 0,005, составляет приблизительно 0,84, и; • Медиана,
выбранная как наилучшая мера центральной тенденции из-за сильной асимметрии распределения, равна 0,011.

Заметим, что между классическим результатом (пропорцию можно оценить как нулевую) и байесовским решением задачи имеется качественное различие, которое позволяет вычислить вероятностное распределение параметра, математически переведя неопределенность относительно него. Этот метод, в частности, сообщает нам, что при наличии определенной информации правильная оценка процентной доли составляет 0,011. Обратите внимание, что байесовская оценка зависит (как и неопределенность) от количества наблюдаемых случаев и эквивалентна наблюдению 0 случаев ВИЧ среди 2 200 или 20,0 учащихся в классическом случае.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

теорема Байеса

Байес, Томас

Условная возможность

Вывод

Совместная функция плотности

Максимальная вероятность

Вероятность

ССЫЛКИ

Бернардо, Дж. М., Смит, АСМ: байесовская теория. Уайли, Чичестер (1994)

Берри, Д.А.: Статистика с точки зрения Байеса. Уодсворт, Белмонт, Калифорния (1996)

Бокс, ГЭП, Тиао, ГП: Байесовский вывод в статистическом анализе. Аддисон-Уэсли, Рединг, Массачусетс (1973)

Карлин, Б., Луис, Т.А.: Байесовские и эмпирические байесовские методы. Чепмен и Холл /CRC, Лондон (2000)

Финетти, Б. де: Теория вероятностей; Критическая вводная обработка. Транс. Мачи, А., Смит, А. Уили, Нью-Йорк (1975)

Гельман А., Карлин Дж. Б., Стерн Х. С., Рубин Д. Б.: Байесовский анализ данных. Чепмен и Холл / CRC, Лондон (2003)

Линдли, Д.В.: Философия статистики. Статистик 49, 293–337 (2000).

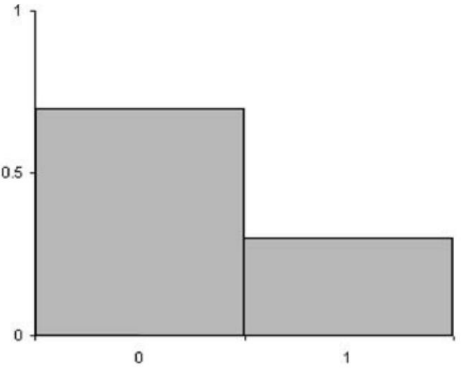
Роберт, СР: Le choix Bayesien. Спрингер, Париж (2006)

Распределение Бернулли

Случайная величина X следует за Бернулли распределение с параметром p , если его функция вероятности имеет вид:

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

где p и q представляют, соответственно, вероятности «успеха» и «неудачи», обозначенные значениями 1 и 0.



Закон Бернулли, p = 0,3, q = 0,7

Распределение Бернулли является дискретным распределение вероятностей.

Machine Translated by Google

ИСТОРИЯ

См. биномиальное распределение.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение распределения Бернулли по определению:

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

$$= 1 \cdot p + 0 \cdot q = p .$$

Дисперсия распределения Бернулли равна по определению:

$$Var(X) = E[X^2] - (E[X])^2 = 1^2 \cdot p + 0^2 \cdot q - p^2 = p - p^2 = pq.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Распределение Бернулли используется, когда

Случайный эксперимент имеет только два возможных результата: «успех» или «неудача». Эти результаты обычно обозначаются цифрами 0 и 1.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Биномиальное распределение
- Дискретное распределение вероятностей

Семья Бернулли

Родом из Базеля, семья Бернулли дала науке несколько математиков. Бернулли Жак (1654–1705) изучал богословие в Базельском университете по завещанию отца, а затем несколько лет путешествовал, преподавая и продолжая собственные исследования. Решительно сосредоточившись на математике, он стал профессором Базельского университета в 1687 г. Согласно Стиглеру (1986),

Бернулли, Жак (называемый Бернулли, Джеймс в большинстве английских работ) является отцом количественного определения неопределенности. Только в 1713 году, через семь лет после его смерти, по наущению его племянника Николя был опубликован его главный труд Ars Conjectandi . Эта работа разделена на четыре части: в первой автор комментирует знаменитый трактат Гюйгенса; вторая посвящена теории перестановок и комбинаций; третья к решению разнообразных задач об азартных играх; и, наконец, в четвертом обсуждается применение теории вероятностей к вопросам морального интереса и экономической науки.

Большое количество работ Бернулли, Жака так и не были опубликованы.
Жан Бернулли (1667–1748), который больше интересовался математикой, чем медициной.
В соответствии с предназначенной ему отцом карьерой, он также стал профессором университета сначала в Гронингене в Нидерландах, а затем в Базеле, где занял кафедру, освободившуюся после смерти его брата.

Два брата работали над дифференциальной интегральное исчисление и задачи минимизации, а также изучал функции.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Стиглер, С.: История статистики, измерение неопределенности до 1900 года. Белкнап, Лондон (1986), стр. 62–71.

Бернулли, Якоб

Бернулли, Якоб (или Жак, или Якоб, или Джеймс) (1655–1705) и его брат Жан были пионерами исчисления Лейбница. Якоб переформулировал проблему вычисления ожидания в исчисление вероятностей. Он также сформулировал слабый закон больших чисел.

Machine Translated by Google

берс, на основании которых современная вероятность и основана статистика. Бернулли, Якоб назначен мэтром литературы в 1671 г. Во время его исследования в области богословия, которые он прекратил в 1676 г. он изучал математику и астрономию вопреки воле своего отца (Николауса Бернулли).

В 1682 году Лейбниц опубликовал (в Acta Eru ditorium) метод, который можно было использовать для определения интегралов алгебраических функций, краткое изложение дифференциального исчисления в алгоритмической схеме и некоторые замечания по поводу основной идеи интеграла исчисление. Эта статья привлекла внимание Якоба и его брата, и они значительно улучшили уже проделанную работу.
по Лейбницу. Сам Лейбниц признавал, что исчисление бесконечно малых было в основном основано братьями Бернулли, а не им самим.
Действительно, в 1690 году Якоб ввел термин "интеграл."
В 1687 году он был назначен профессором Математика в Базельском университете, где он оставался до своей смерти в 1705 году.
Ars conjectandi — это название того, что принято считать самым оригинальным трудом Бернулли. работа. Он состоит из четырех частей: первая содержит переиздание « De Ratiociniis» Гюйгена. в Ludo Aleae (опубликовано в 1657 г.), в котором завершается с помощью важных модификаций.
Во второй части своей работы Бернулли обращается к комбинаторной теории. Третья часть содержит 24 примера, которые помогают проиллюстрировать модифицированная концепция ожидаемой стоимости.
Наконец, четвертая часть самая интересная и оригинален, хотя Бернулли и не есть время закончить. Именно в этой части Бернулли различает два способа определения (точного или приблизительного) классического мера вероятности.
Около 1680 года Бернулли, Якоб также стал интересуюсь стохастикой. Эволюция его за идеями можно следить в его научном журнале Медитации.

Некоторые из основных работ и статей Бернулли, Якоб включают:

1677 Meditationes, Annotationes, Ani madversiones Theologicae & Philo sophicae, a me JB. concinnatae & collectae ab anno 1677. Университеты библиотека, Базель, LI a 3.

1713 Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infini tis, et Epistola Gallice scripta de ludo Пилы прямоугольные. Impensis Thurni siorum, Fratrum, Базель.

Испытание Бернулли

Испытания Бернулли — это повторные испытания эксперимент , который подчиняется следующим правилам:
1. Каждое испытание приводит либо к успеху, либо к провалу. уре;
2. Вероятность успеха одинакова за каждое испытание; вероятность успеха обозначается p, а вероятность отказа q = 1 p;

3. Испытания независимы.

ИСТОРИЯ

Испытания Бернулли получили свое название от швейцарский математик Бернулли Якоб (1713 г.).
Бернулли, Якоб (1654–1705) был старшим из четырех братьев, и это была воля его отца что он должен изучать богословие. Когда у него было закончил богословие в Базеле в 1676 г., он ненадолго покинул город, чтобы вернуться в 1680 году, чтобы посвятить себя математике матика. Он получил кафедру Мате математики в Базельском университете в 1687 году.

Machine Translated by Google

ПРИМЕРЫ

Самый простой пример испытания Бернулли — подбрасывание монеты. Если выпадение «орла» считается успехом (S), а выпадение «решки» — неудачей (F), то имеем:

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

и

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

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Распределение Бернулли

Биномиальное распределение

ССЫЛКИ

Бернулли, Дж.: ArsConjectandi, Opus Posthu мама. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Базель (1713 г.)

Росс, С.М.: Введение в вероятностные модели, 8-е изд. Джон Уайли, Нью-Йорк (2006 г.)

Теорема Бернулли

Теорема Бернулли утверждает, что относительная частота успеха в последовательности испытаний Бернулли приближается к вероятности успеха по мере увеличения числа испытаний в сторону увеличения. бесконечность.

Это упрощенная форма закона больших чисел , вытекающая из неравенства Чебышева.

ИСТОРИЯ

Теорема Бернулли, иногда называемая «слабым законом больших чисел», была впервые описана Бернулли, Якобом (1713 г.) в его работа Ars Conjectandi, которая была опубликована (с помощью его племянника Николауса) через семь лет после его смерти.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если S представляет количество успехов, полученных во время n испытаний Бернулли, и если p - вероятность успеха, то мы имеем:

$$\lim_{n} P\left(\left|\frac{C}{n} - p\right| \geq \varepsilon\right) = 0,$$

или

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

где ε > 0 и сколь угодно мало.

Аналогичным образом мы можем написать:

$$\frac{C}{n} \approx p,$$

это означает, что относительная частота успеха стремится к вероятности успеха, когда n стремится к бесконечности.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Распределение Бернулли

испытание Бернулли

Конвергенция

Закон больших чисел

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бернулли, Дж.: ArsConjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Базель (1713 г.)

Реньи, А.: Теория вероятностей. Северная Голландия, Амстердам (1970)

Machine Translated by Google

Бета-распределение

Случайная величина X следует бета-распределению с параметрами α и β, если его функция плотности имеет вид:

$$e(x) = \frac{1}{B(\alpha, \beta)(x \quad a)\alpha \quad 1(b \quad x)\beta \quad 1} \cdot (b \quad a)(\alpha+\beta \quad 1),$$

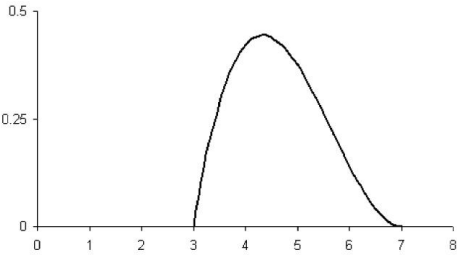
$$a \quad x \quad б \quad , \quad \alpha > 0 \text{ и } \beta > 0 \quad ,$$

где:

$$B(\alpha, \beta) = \int_0^1 t\alpha \quad 1(1 \quad t)\beta \quad 1dt$$

$$"=" \quad \frac{(\alpha)(\beta)}{(\alpha + \beta)},$$

– гамма-функция (см. гамма-распределение).



Бета-распределение α = 2, β = 3, a = 3, b = 7

Бета-распределение представляет собой непрерывное распределение вероятностей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Функция плотности стандартной бета распределение получается путем выполнения изменение переменной $Y = \frac{X-a}{б-a}$:

$$\phi(y) = \frac{1}{B(\alpha,\beta)y\alpha \quad 1} \cdot (1 \quad y)\beta \quad 1 \quad \text{если } 0 < y < 1 \quad .$$

$$0 \quad \text{если не}$$

Рассмотрим X, случайную величину , которая следует стандартный бета-дистрибутив. Ожидаемый _ значение и дисперсия X , соответственно, предоставлено:

$$E[X] = \alpha \frac{\alpha}{+ \beta},$$

$$Var(X) = (\alpha \frac{\alpha \cdot \beta}{+ \beta)2 \cdot (\alpha + \beta + 1)} \cdot$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Бета-распределение является одним из самых часто используется для корректировки эмпирических распределений где диапазон (или интервал изменения) [a, b] известен.

Вот некоторые частные случаи, когда бета-версия используется распределение, связанное с другими непрерывными распределениями вероятностей:

- Если X1 и X2 — два независимых случайных переменные, каждая из которых распределена согласно гамма - распределение с параметрами (α1, 1) и (α2, 1) соответственно случайная величина

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

  распределяется по бета-распределению с параметрами (α1, α2).
- Бета-распределение становится единым распределение , когда

$$\alpha = \beta = 1 \quad .$$

- Когда параметры α и β стремятся к бесконечности, бета-распределение стремится к стандартному нормальному распределению ция.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Непрерывное распределение вероятностей
Гамма-распределение
Нормальное распределение
Равномерное распределение

Machine Translated by Google

Со статистической точки зрения систематическая ошибка определяется как разница между ожидаемым значением статистики и истинным значением соответствующего параметра. Следовательно, смещение является мерой систематической ошибки оценки . Если мы посчитаем среднее значение большое количество непредвзятых оценок, мы найдем правильное значение. Смещение указывает расстояние оценщика от истинного значение параметра.

ИСТОРИЯ

Концепция несмещенной оценки исходит от Gauss, CF (1821), в то время, когда он работал над методом наименьших квадратов .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Мы не должны путать предвзятость оценки тора параметра с его степенью точности, что является измерением выборки ошибка.

Существует несколько типов систематической ошибки: систематическая ошибка отбора (из-за систематических различий между сравниваемыми группами), систематическая ошибка исключения (из-за систематического исключения определенных лиц из исследования) или аналитическая систематическая ошибка (из-за способа оценки результатов).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим статистику T , используемую для оценки параметр θ . Если E[T] = θ + b(θ) (где E[T] представляет ожидаемое значение T), то величина b(θ) называется смещением статистики T.

Если b(θ) = 0, мы имеем E[T] = θ, и T — несмещенная оценка θ.

ПРИМЕРЫ

Рассмотрим X1, X2,...,Xn — последовательность независимых случайных величин , распределенных по одному и тому же закону вероятности со средним значением μ и конечной дисперсией σ2. Мы можно вычислить смещение оценщика

$$S2 = \frac{1}{н} \cdot \sum_{я=1}^{н} (xi \quad \bar{x})^2 ,$$

используется для оценки дисперсии σ2 совокупности следующим образом :

$$E[S2] = E \quad \frac{1}{н} \sum_{я=1}^{н} (xi \quad \bar{x})^2$$

$$= E \quad \frac{1}{н} \sum_{я=1}^{н} (xi \quad \mu)2 \quad (x^{-} \quad \mu)2$$

$$"=" \quad \frac{1}{н} E \sum_{я=1}^{н} (xi \quad \mu)2 \quad E(x^{-} \quad \mu)2$$

$$"=" \quad \frac{1}{н} \sum_{я=1}^{н} E(xi \quad \mu)2 \quad E(x^{-} \quad \mu)2$$

$$= Var(xi) \quad Var(x^{-}) \quad \sigma2$$

$$\sigma2 \quad = \frac{}{н} .$$

Тогда смещение S2 равно $\frac{\sigma2}{н}$.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Оценка

Оценщик

Ожидаемое значение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Gauss, CF: Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, части 1, 2 и доп. Верке 4, 1–108 (1821, 1823, 1826)

Machine Translated by Google

Gauss, CF: Méthode des Moandres Car
рез. Воспоминания о Combinaison des
Наблюдения. Французский перевод, пар. Ж.
Бертран. Малле-Башелье, Париж (1855 г.)

Бьенеме, Ирене-Жюль

Бьенеме, Ирене-Жюль (1796–1878)
поступил в министерство финансов Франции и
стал генеральным инспектором по финансам
в 1836 г., хотя он потерял работу
в 1848 г. по политическим причинам. Вскоре после
этого он стал давать уроки на факультете
науки в Париже.

Последователь Лапласа, Пьер Симон
де, Бьенеме доказал Бьенеме–
Неравенство Чебышева несколько лет назад
Чебышев, Пафнутий Львович. Отредактированная
версия статьи Бьенеме 1867 г.
предшествует французской версии чебычевского
доказательство. Затем это неравенство использовалось в
Неполное доказательство центральности Чебышева
предельная теорема, которую позже доработал
Марков, Андрей Андреевич.
Более того, Бьенеме правильно сформулировал
Теорема для ветвящихся процессов в 1845 году.
Его самая известная общественная работа, вероятно,
исправления, которые он внес в использование Дювиллара
таблица смертности.

Некоторые основные работы и статьи Bien aymé,
Irénée-Jules:

1853 г. Соображения по поводу аппуи-де-ла-декоу-
верте-де-Лаплас-сюрлалои-де-вероятность
в методе автомобиля
рез. Comptes Rendus de l'Académie
des Sciences, Париж, 37, 5–13; отредактировано
в 1867 г. в Journal de Liouville
предшествующий доказательству Бьенеме –
Неравенство Чебышева в J. Math.
Чистый. Приложение, 12, 158–176.

Двоичные данные

Двоичные данные возникают, когда интересующая
переменная может принимать только два значения.
Эти два значения обычно представлены 0 и 1, даже
если переменная не является количественной.
Пол, наличие или отсутствие характеристики, успех
или неудача эксперимента — вот лишь несколько
примеров переменных .
которые приводят к двоичным данным. Эти переменные
называются дихотомическими переменными.

ПРИМЕРЫ

Метеоролог хочет знать, насколько надежен
его прогнозы. Для этого он изучает случайную
переменную , представляющую предсказание.
Эта переменная может принимать только два значения:

$$X = \begin{cases} 0, \text{ если предсказание было неверным} \\ 1, \text{ если предсказание было правильным} \end{cases}.$$

Метеоролог делает прогнозы на
срок 50 дней подряд. Оказалось, что предсказание
оказывается правильным 32 раза, а неверным — 18
раз.

Чтобы узнать, лучше ли его предсказания
чем те, которые можно было получить
подбрасывая монетку и предсказывая погоду
в зависимости от того, выпадет ли орел или решка,
он решает проверить нулевую гипотезу H0,
что доля p правильных предсказаний равна
равна 0,5, вопреки альтернативной гипотезе H1,
что эта пропорция отлична от 0,5:

$$H0 : p = 0,5$$

$$H1 : p = 0,5 .$$

Рассчитаем значение _

$$x2 = \frac{\text{наблюдаемая частота} \quad \text{теоретическая частота2}}{\text{теоретическая частота}},$$

Machine Translated by Google

где наблюдаемые частоты равны 18 и 32, а теоретические частоты равны, соответственно 25 и 25:

$$x2 = \frac{(32 \quad 25)2 + (18 \quad 25)2}{25}$$

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

Предполагая, что применима центральная предельная теорема , мы сравниваем это значение с значение распределения хи-квадрат с одной степенью свободы. Для уровня значимости 5% мы находим в хи-квадрате стол:

$$x2_{1,0,95} = 3{,}84 .$$

Следовательно, поскольку 3,92 > 3,84, нуль гипотеза отвергается, а это означает, что метеоролог предсказывает погоду лучше, чем монета.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Распределение Бернулли

Категориальные данные

Таблица сопряженности

Данные

Дихотомическая переменная

Тест отношения правдоподобия

Логистическая регрессия

Переменная

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кокс, Д.Р., Снелл, Э.Дж.: Анализ двоичных данных. Чепмен и Холл (1989)

Бишоп, Ю.М.М., Финберг, С.Е., Холланд, PW: Дискретный многомерный анализ: теория и практика. Массачусетский технологический институт Пресс, Кембридж, Массачусетс (1975)

Биномиальный

Алгебраические суммы, содержащие переменные , называются многочленами (от греческого «поли», означает «несколько»). Выражение, содержащее два члена, называется биномом (от латинское «би», что означает «двойной»). моном (от греческого «моно», что означает «уникальный») выражение с одним термином и трехчленом (от латинского «tri», что означает «тройной») содержит три элемента.

ИСТОРИЯ

См. арифметический треугольник.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Квадрат двучлена легко вычисляется

с использованием

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

Биномиальные формулы для показателей степени выше чем два также существуют, например:

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

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

Мы можем написать обобщенную биномиальную формулу следующим образом:

$$(a + b)^n = an + nan \quad 1b + \frac{n(n-1)}{2} an \quad 2b2 + \cdots + 2 \quad \frac{n(n-1)}{} a2bn \quad 2 + nabn \quad 1 + bn$$

$$= C0_n an + C1 \quad net^n \quad 1 b + C2 nan \quad 2b2 + \cdots + Cn \quad 1 abn \quad 1 + Cn nbn$$

$$"=" \sum_{k=0}^{n} Ck nan \quad k bk .$$

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Бином (p + q), возведенный в степень n , дает название биномиальному распределению потому что она соответствует полной вероятности , полученной после n испытаний Бернулли.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Арифметический треугольник
- Биномиальное распределение
- Комбинация

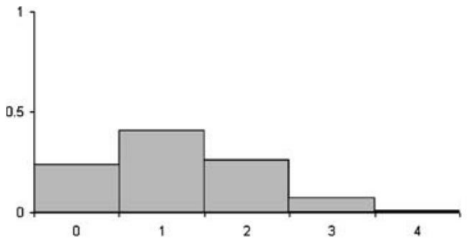
Биномиальное распределение

Случайная величина X следует биномиальному распределение с параметрами n и p , если его функция вероятности имеет вид:

$$P(X = x) = C_x^n \cdot p^x \cdot q^{n-x},$$

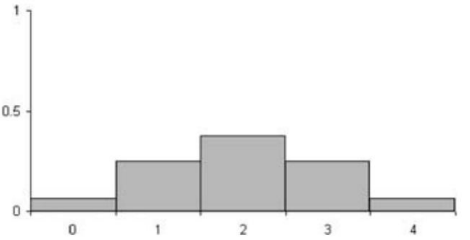
$$x = 0, 1, 2,\ldots,n.$$

Следовательно, если событие включает в себя x «успехов» и (n x) «неуспехов», где p — вероятность «успеха», а q = 1 — p — вероятность «неудачи», биномиальное распределение позволяет вычислить вероятность получить x успехов из n независимых испытаний.



Биномиальное распределение, p = 0,3, q = 0,7, n = 3

Биномиальное распределение с параметрами n и p, обозначаемое B(n, p), представляет собой дискретное распределение вероятностей.



Биномиальное распределение, p = 0,5, q = 0,5, n = 4

ИСТОРИЯ

Биномиальное распределение является одним из старых наиболее известные распределения вероятностей. Он был открыт Дж. Бернулли в его работе под названием Ars Conjectandi (1713 г.). Эта работа разделена на четыре части: в первой автор комментирует трактат Гюйгенса; вторая часть посвящена теории пермутаций и комбинаций; третья посвящена решению различных задач, связанных с азартными играми; наконец, в четвертой части он предлагает применить теорию вероятностей к моральным вопросам и к экономической науке.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если X1, X2,...,Xn — n независимых случайных величин, подчиняющихся распределению Бернулли с параметром p, то случайная величина

$$X = X1 + X2 + \ldots + Xn$$

следует биномиальному распределению B(n, p). Для вычисления ожидаемого значения X будет использоваться следующее свойство, где Y и Z — две случайные величины:

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

Таким образом, мы имеем:

$$E[X] = E[X1 + X2 + \ldots + Xn]$$

$$= E[X1] + E[X2] + \ldots + E[Xn]$$

знак равно n + n + ... + n знак равно np .

Machine Translated by Google

Чтобы вычислить дисперсию X, следующее будет использоваться свойство, где Y и Z — два независимые переменные:

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

Таким образом, мы имеем:

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

Таблицы биномиального распределения (для распределения вероятностей и распределения функция) были опубликованы Рао и другие. (1985), а также Национальным бюро стандартов (1950). Расширенный таблицы функций распределения можно найти в Анналы вычислительной лаборатории (1955).

ПРИМЕРЫ

Монета подбрасывается десять раз. Рассмотрим случайную переменную X, которая представляет количество раз, когда монета выпадает «решкой».

Таким образом, мы имеем:

количество испытаний:	$n = 10$
вероятность _ один успех:	$p = \frac{1}{2}$ (хвосты получены)
вероятность одного отказ:	$q = \frac{1}{2}$ (головы получены)

Вероятность выпадения решки x раз среди десяти испытаний дается

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

Вероятность выпадения решки ровно восемь раз равно:

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

Случайная величина X следует за бино миальное распределение $B(10, \frac{1}{2})$.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Распределение Бернулли

Биномиальная таблица

Дискретное распределение вероятностей

Отрицательное биномиальное распределение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бернулли, Дж.: ArsConjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Базель (1713 г.)

Гарвардский университет: таблицы кумулятивного биномиального распределения вероятностей, об. 35. Летопись вычислительной лаборатории, издательство Гарвардского университета, Кэмбридж, Массачусетс (1955).

Национальное бюро стандартов.: Таблицы Биномиальное распределение вероятностей. Министерство торговли США. Применяемый Математика, серия 6 (1950)

Рао, Ч.Р., Митра, С.К., Маттай, А., Рама Мурти, К.Г.: Формулы и таблицы для Статистическая работа. Статистическая публикация Компания, Калькутта, стр. 34–37 (1966).

Биномиальная таблица

Биномиальная таблица дает значения для функция распределения случайного _

Machine Translated by Google

переменная , следующая за биномиальным распределением
бутион.

ИСТОРИЯ

См. биномиальное распределение.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть случайная величина X следует за бинарным
миальное распределение с параметрами n и p.
Его функция вероятности определяется выражением:

$$P(X = x) = C_{x\ \ н} \cdot p^{x} \cdot q^{n\ \ x},$$
$$x = 0, 1, 2, \ldots, n,$$

где $C_{x\ \ н}$ - биномиальный коэффициент, равный
к $\frac{}{н!\ x!(n\ \ x)!}$, параметр p – вероятность
успеха, а q = 1 p является дополнительным
вероятность, которая соответствует вероятности
отказа (см. нормальное распределение).
Функция распределения случайных
переменная X определяется:

$$P(X \quad x) = \sum_{я=0}^{Икс} C_{н}^{и} \cdot п^{и} \cdot q^{n\ \ i},$$
$$0 \quad x \quad п \quad .$$

Биномиальная таблица дает значение
P(X x) для различных комбинаций x,
н и p.
При увеличении этот расчет становится утомительным.
К счастью, мы можем использовать некоторые очень хорошие
вместо этого приближения. Если min(np, n(1
p)) > 10, мы можем аппроксимировать его
нормальное распределение:

$$P(X \quad x) = \varphi\left(\frac{x + 2\ \frac{1}{}\ нп}{\sqrt{npq}}\right),$$

где φ — функция распределения
стандартное нормальное распределение и поправка
на непрерывность $\frac{1}{2}$ Включено.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Биномиальная таблица используется для выполнения
непараметрических тестов статистических данных ,
распределенных в соответствии с биномиальным
распределением, особенно теста на знак и теста на знак.
биномиальный тест.
Национальное бюро стандартов (1950 г.)
опубликованы индивидуальные и кумулятивные
вероятности биномиального распределения для n 49,
в то время как кумулятивное биномиальное распределение
вероятности для n 1000 приведены в
Анналы вычислительной лаборатории
(1955).

ПРИМЕРЫ

См. Приложение Б.
Мы можем проверить, что для n = 2 и p = 0,5:

$$P(X \quad 1) = \sum_{я=0}^{1} Ки\ 2(0{,}5)^{я}(0{,}5)^{2\ \ i}$$
$$= 0{,}75 .$$

или для n = 5 и p = 0,05:

$$P(X \quad 3) = \sum_{я=0}^{3} Ки\ 5(0{,}05)^{я}(0{,}95)^{5\text{-я}}$$
$$= 1.0000 .$$

Пример применения биномиальной таблицы см.
в биномиальном тесте.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Биномиальное распределение
- Биномиальный тест
- Подписать тест
- Статистическая таблица

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Национальное бюро стандартов.: Таблицы
Биномиальное распределение вероятностей.
Министерство торговли США. Применяемый
Математика, серия 6 (1950)

Machine Translated by Google

Гарвардский университет: Таблицы кумулятивного биномиального распределения вероятностей, том. 35. Летопись вычислительной лаборатории, издательство Гарвардского университета, Кэмбридж, Массачусетс (1955).

Биномиальный тест

Биномиальный тест — это параметрический тест гипотезы , который применяется, когда совокупность может быть разделена на два класса: каждый наблюдатель часть этой популяции будет принадлежать к той или иной из этих двух категорий.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ Мы рассматриваем выборку из n независимых испытаний. Каждое испытание принадлежит либо к классу C1 , либо к классу C2. Заметим количество наблюдений n1 , попадающих в C1 , и количество наблюдений n2 , попадающих в C2.

Каждое испытание имеет вероятность p принадлежности к классу C1, где p одинаково для всех n испытаний, и вероятность q = 1 p принадлежности к классу C2.

Гипотезы

Биномиальный критерий может быть двусторонним . тест или односторонний тест. Если p0 является предполагаемым значением p (0 p0 1), гипотезы выражаются следующим образом:

A: Двусторонний чехол

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

B: односторонний чехол

$$H0 : p \quad p0 ,$$

$$H1 : p > p0 .$$

C: Односторонний случай

$$H0 : p \quad p0 ,$$

$$H1 : p < p0 .$$

Правила принятия решений

Случай A

Для n < 25 мы используем биномиальную таблицу с n и po в качестве параметров.

Поэтому сначала ищем ближайшее значение к $\frac{\alpha}{2}$ inthistable (которую мы обозначаем α1), где α — уровень значимости. Обозначим значение, соответствующее α1, через t1.

Далее находим в таблице значение 1 α1 = α2 . Обозначим значение, соответствующее α2, через t2.

Мы отвергаем H0 на уровне α, если

$$n1 \quad t1 \text{ или } n1 > t2 ,$$

где n1 — количество наблюдений, попадающих в класс C1.

Когда n больше 25, мы можем использовать нормальное распределение в качестве аппроксимации биномиального распределения.

Параметры биномиального распределения являются:

$$\mu = n \cdot p0 ,$$

$$\sigma \text{ знак равно } \sqrt{n \cdot p0 \cdot q0} .$$

Случайная величина Z , следующая за стандартное нормальное распределение тогда равно:

$$Z = \frac{X - \text{мю}}{\text{о}} \text{ "=" } \frac{X \quad n \cdot p0}{\sqrt{n \cdot p0 \cdot q0}} ,$$

где X — случайная биномиальная величина с параметры n и p0.

Приближения для значений t1 и t2 тогда:

$$t1 = n \cdot p0 + z\alpha1 \cdot \sqrt{n \cdot p0 \cdot q0} , t2 = n \cdot p0 + z\alpha2 \cdot \sqrt{n \cdot p0 \cdot q0} ,$$

где zα1 и zα2 – значения, найденные в таблице нормалей , соответствующие уровням α1 и α2.

Machine Translated by Google

Случай Б

Для n < 25 мы принимаем за t значение в биномиальной таблице, соответствующее 1 α, где α — уровень значимости (или ближайшее значение), а n и p0 — параметры, описанные ранее. Мы отвергаем H0 на уровне α, если

$$n1 > t \quad ,$$

где n1 — количество наблюдений, попадающих в класс C1.

Для n 25 мы можем сделать приближение:

$$t = n \cdot p0 + z\alpha \quad \sqrt{n \cdot p0 \cdot q0} \ ,$$

где zα можно найти в таблице нормалей для 1 α.

Правило решения в этом случае такое же, как что для n < 25.

Случай С

При n < 25 это значение в биномиальной таблице, соответствующее α, где α — уровень значимости (или ближайшее значение), а при n и p0 — те же параметры, что описаны ранее.

Мы отвергаем H0 на уровне α, если

$$n1 \quad t \quad ,$$

где n1 — количество наблюдений, попадающих в класс C1.

Для n 25 делаем приближение:

$$t = n \cdot p0 + z\alpha \quad \sqrt{n \cdot p0 \cdot q0} \ ,$$

где zα находится в нормальной таблице для уровня значимости α. Тогда правило принятия решения такое же, как описано ранее.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Два основных условия, которые необходимо соблюдать при выполнении биномиального теста:

1. n наблюдений должны быть взаимно независимый;
2. Каждое наблюдение имеет вероятность p попасть в первый класс. Эта вероятность также одинакова для всех наблюдений.

ПРИМЕРЫ

Станок считается исправным, если не более 5% произведенных на нем деталей имеют дефекты. Нулевая гипотеза, обозначенная H0, описывает эту ситуацию, тогда как альтернативная гипотеза, обозначенная H1, означает, что машина не работает:

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

Выполняя тест, берем выборку из 10 штук и отмечаем, что бракованных штук три (n1 = 3). Поскольку гипотезы соответствуют одностороннему критерию (случай B), используется решающее правило B. Если мы выбираем уровень значимости α = 0,05, значение t в биномиальной таблице равно 1 (при n = 10 и p0 = 0,05) .

Мы отвергаем H0 , поскольку n1 = 3 > t = 1, и делаем вывод, что машина не работает.
Затем снова проводим тест, но уже на 100 штук. Мы замечаем, что есть 12 дефектных частей. В этом случае значение часто может быть аппроксимировано:

$$t = n \cdot p0 + z\alpha \quad \sqrt{n \cdot p0 \cdot q0} \ ,$$

где zα можно найти из таблицы нормалей при 1 α = 0,95. Тогда мы имеем:

$$t = 100 \cdot 0{,}05 + 1{,}64 \quad \sqrt{100 \cdot 0{,}05 \cdot 0{,}95} = 8{,}57 \ .$$

Мы снова отвергаем H0, так как n1 = 12 > t = 8,57.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Биномиальная таблица

Тест на пригодность

Проверка гипотезы

Односторонний тест

Параметрический тест

Двусторонний тест

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Абди, Х.: Биномиальное распределение: биномиальные и знаковые тесты. В: Салкинд, Нью-Джерси (ред.) Энциклопедия измерений и статистики. Мудрец, Тысяча дубов (2007)

Биостатистика

Биостатистика – это научная область, в которой применяются статистические методы, ответить на вопросы, связанные с биологией человека и медицины (приставка «био» происходит от греческое «биос», что означает «жизнь»). Области биостатистики в основном эпидемиологии, клинических и биологических исследований, а также используется при изучении этика этих испытаний.

ИСТОРИЯ

Биостатистика началась в середине семнадцатого века, когда Петти, сэр Уильям (1623–1687) и Граунт, Джон (1620–1674) создал новые методы анализа лондонских ведомостей о смертности. Они применили эти новые методы анализа смертности, рождаемости и переписи населения , создавая области биометрии. Затем, в середине девятнадцатого века, работы Менделя, Грегор изучал наследственность у растений. Его наблюдения и результаты основывались на систематически собираемых данных, а также на применение численных методов описания закономерностей наследственной передачи. Гальтон, Фрэнсис и Пирсон, Карл были два самых важных человека, связанных с развитием этой науки. Они использовали новые концепции и статистические методы при исследовании сходства в физическом, психологическом и поведенческом данных между родителями и их детьми. Пирсон, Гальтон и Уэлдон, Уолтер Франк Рафаэль (1860–1906) стал одним из основателей журнала «Биометрика». Фишер, Рональд Эйлмер, во время сельскохозяйственных исследований, проведенных в Предлагаемая экспериментальная станция в Ротамстеде метод случайной выборки , при котором животные были разделены на разные группы и выделены различные обработки, маркировка первые исследования в клинические испытания.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Области биостатистики в основном эпидемиологии, клинических испытаний и этических связанные с ними вопросы, а также биологические испытания. Одна из областей, где биостатистики

Для анализа конкретного вопроса в косвенных измерениях использовались ка льские концепции. стойкости вещества (например, витаминов и гормонов) при введении его живому существу. Таким образом, биостатистика в основном относится к статистике, используемой для решения проблемы, возникающие в биомедицинских науках энц.

Наиболее часто используемые статистические методы в биостатистике включают методы повторной выборки (бутстрап, складной нож), многовариантный анализ, различные методы регрессии, оценки неопределенности, связанные с оценкой, и обработка отсутствующих данных.

Самые известные научные журналы к биостатистике относятся: Статистика в Медицина, Биометрический журнал, Контролируемые клинические испытания, Журнал биофармацевтической статистики, Статистические методы в медицинских исследованиях и биометрия.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Демография
- Эпидемиология

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Армитидж П., Колтон Т.: Энциклопедия биостатистики. Уайли, Нью-Йорк (1998)

Граунт, Дж.: Естественные и политические наблюдения, упомянутые в следующем указателе и сделанные по спискам смертности: в отношении правительства, религии, торговли, роста, воздуха, болезней и некоторых изменений в указанном городе. То. Ройкрофт для Джона Мартина, Джеймса Аллестри и Тхо. Дикас (1662 г.)

Гринвуд, М.: Медицинская статистика от Граунта до Фарра. Издательство Кембриджского университета, Кембридж (1948)

Петти, В.: Еще одно эссе по политической арифметике, касающееся роста Лондона. Келли, Нью-Йорк (1682 г.) (2-е изд. 1963 г.)

Петти, В.: Наблюдения за дублинскими смертными приговорами 1681 г. и состоянием этого города. Келли, Нью-Йорк (1683 г.) (2-е изд. 1963 г.)

Блокировать

Блоки — это наборы, в которых экспериментальные единицы сгруппированы таким образом, чтобы единицы были максимально похожи внутри каждого блока. Мы можем ожидать, что экспериментальная ошибка , связанная с блоком, будет меньше, чем получается, если одинаковое количество единиц располагались случайным образом в пределах всего экспериментального пространства.

Блоки, как правило, определяются с учетом как управляемых причин, связанных с изучаемыми факторами, так и причин, может быть трудно или невозможно поддерживать постоянное значение для всех экспериментальных единиц. Затем вариации между блоками устраняется, когда мы сравниваем эффекты факторов .

Можно использовать несколько типов перегруппировок, чтобы уменьшить влияние одного или нескольких источников ошибка. Рандомизированный дизайн блока получается, если есть только один источник ошибки.

ИСТОРИЯ

Блочная концепция, используемая в области планирования эксперимента, возникла в исследованиях, проведенных Фишером Р.А. (1925 г.), когда он был руководителем экспериментальной станции в Ротамстеде . в том смысле, что плодородие систематически изменяется от одной точки земли к другой.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- План экспериментов
- Экспериментальная установка
- Фактор
- Греко-латинская квадратная конструкция
- Латинские квадратные узоры
- Рандомизированный блочный дизайн

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Статистические методы для научных работников. Оливер и Бойд, Эдин Бург (1925)

Бонферрони, Карло Э.

Бонферрони Карло Эмилио родился в 1892 году в Бергамо, Италия. Он получил степень по математике в Турине и завершил свое образование, проведя год в университете в

Machine Translated by Google

Вена и Eidgenössiche Technische HochschuleinZurich.HewasaВоенный офицер во время Первой мировой войны и после после окончания войны стал ассистентом профессора в Туринском политехническом институте. В 1923 году он получил кафедрой финансовой математики в Эко номики в Бари, где он был Ректор в течение семи лет. Он, наконец, перевелся во Флоренцию в 1933 году, где он занимал свое кресло до его смерти.

Бонферрони занимался различными вопросами, включая актуарную математику, вероятность и статистическая математика, анализ, геометрия и механика. Он дал свое имя два неравенства Бонферрони, облегчающие обработка статистических зависимостей. Эти впервые появился в 1936 г. в статье Teoria statistica delle classi e calcolo.

делле вероятностный. Развитие этих неравенства породили множество новых литературы в этой области.

Бонферрони умер в 1960 году во Флоренции, Италия. Основная статья Карло Эмилио Бонферрони:

1936 Теория статистики делле класси и расчет делле вероятностный. Опубл. Р. Истит. Супер. науч. Экон. ком. Флоренция, 8, 1–62.

Начальная загрузка

Термин бутстрап описывает семейство методы, которые в основном используются для оценки стандартной ошибки, систематической ошибки и доверительного интервала параметра (или бо чем один параметр). Он основан на n независимых наблюдениях за случайной величиной. с неизвестной функцией распределения F, и особенно полезен, когда оцениваемые параметры относятся к сложной функция Ф.

Основная идея начальной загрузки (которая чем-то похожа на идею складного ножа метод) заключается в оценке F с использованием возможного распределения Fˆ, а затем провести повторную выборку из Fˆ. Процедуры начальной загрузки обычно требуют использования компьютеров, так как они могут выполнять большую количество симуляций за относительно короткий время. В методах начальной загрузки автоматическое моделирование заменяет аналитические расчеты, используемые в «традиционных» методах расчета. оценки, а в некоторых случаях они могут предоставить больше свободы, например, когда мы делаем не хотим (или не можем) принять гипотезу о структуре распределения данные. Некоторые методы начальной загрузки включены в статистическом программном обеспечении, таком как S-PLUS, SAS и МАТЛАБ.

ИСТОРИЯ

Происхождение слова «бутстрап» в отношении методов, описанных здесь ярко иллюстрирует рефлексивную природу генерируемые ими вторичные образцы (т. построенные из Fˆ). Это происходит из литературный персонаж барон Мюнхгаузен (из «Удивительных приключений барона Мюнхгаузена Распе, РЭ), попавшего в ладно, но вытащил себя за собственные ремешки ботинок (свои шнурки).

История бутстрапа, как и многих методов самостатистики, начинается в 1979 г. публикация статьи Эфрона (1979), который оказал большое влияние на многих исследователи. Эта статья вызвала публикацию огромного количества статей о Тео рия и приложения бутстрапа.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Были предложены различные типы начальной загрузки; цель здесь не в том, чтобы просмотреть все это исчерпывающе, но вместо того, чтобы дать представление видов применяемых методов.

Machine Translated by Google

Пусть x1, x2,...,xn — n независимых наблюдений случайной величины с неизвестной функция распределения Ф.

Нас интересует оценка неизвестный параметр δ, который зависит от F и надежности (другими словами, его смещения, его дисперсии и его доверительного интервала). Оценка δ дается статистикой t = t (F), которая зависит от F. Поскольку F неизвестно, находим оценку Fˆ для F по выборке x1, x2,...,xn, и находим оценить δ, используя t Fˆ . Классическими примерами t являются математическое ожидание, дисперсия и квантили F. Обозначим случайную величину, соответствующую статистике t (которая зависит от x1, x2,...,xn), через T.

Мы различаем следующие типы бутстрапа: •

Параметрический бутстрап. Если F = Fθ является членом параметрического семейства функций распределения, мы можем оценить (вектор) θ, используя оценку максимального правдоподобия θˆ. Мы естественно оцениваем F через Fθˆ. • Непараметрический бутстрап. Если F не входит в параметрическое семейство, оно оценивается с помощью эмпирической функции распределения, рассчитанной на основе выборки размера. Эмпирическая функция распределения по x определяется следующим образом:

$$\hat{F}(x) = \frac{\text{количество наблюдений } x_i \quad x}{n}.$$

Идея бутстрапа заключается в следующем.

1. Рассмотрим R независимых выборок типа x $x^*_1, x^*_2, \ldots, x^*_n$ случайной величины распределения Fˆ.
2. Вычислить оценочные значения $t^*_1, t^*_2, \ldots, t^*_p$ для параметра δ на основе gen проверенные образцы начальной загрузки. Сначала вычислите оценку Fˆ на основе выборки i (i = 1,...,R), а затем установить:

$$t_i = tF\hat{}_i .$$

Следуя методу начальной загрузки, смещение b и дисперсия v t оцениваются в паре:

$$\text{bboot} = \bar{t}^* \quad \tau = \frac{1}{p}\sum_{i=1}^{p} \tau_i \quad \tau,$$

$$\text{загрузка} = \frac{1}{p-1}\sum_{i=1}^{p} (\tau_i^* \quad \bar{t})^2 .$$

Следует отметить, что здесь вмешиваются два типа ошибок:

- Статистическая ошибка, возникающая из-за того, что мы исходим из смещения и дисперсии для Fˆ, а не для реального F, который неизвестен.
- Ошибка симуляции, возникающая из-за того, что количество симуляций не достаточно высок. На самом деле достаточно провести повторную выборку от R = 50 до 200 раз.

Существует множество возможных методов получения доверительных интервалов с помощью бутстрапа, и здесь мы приводим лишь некоторые из них: 1. Нормальные интервалы. Предположим, что T следует гауссовскому (нормальному) распределению. Надеясь, что T также следует нормальному распределению, мы затем находим следующий доверительный интервал для скорректированного (смещения) значения tb = t b для t:

$$tb \quad \sqrt{\text{vboot}} \cdot z1 \quad \alpha/2,$$

$$tb + \sqrt{\text{vboot}} \cdot z1 \quad \alpha/2 ,$$

где zγ – квантиль γ гауссовского распределения N(0, 1). В качестве предела допуска α используем значение из таблицы нормалей z1 α $\frac{}{2}$, что дает α = 5%: $z_{1\ 0{,}05\frac{}{2}}$ = 1,96.

Machine Translated by Google

2. Начальные интервалы. Предположим, что T δ не зависит от неизвестной переменной. В этом случае мы вычисляем α-квантиль aα T δ . Для этого упорядочим значения τ_1^* t, t_2 τ, . . . , t R t в порядке возрастания:

$$\tau^*_{(1)} \quad \tau, \tau^* \quad \tau, \ldots, \tau^*_{(P)} \tau_{(2)} ,$$

а затем читаем квантили inter t и a1 α есть aα = t $\tau^*_{((R+1)\alpha)}$ = $\tau_{((R+1)(1\ \alpha))}$ τ. И насколько

$$1 \quad 2\alpha = \Pr(a\alpha \quad T \quad \delta \quad a1 \quad \alpha) = \Pr\ 2t - t^*_{((R+1)(1\ \alpha))} \quad \delta \quad 2\tau - \tau^*_{((R+1)\alpha)} ,$$

доверительный интервал для t на уровне значимости α равен:

$$2\tau - \tau^*_{((R+1)(1\ \alpha)),} \quad 2\tau - \tau^*_{((R+1)\alpha)} .$$

3. Стьюдентизированные интервалы начальной загрузки. Можно улучшить предыдущую оценку С учетом

$$Z = \frac{T - \delta}{\bar{v}} ,$$

где v — дисперсия часто, которая должна быть оценивается (обычно с использованием известных методов, таких как дельта-методы или складной нож). метод). Доверительный интервал для t находится аналогично этому описано ранее:

$$\tau - \bar{v} \cdot \tau^*_{((R+1)(1\ \alpha)),} \quad \tau + \bar{v} \cdot \tau^*_{((R+1)\alpha)} ,$$

где z $_{((R+1)(1\ \alpha))}$ и z $_{((R+1)\alpha)}$ являются “эмпирически” (1 α и α) квантили z $_{я}$ "=" $\frac{\tau^*_я - t}{v}$ i=1,...,R.

ПРИМЕРЫ

Рассмотрим типичный пример непараметрического бутстрапа (взято из работы Дэвисон, А. и Хинкли, Д.В. (1997)). данные относятся к населению десяти американских городов в 1920 (U) и 1930 (X):

и	138	93	61	179	48					
х	143	104	69	260	75					

и	37	29	23	30	2					
х	63	50	48	111	50					

Нас интересует значение δ статистики T = что позволит нам вывести населения в 1930 г. по сравнению с населением в 1920 г. Мы оценить δ, используя $t = \frac{\text{Икс}}{\text{ты}}$ = 1,52, но что неопределенность в этой оценке? Начальная загрузка позволяет нам уэс $\tau^*_я$ смоделировать значение (i = 1,...,R) δ, выбранное из двумерное распределение Y = (U, X) с эмпирическая функция распределения F^, которая дает вес $\frac{1}{10}$ к каждому наблюдению. На практике это вопрос повторной выборки с вернуть десять значений данных и вычислить соответствующий

$$\tau^*_я \text{ "=" } \frac{\sum_{j=1}^{10} x^*_{ij}}{\sum_{j=1}^{10} и^*_{ij}} , \quad я = 1,...,R,$$

где x^*_{ij} и u^*_{ij} – значения переменной Способны X и U. Это позволяет приблизить смещение и дисперсию t с использованием формул, приведенных ранее.

В следующей таблице приведены результаты моделирования, полученные при R = 5:

	Дж					
	1	2	3	4	5	6	7
$и^*_{1j}$	93	93	93	61	61	37	29
x^*_{1j}	104	104	104	69	69	63	50
$и^*_{2j}$	138	61	48	48	37	37	29
x^*_{2j}	143	69	75	75	63	63	50

Machine Translated by Google

	Дж 1234567
$и^*_{3j}$	138 93 179 37 30 30 30
$х^*_{3j}$	143 104 260 63 111 111 111
$и^*_{4j}$	93 61 61 48 37 29 29
$х^*_{4j}$	104 69 69 75 63 50 50
$и^*_{5j}$	138 138 138 179 48 48 48
x5j	143 143 143 260 75 75 75

	Дж			Сумма
	8	9	10	
$и^*_{1j}$	29	30	2	$и^*_{1j}$ = 528
$х^*_{1j}$	50	111	50	$х^*_{1j}$ = 774
$и^*_{2j}$	30	30	2	$и^*_{2j}$ = 460
$х^*_{2j}$	111	111	50	$х^*_{2j}$ = 810
$и^*_{3j}$	30	2	2	$и^*_{3j}$ = 571
$х^*_{3j}$	111	50	50	$х^*_{3j}$ = 1114
$и^*_{4j}$	23	23	2	$и^*_{4j}$ = 406
$х^*_{4j}$	48	48	50	$х^*_{4j}$ = 626
$и^*_{5j}$	29	23	30	$и^*_{5j}$ = 819
x5j *	50	48	111	$х^*_{5j}$ = 1123

Итак, $т^*_я$ являются:

$т^*_1 = 1{,}466$, $т^* = 1{,}761\ 2$, $т^* = 1{,}951\ 3$,
$т^*_4 = 1{,}542$, $т^*_5 = 1{,}371$.

Отсюда легко вычислить смещение и дисперсия t с использованием приведенных формул ранее:

б = 1,62 - 1,52 = 0,10 и v = 0,0553 .

Эти результаты можно использовать для расчета нормальных доверительных интервалов, предполагая, что T δ имеет нормальное распределение N (0,1, 0,0553). Получаем для значимости уровне 5%, доверительный интервал для t следующий: [1,16, 2,08].
Небольшое количество рассматриваемых городов и симуляции означают, что у нас не может быть много уверенность в полученных значениях.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Метод Монте-Карло

Повторная выборка

Моделирование

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Дэвисон, АС, Хинкли, DV: Bootstrap Методы и их применение. Издательство Кэмбриджского университета, Кембридж (1997)

Эфрон, Б.: Методы начальной загрузки: другой посмотри на складной нож. Анна. Стат. 7, 1–26 (1979)

Эфрон, Б., Тибширани, Р. Дж.: Введение к бутстрапу. Чепмен и Холл, Нью Йорк (1993)

Холл, П.: Bootstrap и Эджворт Расширение. Springer, Берлин Гейдельберг Нью-Йорк (1992)

Шао, Дж., Ту, Д.: Складной нож и ремешок для ботинок. Springer, Берлин Гейдельберг Новый Йорк (1995)

Боскович, Роджер Дж.

Боскович, Роджер Джозеф родился в Рагузе. (ныне Дубровник, Хорватия) в 1711 г. Он посещал Collegium Ragusinum, затем он отправился в Рим, чтобы учиться в Collegio Романо; оба колледжа были иезуитскими. Он умер в 1787 году в Милане.

В 1760 г. Боскович разработал геометрический метод нахождения простой регрессии. линия L1 (для исправления ошибок звездных наблюдений). Лаплас, Пьер Симон де не одобрял тот факт, что Боскович использовал геометрическую терминологию и перевел этот метод в алгебраическую версию в 1793.

Machine Translated by Google

Основная работа Босковича, Роджера Джозеф:

1757 г. De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex instancearia etiam sensorum impress sa. Bononiensi Scientiarium et Ar tium Instituto Atque Academia Commentarii, Tomus IV, стр. 353–396.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Уайт, Ланселот Ло: Роджер Джозеф Боскович, Очерки его жизни и творчества к 250-летию со дня его рождения. Аллен и Анвин, Лондон (1961)

Коробка, ЕР Джордж

Бокс, Джордж Э.П. родился в 1919 году в Англии. Он работал химиком в Великобритании.

Армейские инженеры во время Второй мировой войны. После войны он получил степень в области математики и статистика Университетского колледжа Лондона. В 1950-х годах он работал приглашенным профессором в США в Институте статистики.

в Университете Северной Каролины. В 1960 году он переехал в Висконсин, где служил первый председатель отдела статистики. Он получил медаль Британской империи в 1946 году и медалью Шухарта в 1968 году.

Его основной интерес заключался в экспериментальной статистике и планировании экспериментов. Его книга Статистика для экспериментаторов, в соавторстве с Хантером, Дж. Стюартом и Хантером, Уильямом G., является одним из наиболее рекомендуемых текстов в этой области. Коробка тоже написала вовремя анализ серии.

Рекомендуемая публикация Бокс, Джордж:

2005 (с Хантером, Уильямом Г. и Хантер, Дж. Стюарт) Статистика для Exper imenters: дизайн, инновации и Discovery, 2-е изд. Wiley, Нью-Йорк

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

План экспериментов

Коробка Сюжет

Блочная диаграмма — это способ представления следующих пяти величин для набора данных: медиана ; первый квартиль и третий квартиль; максимальные и минимальные значения.

Блочная диаграмма — это диаграмма (коробка), которая иллюстрирует ставки:

- Мера центральной тенденции (в основная медиана);
- Изменчивость и;
- Симметрия.

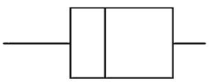
Он часто используется для сравнения нескольких наборов наблюдения.

ИСТОРИЯ

Сюжет коробки и усов, или сюжет коробки, был введенный Тьюкейном в 1972 г., наряду с другими методами полуграфического представления данных, одним из наиболее известных из которых является схема стебля и листа.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Существует несколько способов представления коробчатой диаграммы. Мы представим его следующим образом:



Центральный прямоугольник представляет 50% площади. наблюдения. Он известен как межквартальный диапазон. Нижний предел этого прямоугольника фиксируется в первом квартиле, а верхний предел

Machine Translated by Google

в третьем квартиле. Положение медианы обозначено линией, проходящей через прямоугольник.

Отрезок линии («ус») соединяет каждый квартиль к соответствующему экстремуму (минимуму или максимуму, если выбросы не подарок; см. ниже) значение с каждой стороны прямоугольник.
В этом представлении выбросы обрабатываются особым образом. Когда наблюдения будут очень разбросаны, мы определяем два значения, называемые «внутренние пределы» или «бакенбарды»:

int.lim.1 = 1-й квартиль
(1,5 · межквартильный размах)
int.lim.2 = 3-й квартиль
+ (1,5 · межквартильный размах)

Затем для каждого внутреннего предела мы выбираем значение данных, которое ближе всего к пределу, но все же внутри промежутка между внутренним пределом его. Эти два значения данных называются соседними точками. Теперь, когда мы строим коробку график, отрезки линии соединяют квартили на соседние точки. Наблюдения снаружи интервала между внутренними пределами (выбросы) представлены звездочками на графике.

ПРИМЕР

В следующем примере представлены доходы индексы на одного жителя для каждого швейцарского кантона (швейцарский доход = 100) в 1993 г.:

Кантон	Индекс Кантон	Индекс
Цюрих	125,7 Шаффхаус	99,2
Берн	86.2 Аппенцелль Прав.-Внешн.	84,2
Люцерн	87,9 Аппенцелль Rh.-Int.	72,6
Ури	88.2 Сен-Галль	89,3
Швиц	94,5 Граубюнден	92,4
Обвальд	80.3 Аргови	98,0
Нидвальд	108.9 Тургови	87,4
Гларис	101.4 Тессин	87,4
Зуг	170,2 Во	97,4
Фрибург	90,9 Вале	80,5
Солер	88.3 Невшатель	87,3
Бале-Сити	124.2 Женева	116,0
Бале Кампания	105.1 Юра	75,1

Источник: Федеральное статистическое управление (1993 г.).

Теперь мы вычисляем различные квартили:
Блочная диаграмма дает нам информацию о центральной тенденции, о дисперсии распределения и степени симметрии:

- Центральная тенденция:
  Медиана равна 90,10 .
- Рассеивание:
  Межквартильный интервал указывает на интервал, который содержит 50% наблюдения ции, и эти наблюдения являются ближе всего к центру распространения.
  В нашем примере имеем:
  – 50% кантонов Швейцарии имеют индекс, попадающий в интервал [102,33 87,03] (при ширине 15,3).

Границы коробчатой диаграммы следующие:

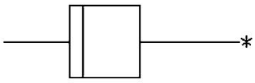
lim.int.1 = 1-й квартиль
(1,5 · межквартильный интервал)
= 87,03 - 1,5 · 15,3 = 64,08 .

lim.int.2 = 3-й квартиль
+ (1,5 · межквартильный интервал)
= 102,33 + 1,5 · 15,3 = 125,28 .

Значения выше 125,28 являются выбросами. Их два: 125,7

Machine Translated by Google

(Цюрих) и 170,2 (Зуг). Коробчатый сюжет
затем следующим образом:



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Графическое представление
- Исследовательский анализ данных
- Мера местоположения
- Мера центральной тенденции

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Tukey, JW: Некоторые графические и полуграфические дисплеи. В: Бэнкрофт, Т.А. (ред.) Статистические статьи в честь Джорджа У.
Снедекор. Издательство государственного университета Айовы, Эймс, ИА, стр. 293–316 (1972).

Machine Translated by Google

Категориальные данные

Категориальные данные состоят из подсчетов наблюдений, попадающих в указанные классы.

Мы можем различать различные типы категориальных данных:

- Бинарные, характеризующие наличие или отсутствие имущества;
- Неупорядоченный многокатегориальный (также называемый «номинальным»);
- Упорядоченный многокатегориальный (также называемый «порядковый»);
- Целые числа.

Мы представляем категориальные данные в виде таблицы сопряженности.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Переменные, которые по существу непрерывны, также могут быть представлены как категориальные переменные.

Одним из примеров является «возраст», который представляет собой непрерывную переменную, но возрасты по-прежнему можно сгруппировать в классы, чтобы их можно было представлять в виде категориальных данных.

ПРИМЕРЫ

В опросе общественного мнения для одобрения или неодобрения нового закона голоса могут быть либо «за», либо «против». Мы можем представить результаты в виде таблицы непредвиденных обстоятельств:

	да	нет
Голосов	8546	5455

Если разделить работников предприятия по профессиям (а представлено не менее трех профессий), то данные, которые мы получим, являются неупорядоченными многокатегориальными данными (естественной упорядоченности профессий нет).

Напротив, если нас интересует количество людей, достигших различных уровней образования, то, вероятно, будет иметь место естественное упорядочение категорий: «начальное, среднее» и затем университетское. Таким образом, такие данные будут примером упорядоченных многокатегориальных данных.

Наконец, если мы сгруппируем сотрудников по категориям на основе размера семьи каждого сотрудника (то есть количества членов семьи), мы получим категориальные данные, где категории представляют собой целые числа.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Анализ категорийных данных
- Двоичные данные
- Категория
- Данные
- Дихотомическая переменная
- Качественная категориальная переменная
- Случайная переменная

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

См. анализ категорийных данных.

Machine Translated by Google

Категория

Категория представляет собой набор людей или объектов, имеющих общие характеристики.

Если мы хотим изучить людей в популяции, мы можем рассортировать их по «естественным» категориям, например по полу (мужчины и женщины), или по категориям, определяемым другими критериями, такими как призвание (менеджеры, секретари, фермеры...).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Двоичные данные
- Категориальные данные
- Дихотомическая переменная
- Население
- Случайная переменная
- Переменная

Распределение Коши

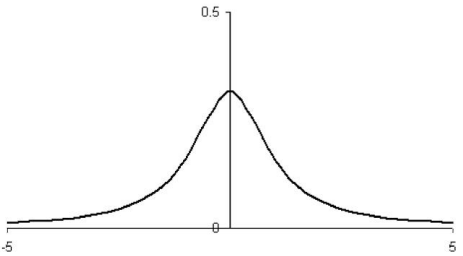
Случайная величина X подчиняется распределению Коши, если ее функция плотности имеет вид:

$$f(x) = \frac{1}{\pi\theta} \cdot \frac{1}{1 + \left(\frac{x - a}{\theta}\right)^2},$$

$\theta > 0$.

Параметры α и θ являются параметрами местоположения и дисперсии соответственно.
Распределение Коши симметрично относительно x = α, который представляет собой медиану. Первый квартиль и третий квартиль задаются как α ± θ.

Распределение Коши является непрерывным распределением вероятностей.



Распределение Коши, θ = 1, α = 0

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение E[X] и дисперсия Var(X) не существуют.

Если α = 0 и θ = 1, распределение Коши идентично распределению Стьюдента с

одна степень свободы.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Его важность в физике в основном связана с тем, что он является решением дифференциальной

уравнение, описывающее силовой резонанс.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Непрерывное распределение вероятностей
- Студенческое распределение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Коши, Алабама: Sur les résultats moyens d'observations de même nature, et sur les résultats les plus probables. ЧР акад. науч. 37, 198–206 (1853 г.)

Причинная эпидемиология

Цель причинно-следственной эпидемиологии состоит в том, чтобы определить, как причина связана со следствием в отношении здоровья человека.

Другими словами, это изучение причин болезней, включающее в себя попытки найти статистические данные.

Machine Translated by Google

доказательства причинно-следственной связи или ассоциации между болезнью и фактором предложил вызвать болезнь.

ИСТОРИЯ

См. эпидемиологию.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

См. причину и следствие в эпидемиологии, соотношении шансов и шансов, относительном риске, атрибутивном риске, предотвратимом риске, уровне заболеваемости, коэффициенте распространенности.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Исследования взаимосвязи между курением табака и развитием рака легких, а также взаимосвязи между ВИЧ и СПИДом являются примерами причинной эпидемиологии. Исследование причинно-следственной связи часто бывает очень сложным, требующим множества исследований, а также включения и комбинирования различных наборов данных, начиная с биологических экспериментов и экспериментов на животных и заканчивая клиническими испытаниями. Хотя причины не всегда могут быть точно идентифицированы, знание факторов риска, связанных с болезнью, и, следовательно, группы риска позволяет нам принимать превентивные меры, которые могли бы сохранить здоровье.

ПРИМЕРЫ

В качестве примера можно исследовать связь между курением и развитием рака легких. Рассмотрим исследование 2000 человек: 1000 курильщиков и 1000 некурящих. Распределение по возрасту и пропорции мужчин и женщин идентичны в обеих группах. Проанализируем сводку данных, полученных за многие годы. Результаты представлены в следующей таблице:

	Курильщики	Не курильщики
Количество случаев рака легких (/1000)	50	10
Доля людей в этой группе, которые заразились рак легких	5%	1%

Если мы сравним долю курильщиков, заболевших раком легких, с долей некурящих, которые заболеют, то получим 5%/1% = 5, поэтому мы можем заключить, что риск развития рака легких у курильщиков в пять раз выше, чем у некурящих. Обычно мы также оцениваем значимость этого результата, вычисляя значение p . Предположим, что критерий хи-квадрат дал p < 0,001. Обычно считается, что если p-значение меньше 0,05, то полученные результаты являются статистически значимыми.

Предположим, что мы проводим то же исследование, но размерность каждой группы (курящих и некурящих) составляет 100, а не 1000, и мы наблюдаем те же пропорции людей, заболевших раком легких:

	Курильщики	Не курильщики
Количество случаев рака легких (/100)	5	1
Доля лиц в этом группа с контрактурой легких рак	5%	1%

Если мы проведем тот же статистический тест, значение p окажется равным 0,212. Поскольку это больше 0,05, мы не можем сделать какой-либо твердый вывод о существовании значимой статистической связи между курением и раком легких. Это показывает, что для того, чтобы иметь статистически значимую

Machine Translated by Google

не может уровень разницы между результатами

для разных популяций, полученных в результате эпидемиологических исследований, обычно необходимо изучение больших выборок.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

См. эпидемиологию.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

См. эпидемиологию.

Причина и следствие в эпидемиологии

В эпидемиологии «причина» — это агент (микробные микробы, загрязненная вода, курение и т. д.), который изменяет здоровье, а «следствие» описывает то, как агент изменяет здоровье. Агент часто является потенциально патогенным (в этом случае он известен как «фактор риска»).

Таким образом, эффект представляет собой сравнение рисков. В этом контексте мы можем определить два разных типа риска:

- Абсолютный эффект причины выражает увеличение риска или дополнительное
  число случаев заболевания, которые являются результатом или
  может возникнуть в результате воздействия этой причины. Он измеряется атрибутивным риском и его производными.
- Относительный эффект причины выражает силу связи между возбудителем и болезнью.

Причина, которая сама по себе производит следствие, называется достаточной.

ИСТОРИЯ

Определены понятия «причина» и «следствие».

при зарождении эпидемиологии, которое произошло в XVII в.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Формально имеем:

$$\text{Абсолютный эффект} = \text{риск воздействия} - \text{риск для неэкспонированных}.$$

Абсолютный эффект выражает избыточный риск или случаи заболевания, которые возникают (или могут возникнуть) в результате воздействия причины.

$$\text{Относительный эффект} = \frac{\text{риск для выставленных}}{\text{риск для неэкспонированных}}.$$

Относительный эффект выражает силу связи между болезнью и

причина. Он измеряется с использованием относительного риска и отношения шансов.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Строго говоря, силы связи между тем или иным фактором и болезнью недостаточно для установления причинной связи между ними. Нам также необходимо учитывать:

- «критерий временности» (мы должны быть уверены, что предполагаемая причина предшествует следствию), и;
- Фундаментальные и экспериментальные элементы исследования, которые позволяют нам быть уверенными, что предполагаемый причинный фактор на самом деле не является «фактором путаницы» (то есть фактором, который не является причинным, но статистически связан с неопознанным реальным причинным фактором).

Этим относительным и абсолютным эффектам соответствуют два типа причинности. как правило, это не позволяет нам заранее судить о клиническом или общественном воздействии соответствующего эффекта. Оно может быть сильным, когда риск подвергнуться воздействию очень высок или когда риск остаться необлученным очень низок.

Абсолютная причинно-следственная связь выражает клиническое или общественное воздействие соответствующего эффекта,

Machine Translated by Google

и поэтому позволяет нам ответить на вопрос: если бы мы подавили причину, то что уровень воздействия на население (в период случаев болезни) можно было бы избежать? Если пациент бросил курить, что бы снижение риска развития легкого рак или инфаркт миокарда быть?

Фактор риска, связанный с высоким относительным эффект, но который касается только небольшого числа людей, вызовет меньше болезней и смерть, то фактор риска, который связан с меньшим относительным эффектом, но где много подвергается большему количеству лиц. Таким образом, ясно, что значение причинно-следственной связи отношение меняется в зависимости от того, рассматривают относительную или абсолютную причинность.

Мы также должны отметить здесь важный момент о причинно-следственных связях. Может быть много причины той же болезни. Хотя все эти причины способствуют одному и тому же результату, они могут также взаимодействовать. Главное последствие этого причинное взаимодействие заключается в том, что мы не можем предвосхитить эффект одновременного воздействия причин A и B (обозначается A+B), основанный на том, что мы знать о влиянии воздействия только A или только Б. В отличие от случая для независимых причины, мы должны оценивать совместное действие, а не ограничимся отдельными анализами взаимодействующих причин.

ПРИМЕРЫ

Относительный эффект и абсолютный эффект

могут быть истолкованы по-разному, как показывает следующий пример.

Предположим, у нас есть две популяции P1 и P2, в каждом по 100 000 особей. В популяции P1 риск заражения данным заболеванием составляет 0,2% для подвергшихся воздействию и 0,1% для лиц, подвергшихся воздействию. неэкспонированный. В популяции P2 риск экспонированный составляет 20%, а для неэкспонированного составляет 10%, как показано в следующей таблице:

Население	Риск для незащищенный (%)	Риск для неэкспонированный (%)
	А	Б
P1	0,2	0,1
P2	20	10

Население	Родственник эффект $\frac{A}{Б}$	Абсолют эффект (%) С "=" А Б	Избегаемый случаи С ×100000
P1	2.0	0,1	100
P2	2.0	10	10000

Относительный эффект одинаков для популяций П1 и П2 (отношение риска для подвержены риску для не подвергавшихся опасности 2), но воздействие одних и тех же профилактических мер будет очень разным в двух популяциях, потому что абсолютный эффект равен десяти. раза важнее в P2: количество потенциально предотвратимые случаи поэтому 100 в популяции P1 и 10000 в популяции P2.

Теперь рассмотрим заболеваемость легких рак у людей, которые выкуривать 35 и более сигарет в день: 3,15/1000/год. Хотя эта скорость может показаться маленький, это маскирует тот факт, что есть сильный относительный эффект (риск в 45 раз больше для курящих, затем для некурящих) из-за тот факт, что рак легких очень редко встречается у некурящих (уровень заболеваемости среди некурящих составляет 0,07/1000/год).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Атрибутивный риск
- Предотвратимый риск
- Уровень заболеваемости
- Шансы и отношение шансов
- Уровень распространенности
- Относительный риск

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Лилиенфельд, А.М., Лилиенфельд, Д.Э.: Основы эпидемиологии, 2-е изд. Кларен Дон, Оксфорд (1980)

MacMahon, B., Pugh, TF: Эпидемиология: Принципы и методы. Маленький Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Ротманн, Дж. К.: Эпидемиология. Введение. Издательство Оксфордского университета (2002)

Перепись

Перепись – это операция, состоящая из наблюдения за всеми особями популяции . Слово «перепись» может относиться к переписи населения, другими словами, к переписи населения.
количество, но также может относиться к запросам (называемым «исчерпывающие» запросы), где мы извлекаем информацию о популяции путем наблюдения за всеми особями популяции.
Ясно, что такие запросы будут очень дорогими для очень больших групп населения. Поэтому исчерпывающие исследования проводятся редко; отбор проб, который заключается в наблюдении за лишь часть населения (так называемая образец), обычно предпочтительнее.

ИСТОРИЯ

Переписи возникли вместе с великими цивилизациями древности, когда большие площади империй и сложности, связанной с для управления ими требовалось знание задействованных популяций.
Среди древнейших цивилизаций она Известно, что переписи проводились в Шумере (между 5000 и 2000 г. до н.э.), где вовлеченные люди сообщали списки люди и товары на глиняных столах клинописью персонажи.

Переписи производились также в Месопотамии (около 3000 г. до н. э.), а также в Древний Египет от первой династии по приходам; эти переписи проводились из-за военные и фискальные цели. При Амасисе II все должны были (с риском для жизни) объявить свою профессию и источник (и) rev энуэ.

Ситуация в Израиле была более сложной: переписи были иногда обязательными и иногда запрещено из-за Старого Теста мент. Это влияние на христианскую цивилизацию ция длилась довольно долго; в середине Века, св. Августин и св. Амбруаз были по-прежнему осуждает переписи.
В Китае была проведена перепись по крайней мере с 200 г. до н.э., в разных формах и для разных целей. Хехт, Дж. (1987) сообщил основные переписи:

1. Династия Хань (200 лет до н.э. – 200 лет до н. э.). AD): переписи населения были связаны с система воинской повинности.
2. От периода Троецарствия до пяти династий (221–959 гг. Н.э.): связано с системой. территориального распределения.
3. Династии Сун и Юань (960–1368 гг.). AD): переписи проводились для фискальных целей.
4. Династия Мин (1368–1644 гг. н.э.): «желтый реестры» были созданы на десятилетний переписи. Они перечисляли имя, профессию, пол и возраст каждого человека.
5. Династия Цин (с 1644 г. н.э.): переписи населения. проводились с целью изучения миграции населения.

В Японии в средние века различные используются виды переписи. Первая перепись, вероятно, проводилась по правилу императора Судзиня (86 г. до н.э.).
Наконец, в Индии политическая и экономическая научный трактат под названием «Артхашастра» (прибыль договор) дает некоторую информацию об использовании

Machine Translated by Google

чрезвычайно подробной записи. Этот трактат был написан Каутильей, премьер-министром в правление Чандрагупты Маурьи (313–289 гг. до н. э.).

Другая очень важная цивилизация, инки, также использовала переписи. Они использовали систему статистики под названием «quipos». Каждый quipo был и инструментом, и реестром информации. Цвета, комбинации и узлы на шнурах, образованные из ряда шнуров, имели точное значение. Кипо передавались специально посвященным охранникам, которые собирали всю эту статистику.

В Европе древние греки и римляне также практиковали переписи. Аристотель сообщил, что греки пожертвовали меру пшеницы за рождение и меру ячменя за смерть богине Афине. В Риме первая перепись была проведена по приказу царя Сервия Туллия (578–534 гг. до н. э.) с целью контроля доходов и, следовательно, повышения налогов.

Позже, в зависимости от страны, переписи практиковались с разной периодичностью и в разных масштабах.

В 786 году Карл Великий приказал пересчитать всех своих подданных старше двенадцати лет; подсчет населения был также начат в Италии в двенадцатом столетии; во многих городах в XV в.

века, включая Нюрнберг (в 1449 г.) и Страсбург (в 1470 г.). В шестнадцатом веке Франция ввела регистры семейного положения.

В семнадцатом веке развились три различные школы мысли: немецкая школа, связанная с описательной статистикой, французская школа, связанная с идеологией и методологией переписи, и английская школа, которая привела к современной статистике.

В истории переписей населения Европы есть страна, занимающая особое место. В 1665 году Швеция инициировала реестры приходов. ионеры, которых содержали пастыри; в 1668 г. указ сделал обязательным учет еди в этих реестрах, и реестры из религиозных стали административными. В 1736 г. появился еще один указ, в котором говорилось, что губернатор каждой провинции должен сообщать парламенту о любых изменениях в населении провинции. Статистические данные о населении Швеции были официально признаны 3 февраля 1748 г. в связи с созданием «Tabellverket» (административных таблиц).

Первую сводку по всем шведским провинциям, проведенную в 1749 г., можно считать первой полноценной переписью в Европе, а 11 ноября 1756 г.

«Tabellkommissionen», первого официального статистического отдела. С 1762 г. эти таблицы цифр ведутся Академией наук.

Первоначально шведские переписи проводились ежегодно (1749–1751 гг.), а затем каждые три года (1754–1772 гг.), но с 1775 г. они проводятся каждые пять лет.

В конце XVIII в. был создан официальный институт переписей населения.

Франция (в 1791 г.). В 1787 г. принцип ценза был прописан в Конституционной хартии США (УК США).

См. также официальную статистику.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Сбор данных
- Демография
  - Официальная статистика
- Население
- Образец
- Опрос

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Hecht, J .: L'idee du dénombrement jusqu'à la Révolution. Pour une histoire de la

Machine Translated by Google

статистика, том 1, стр. 21–82.
Экономика / INSEE (1978)

Центральная предельная теорема

Центральная предельная теорема является фундаментальной теорема статистики. В простейшем виде он предписывает, чтобы сумма достаточно большого количество независимых одинаково распределенных случайных величин примерно равно нормальное распределение.

ИСТОРИЯ

Центральная предельная теорема была впервые сформулирована в рамках биноми

al распределение Муавра , Авраама де (1733 г.). Лаплас, Пьер Симон де (1810) сформулировал доказательство теоремы. Пуассон, Симеон Дени (1824) также работал по этой теореме, а Чебышев, Пафнутий Львович (1890–1891) дал строгую демонстрация этого в середине девятки десятый век.

В начале XX века русский математик Ляпунов,

Александр Михайлович (1901) создал общепризнанная форма центральной предельной теоремы путем введения ее характеристических функций. Марков Андрей Андреевич (1908) также работал над ним и первым обобщить теорему на случай независимых переменных.

Согласно Le Cam, L. (1986), квалификатор «центральный» ему присвоил Джордж Полиа. (1920) из-за важной роли, которую он играет в теории вероятностей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X1, X2,...,Xn — n независимых случайных величин, одинаково распределенных ed (с любым распределением) со средним значением μи конечная дисперсия σ2.

Определим сумму Sn = X1 + X2 + ...+ Xn и устанавливаем соотношение:

$$\frac{Sn \quad n \cdot \mu}{\sigma \cdot \sqrt{n}},$$

где $n \cdot \mu$ и $\sigma \cdot \sqrt{n}$ представляют собой среднее значение, а стандартное отклонение Sn соответственно . Центральная предельная теорема устанавливает, что распределение этого соотношения стремится к стандартному нормальное распределение, когда n стремится к бесконечности. Это значит, что:

$$n \quad \frac{Sn \quad n \cdot \mu}{\sigma \quad n} \quad x \quad n \quad + \quad (x)$$

где (x) – функция распределения стандартного нормального распределения, выражается:

$$(x) = \int_{\text{Инс}} \frac{1}{\sqrt{2\pi}} \text{ опыт } \frac{x2}{2} \quad \text{дх} ,$$

$$< x < \quad .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Центральная предельная теорема дает простой метод приближенного вычисления вероятностей, связанных с суммами случайных величин способен.

Помимо своего интереса по отношению к сам теорема Плинга , где суммы и средние значения играют важную роль, центральная теорема используется для аппроксимации нормального распределения , полученные суммированием идентичных распределений. Мы можем, например, с с помощью центральной предельной теоремы используйте нормальное распределение для аппроксимации биномиальное распределение, распределение Пуассона, гамма - распределение, распределение хи -квадрат, распределение Стьюдента , гипергеометрическое распределение, распределение Фишера и логнормальный распределение.

Machine Translated by Google

ПРИМЕРЫ

В крупной партии электротехнических изделий вероятность выбора бракованного изделия равна $\frac{1}{8}$. Какова вероятность того, что будут выбраны 4 дефектных предмета , когда 25 предметов выбран?

Пусть X — дихотомическая переменная .

результат судебного разбирательства:

$$X = \begin{cases} 1, & \text{если выбранный товар неисправен;} \\ 0, & \text{если это не так.} \end{cases}$$

Случайная величина X следует за Бернулем li с параметром p. Следовательно, сумма Sn = X1+X2+...+Xn подчиняется биномиальному распределению со средним np и дисперсии np(1 p), которые, следуя центральной предельной теореме, могут быть аппроксимированы нормальным распределением со средним μ = np и дисперсия σ2 = np(1 p).
Мы оцениваем эти значения:

$$\mu = n \cdot p = 25 \cdot \frac{1}{8} = 3{,}125$$

$$\sigma^2 = n \cdot p(1 \quad p) = 25 \cdot \frac{1}{8} \cdot \quad 1 \quad \frac{1}{8}$$

$$= 2{,}734 .$$

Затем мы вычисляем P(Sn > 4) в двух разных способы:

1. При биномиальном распределении:

Из биномиальной таблицы вероятность P (Sn 4) = 0,8047. Вероятность P (Sn > 4) тогда:

$$P(Sn > 4) = 1 - P(Sn \quad 4) = 0{,}1953 \quad ,$$

2. При нормальном приближении (получено из центральной предельной теоремы):
Для учета дискретного символа действующее лицо случайной величины Sn, мы должны сделать поправку на непрерывность; то есть мы вычислить вероятность того, что Sn больше тогда 4+ $\frac{1}{2}$ = 4,5

У нас есть:

$$r = \frac{Sn - n \cdot p}{\sqrt{n \cdot p(1 \quad p)}}$$

$$"=" \frac{4{,}5 - 3{,}125}{1{,}654}$$

$$= 0{,}832 .$$

Из обычной таблицы получаем вероятность:

$$P(Z > z) = P(Z > 0{,}832)$$

$$= 1 - P(Z \quad 0{,}832)$$

$$= 1 - 0{,}7967 = 0{,}2033 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Биномиальное распределение
Распределение хи-квадрат
Конвергенция
Теорема сходимости
Распределение Фишера
Гамма-распределение
Гипергеометрическое распределение
Закон больших чисел
Логнормальное распределение
Нормальное распределение
распределение Пуассона
Вероятность
Распределение вероятностей
Студенческое распределение

ССЫЛКА

Лаплас, PS de: Mémoire sur les apquimations des formul qui sont foctions de très grands nombres et sur leur application aux probabilités. Воспоминания де l'Académie Royale des Sciences de Paris, 10. Воспроизведено в: Œuvres de Laplace. 12, 301–347 (1810 г.)

Ле Кам, Л.: Центральная предельная теорема около 1935 г. Стат. науч. 1, 78–96 (1986)

Machine Translated by Google

Ляпунов А. М.: Sur une proposition de la théorie des probabilités. Бюллетень Императорской академии наук Санкт-Петербурга 8, 1–24 (1900)

Марков, А.А.: Расширение теоретических расчетов вероятностей по некоторым количественным показателям лжи в цепи. Мем. акад. науч. СПб. 8, 365–397 (1908).

Moivre, A. de: Approximatio ad summam terminorum binomii (a + b) n, in seriem expansi. Приложение II к Miscellanae Analytica, стр. 1–7 (1733 г.). Фотографически перепечатано в редкой брошюре о Муавре и некоторых его открытиях. Паб

опубликовано Арчибальдом, RC Isis 8, 671–683 (1926)

Poisson, SD: Surla probabilité des résultats moyens des наблюдения. Connaissance des temps pour l'an 1827, стр. 273–302 (1824 г.)

Polyà, G .: Ueber den zentralen Grenzwert satz der Wahrscheinlichkeitsrechnung und dasMomentproblem.Mathematische Zeitschrift 8, 171–181 (1920)

Чебычев, П.Л. (1890–1891). Sur deux théorèmes relatifs aux probabilités. Акта Математика. 14, 305–315

Чебышев Пафнутий Львович

Чебышев, Пафнутий Львович (1821–1894) начал учиться в Московском университете в 1837 году, где на него оказали влияние Зернов, Николай Ефимович (первый русский, получивший степень доктора математических наук) и Брашман, Николай Дмитриевич. После получения степени он не мог найти преподавательской работы в Москве, поэтому отправился в Санкт-Петербург, где организовывал конференции по алгебре и теории вероятностей. В 1859 году он прошел курс вероятностей, данный Буни.

Аковский Виктор Яковлевич в Санкт-Петербургском университете.

Его имя живет благодаря неравенству Чебышева (также известному как неравенство Бьенеме–Чебышева), которое он доказал.

Это было опубликовано на французском языке сразу после того, как Bien aymé, Ирене-Жюль опубликовала статью на ту же тему в Journal de Math ématiques Pures et Appliquées (также называемом Journal of Liouville).

Он инициировал кропотливую работу по установлению общей версии центральной предельной теории. рем и считается основателем

математическая школа Санкт-Петербурга.

Некоторые основные работы и статьи Чебышев, Пафнутий Львович:

1845 г. Очерк элементарного анализа теории вероятностей (диссертация) Журнал Крелла.

1867 Preuve de l'inégalité de Chebychev. Дж. Матем. Чистый. Приложение, 12, 177–184.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Центральная предельная теорема

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Heyde, CE, Seneta, E .: IJ Bienaymé. Ожидается статистическая теория. Спрингер, Берлин Гейдельберг, Нью-Йорк (1977)

Неравенство Чебышева

См. закон больших чисел.

Расстояние хи-квадрат

Рассмотрим таблицу частот с n строками и p столбцами, можно рассчитать профили строк и профили столбцов. Давайте тогда построим

Machine Translated by Google

n или p точек из каждого профиля. Мы можем определить расстояния между этими точками. Евклидово расстояние между компонентами профилей, на которые взвешивается определено (каждый терм имеет вес, т. е. обратное его частоте), называется расстоянием хи-квадрат. Название дистанции вытекает из того, что математическое выражение, определяющее расстояние, идентично с тем, что встречается при разработке Хи-квадрат критерия согласия.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть(fij) — частота i - й строки и j-й столбец в таблице частот с n строками san n столбцы. Расстояние хи-квадрат между две строки i и i задаются формулой:

$$d(i, i') = \sqrt{\sum_{j=1}^{n} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \cdot \frac{1}{f_{.j}}},$$

где

$f_{i.}$ представляет собой сумму компонентов i -го ряд;

$f_{.j}$ представляет собой сумму компонентов j -го столбец;

$\left[\frac{f_{ij}}{f_{i.}}\right]$ — профиль i-й строки для j = 1, 2,...,p.

Аналогично, расстояние между двумя столбцами j и j определяется как:

$$d(j, j') = \sqrt{\sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \cdot \frac{1}{f_{i.}}},$$

где $\left[\frac{f_{ij}}{f_{.j}}\right]$ — профиль j-го столбца для j = 1,...,n.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Расстояние хи-квадрат включает в себя вес, обратно пропорциональный сумма каждой строки (или столбца), которая увеличивает значимость небольших отклонений в строках (или столбцах), которые имеют небольшая сумма по сравнению с теми, у кого больше важная сумма пакет.

Расстояние хи-квадрат имеет свойство дистрибутивной эквивалентности, что означает, что гарантирует, что расстояния между рядами и столбцы инвариантны, когда два столбца (или два ряда) с одинаковыми профилями.

ПРИМЕРЫ

Рассмотрим таблицу непредвиденных обстоятельств , показывающую, как довольные сотрудники, работающие в трех разных компаниях. Давайте установим дистанцию таблица с использованием расстояния хи-квадрат.

Значения исследуемой переменной X могут попасть в одна из трех категорий:

- X1: высокая удовлетворенность;
- X2: среднее удовлетворение;
- X3: низкая удовлетворенность.

Наблюдения, собранные из выборок лица из трех предприятий приведены ниже:

	Буси 1	Буси 2	Буси 3	Общий
X1	20	55	30	105
X2	18	40	15	73
X3	12	5	5	22
Всего	50	100	50	200

Таблица относительных частот получается разделить все элементы таблицы на 200, общее количество наблюдений:

	Буси 1	Буси 2	Буси 3	Общий
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
Итого	0,25	0,5	0,25	1

Machine Translated by Google

Мы можем рассчитать разницу в удовлетворенности сотрудников трех предприятий. Матрица профиля колонны приведена ниже:

	Буси 1	Буси 2	Буси 3	Общий
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
Всего 1		1	1	3

Это позволяет вычислить расстояния между разными столбцами:

$$d2(1, 2) = \frac{1}{0,525} \cdot (0,4 \quad 0,55)^2 + \frac{1}{0,365} \cdot (0,36 \quad 0,4)^2 + \frac{1}{0,11} \cdot (0,24 \quad 0,05)^2 = 0,375423$$

$$д(1, 2) = 0,613$$

Аналогичным образом можно вычислить d(1, 3) и d(2, 3). Полученные расстояния сведены в следующую таблицу расстояний:

	Буси 1	Буси 2	Буси 3
Бизнес 1	0	0,613	0,514
Бизнес 2	0,613	0	0,234
Бизнес 3	0,514	0,234	0

Мы также можем вычислить расстояния между строк, другими словами, разница в удовлетворенности сотрудников; для этого нам понадобится таблица профиля линии:

	Буси 1	Буси 2	Буси 3	Общий
X1	0,19	0,524	0,286	1
X2	0,246	0,548	0,206	1
X3	0,546	0,227	0,227	1
Всего	0,982	1,299	0,719	3

Это позволяет вычислить расстояния между разными строками:

$$d2(1, 2) = \frac{1}{0,25} \cdot (0,19 \quad 0,246)^2 + \frac{1}{0,5} \cdot (0,524 \quad 0,548)^2 + \frac{1}{0,25} \cdot (0,286 \quad 0,206)^2 = 0,039296$$

$$д(1, 2) = 0,198$$

Мы можем вычислить (1, 3) и d (2, 3) аналогичным образом. Различия между степенями удовлетворенности сотрудников, наконец, сведены в следующую таблицу расстояний:

	Ик1	Ик2	Ик3
X1	0	0,198	0,835
X2	0,198	0	0,754
X3	0,835	0,754	0

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Таблица сопряженности
Расстояние

Таблица расстояний

Частота

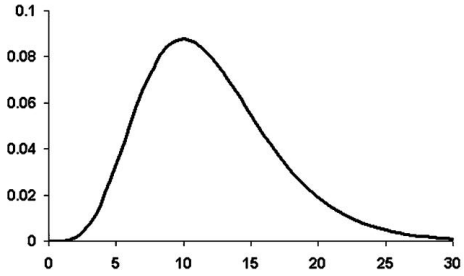
Распределение хи-квадрат

Случайная величина X следует хи-квадрату распределение с n степенями свободы, если его функция плотности :

$$e(x) = \frac{2^{\frac{н}{x}} \quad 1 \quad \text{опыт} \quad \frac{x}{2}}{2^{\frac{н}{2}} \quad \frac{н}{2}}, \quad x \quad 0 \quad ,$$

где гамма-функция (см. Гамма распределение).

Machine Translated by Google



распределение χ2 , ν = 12

Распределение хи-квадрат является непрерывным распределение вероятностей.

ИСТОРИЯ

Согласно Шейнину (1977), хи-квадрат распространение было обнаружено Эрнстом Карлом Аббе в 1863 году. Максвелл получил его за три степени свободы несколько лет назад (1860 г.), а Больцман открыл Дело 1881 года.

Однако, согласно Ланкастеру (1966), Бьенеме получил распределение хи-квадрат в 1838 г. как предел дискретного ранга. домашняя переменная

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

если (N1, N2,...,Nk) подчиняются совместному мультиномиальному распределению параметров n, p1, p2,...,pk.

Эллис показал в 1844 г., что сумма k случайных величин, распределенных согласно к распределению хи-квадрат с двумя степенями свободы следует распределению хи-квадрат с 2k степенями свободы. Генерал результат был продемонстрирован в 1852 г. Бьен Эме.

В этой области большое значение имеют работы Пирсона, Карла . В 1900 году он использовал ци квадратное распределение для аппроксимации статистики хи-квадрат , используемой в различных тестах, основанных на на таблицах непредвиденных обстоятельств.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Распределение хи-квадрат появляется в теория распределенных случайных величин по нормальному распределению. В этом, это распределение суммы квадратов нормальные, центрированные и редуцированные случайные величины (со средним значением , равным 0, и дисперсией равно 1).

Считаем Z1, Z2,...,Zn, n независимыми, стандартные нормальные случайные величины. Их сумма квадратов:

$$X = Z_1^2 + Z_2^2 + \ldots + Z_n^2 \overset{"="}{=} \sum_{i=1}^{n} Z_i^2$$

случайная величина, распределенная по закону распределение хи-квадрат с n степенями свобода.

Ожидаемое значение распределения хи-квадрат определяется выражением:

$$E[X] = n \quad .$$

Дисперсия равна :

$$Var(X) = 2n \quad .$$

Распределение хи-квадрат связано с другими непрерывные распределения вероятностей:

- Распределение хи-квадрат является особым Случай гамма-распределения.
- Если две случайные величины X1 и X2 подчиняются распределению хи-квадрат с соответственно n1 и n2 степенями свободы, тогда случайная величина

$$Y = \frac{X1/n1}{X2/n2}$$

следует распределению Фишера с n1 и n2 степеней свободы.

- Когда число степеней свободы n стремится к бесконечности, распределение хи-квадрат стремится (относительно медленно) к нормальному распределению.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Распределение хи-квадрат используется во многих подходах к проверке гипотез, наиболее важным из которых является критерий согласия , который включает сравнение наблюдаемых частот и гипотетических частот конкретных классов.

Он также используется для сравнения наблюдаемой дисперсии и гипотетической дисперсии нормально распределенных выборок, а также для проверки независимости двух переменных.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Хи-квадрат критерия согласия
- таблица хи-квадрат
- критерий хи-квадрат
- Непрерывное распределение вероятностей
- Распределение Фишера
- Гамма-распределение
- Нормальное распределение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Ланкастер, НО: Предшественники хи-квадрата сына груши. Ауст. Дж. Стат. 8, 117–126 (1966)

Пирсон, К.: О критерии, что данная система отклонений от вероятного в случае коррелированной системы переменных такова, что можно разумно предположить, что она возникла в результате случайной выборки. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, стр. 339–357. Впервые опубликовано в 1900 году в Philos.Mag. (5-я серия) 50, 157–175 (1948)

Шейнин О.Б.: К истории некоторых статистических законов распределения. В: Кендалл, М., Плакетт, Р.Л. (ред.) Исследования по истории статистики и вероятностей, том. II. Гриффин, Лондон (1977)

Хи-квадрат Тест на пригодность

Критерий согласия хи-квадрат, наряду с критерием Колмогорова-Смирнова, является одним из наиболее часто используемых критериев согласия. тесты.

Этот тест направлен на определение того, является ли можно аппроксимировать наблюдаемое распределение конкретным распределением вероятностей (нормальным распределением, распределением Пуассона и т. д.).

ИСТОРИЯ

Критерий согласия хи-квадрат является старейшим и наиболее известным тестом согласия. Впервые он был представлен в 1900 году Пирсоном, Карлом.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X1,...,Xn — выборка наблюдений.

Шаги, используемые для выполнения критерия согласия хи-квадрат, следующие: 1. Сформулируйте гипотезу. Нулевая гипотеза примет следующий вид:

$$H0 : \Phi = \Phi 0 ,$$

где F0 — предполагаемая функция распределения распределения.

2. Распределить наблюдения по k непересекающимся классам:

$$[ай-1, ай) .$$

Обозначим количество наблюдений содержащихся в i-м классе, i = 1,...,k, на ni.

3. Рассчитайте теоретические вероятности для каждого класса на основе предполагаемой функции распределения F0:

$$pi = F0(ai) \quad F0(ai \quad 1) , \qquad я = 1,...,к.$$

Machine Translated by Google

4. Получите ожидаемые частоты для каждого сорт

$$ei = n \cdot pi, \qquad i = 1,\ldots,k,$$

где n — размер выборки .

5. Рассчитайте статистику χ^2 (хи-квадрат):

$$x2 = \sum_{i=1}^{k} \frac{(ni - ei)^2}{ei}.$$

Если H0 истинно, статистика χ^2 следует распределению хи-квадрат с v степенями свободы, где:

$$v = k - 1 - \text{количество предполагаемых параметры}.$$

Например, при проверке качества соответствия нормальному распределению число степеней свободы равно:

- k 1, если среднее значение μ и стандартное отклонение σ совокупности известны;
- k 2, если одно из μилиσ неизвестно и будет оценено для продолжения с тестом;
- k 3, если оба параметра μ и σ неизвестно, и оба оцениваются из соответствующие значения выборки.

6. Отклонить H0 , если отклонение между наблюдаемые и расчетные частоты большой; то есть:

$$\text{если } x2 > x2_{v,\alpha},$$

где $x2_{v,\alpha}$ это значение , указанное в чи квадратный стол для особого значения уровень a.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Чтобы применить хи-квадрат силы наиболее приспособленных, это важно, чтобы n было достаточно большим и чтобы предполагаемые частоты, ei, не слишком малы. Обычно мы утверждаем, что предполагаемые частоты должны быть больше 5, за исключением экстремальные классы, где их может быть меньше тогда 5, но больше 1. Если это ограничение не удовлетворены, мы должны перегруппировать классы в чтобы удовлетворить этому правилу.

ПРИМЕРЫ

Качество соответствия биномиальному Распределение

Бросаем монетку четыре раза и считаем количество раз, когда появляется «орел». Этот опыт проводится 160 раз. Наблюдаемые частоты следующие:

Количество «голов» xi	Количество эксперименты (ни)
0	17
1	52
2	54
3	31
4	6
Общий	160

1. Если опыт проведен правильно и монета не подделана, то распределение количества полученных «орлов» должно подчиняться биномиальному распределению.

   Затем мы формулируем нулевую гипотезу о том, что наблюдаемое распределение может быть аппроксимировано биномиальным распределением, и мы продолжится с тестом на пригодность чтобы определить, можно ли принять эту гипотезу или нет.

2. В этом примере разное количество «головы», которые можно получить за эксперимент (0, 1, 2, 3 и 4), считаются быть классом.

Machine Translated by Google

3. Случайная величина X (количество «орел», полученный после четырех бросков монета) следует биномиальному распределению, если

$$P(X = x) = C_{n}^{x} \cdot p^{x} \cdot q^{n-x},$$

где:

n - количество независимых испытаний = 4;

p - вероятность успеха («головы») = 0,5;

q — вероятность отказа («решка») = 0,5;

C_{n}^{x} это количество комбинаций x объектов из n.

Тогда мы имеем следующую теоретическую вероятности для четырех бросков:

$$P(X = 0) = \frac{1}{16}$$
$$P(X = 1) = \frac{4}{16}$$
$$P(X = 2) = \frac{6}{16}$$
$$P(X = 3) = \frac{4}{16}$$
$$P(X = 4) = \frac{1}{16}$$

4. После проведения эксперимента 160 раз, ожидаемое количество голов для каждого возможного значения X определяется как :

$$e_i = 160 \cdot P(X = x_i).$$

Получаем следующую таблицу:

Количество «головы» x_i	Наблюдаемый частота (o_i)	Ожидал частота (e_i)
0	17	10
1	52	40
2	54	60
3	31	40
4 610		
Общий	160	160

5. Тогда статистика χ2 (хи-квадрат):

$$x2 = \sum_{i=1}^{k} \frac{(o_i - e_i)2}{e_i},$$

где k - количество возможных значений X. _

$$x2 = \frac{(17 \quad 10)2}{10} + \ldots + \frac{(6 \quad 10)2}{10}$$

$$= 12{,}725 .$$

6. Выбрав уровень значимости α, равный 5%, мы найти, что значение $\chi^2_{v,\alpha}$ для к - 1 = 4 степеней свободы это:

$$x2_{4,0,05} = 9{,}488 .$$

Поскольку расчетное значение χ2 больше чем значение, полученное из таблицы, мы отвергаем нулевую гипотезу и заключаем, что биномиальное распределение не дает хорошего приближения к нашему наблюдаемое распределение. Затем мы можем кон предположить, что монеты, вероятно, были подделаны, или что их не правильно кинули.

Качество подгонки к нормальному распределению

Исследованы диаметры кабелей, выпускаемых заводом.
Таблица частот наблюдаемого распределения диаметров приведена ниже:

Диаметр кабеля (в мм)	Наблюдаемая частота o_i
19.70–19.80	5
19.80–19.90	12
19.90–20.00	35
20.00–20.10	42
20.10–20.20	28
20.20–20.30	14
20.30–20.40	4
Общий	140

Machine Translated by Google

1. Мы проводим тест на пригодность для нормальное распределение. Таким образом, нулевая гипотеза состоит в том, что наблюдаемое распределение можно аппроксимировать нормальным распределение.
2. В предыдущей таблице показаны наблюдаемые диаметры делятся на классы.
3. Если случайная величина X (диаметр кабеля) следует нормальному распределению, случайная величина

$$Z = \frac{X - \text{мю}}{\text{о}}$$

следует стандартному нормальному распределению.

Среднее значение μ и стандартное отклонение σ совокупности неизвестны и оцениваются с использованием среднего $\bar{x}$ и стандартное отклонение S выборки :

$$\bar{x} = \frac{\sum_{\text{я}=1}^{7} \delta\text{i} \cdot \text{ни}}{\text{н}} \text{ "=" } \frac{2806{,}40}{140} = 20{,}05 \quad ,$$

$$C = \frac{\sum_{\text{я}=1}^{7} \text{ni} \cdot (\delta\text{i} - \bar{x})^2}{\text{п} - 1} \text{ "=" } \frac{2{,}477}{139}$$

$$= 0{,}134$$

где δi — центры класса es (в данном примере средний диаметр класс; для i = 1: δ1 = 19,75) и n равно общее количество наблюдений.

Затем мы можем рассчитать теоретическую вероятности, связанные с каждым классом. Подробные расчеты для первых двух классы:

$$\text{p1} = \text{P(X} \quad 19{,}8)$$

$$= \text{P}\left(\text{Z} \quad \frac{19{,}8 - \bar{x}}{\text{C}}\right)$$

$$= \text{P (Z} \quad -1{,}835)$$

$$= 1 - \text{P (Z} \quad 1{,}835)$$

$$\text{p2} = \text{P}(19{,}8 \quad \text{X} \quad 19{,}9)$$

$$= \Pi \quad \frac{19{,}8 - \bar{x}}{\text{C}} \quad \text{Z} \quad \frac{19{,}9 - \bar{x}}{\text{C}}$$

$$= \text{P (} \quad 1{,}835 \quad \text{Z} \quad 1{,}089)$$

$$= \text{P (Z} \quad 1{,}835) - \text{P (Z} \quad 1{,}089)$$

Эти вероятности можно найти, обратившись к обычной таблице. Мы получаем:

$$\text{p1} = \text{P(X} \quad 19{,}8) = 0{,}03325$$

$$\text{p2} = \text{P}(19{,}8 \quad \text{X} \quad 19{,}9) = 0{,}10476$$

$$\text{p3} = \text{P}(19{,}9 \quad \text{X} \quad 20{,}0) = 0{,}22767$$

$$\text{p4} = \text{P}(20{,}0 \quad \text{X} \quad 20{,}1) = 0{,}29083$$

$$\text{p5} = \text{P}(20{,}1 \quad \text{X} \quad 20{,}2) = 0{,}21825$$

$$\text{p6} = \text{P}(20{,}2 \quad \text{X} \quad 20{,}3) = 0{,}09622$$

$$\text{p7} = \text{P(X} > 20{,}3) = 0{,}02902$$

4. Ожидаемые частоты занятий затем даются:

$$\text{ei} = \text{п} \cdot \text{пи} \ ,$$

что дает следующую таблицу:

Кабель диаметр (в мм)	Наблюдаемый частота ни	Ожидал частота эй
19.70–19.80	5	4.655
19.80–19.90	12	14.666
19.90–20.00	35	31,874
20.00–20.10	42	40.716
20.10–20.20	28	30.555
20.20–20.30	14	13.471
20.30–20.40	4	4.063
Общий	140	140

5. Тогда статистика χ2 (хи-квадрат):

$$\text{x2} = \sum_{\text{я}=1}^{\text{к}} \frac{(\text{ни} - \text{еи})2}{\text{эй}} \ ,$$

Machine Translated by Google

где k = 7 — количество классов.

$$x2 = \frac{(5 \quad 4{,}655)2}{4.655} + \frac{(12 \quad 14{,}666)2}{14.666} + \ldots + \frac{(4 \quad 4{,}063)2}{4.063} = 1{,}0927 .$$

6. Выбор уровня значимости α = 5%, получаем, что значение $\chi 2_{v,\alpha}$ c k 3 = 7 3 = 4 степени свободы в таблице хи-квадрат:

$$x2_{4,0,05} = 9{,}49 .$$

Поскольку значение, рассчитанное по χ2 , равно меньше значения, полученного из таблица хи-квадрат, мы не отвергаем нуль гипотезу, и мы заключаем, что разница между наблюдаемым распределением а нормальное распределение не значимо при уровне значимости 5%.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Распределение хи-квадрат

таблица хи-квадрат

Тест на пригодность

Проверка гипотезы

Тест Колмогорова – Смирнова

ССЫЛКА

Пирсон, К.: По критерию того, что данный система отклонений от вероятного в случай коррелированной системы переменных таков, что можно разумно предположить, что она возникла в результате случайной выборки. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, стр. 339–357. Впервые опубликовано в 1900 г. Филос.маг. (5-я серия) 50, 157–175 (1948)

Таблица хи-квадрат

Таблица хи-квадрат дает значения получается из функции распределения случайной величины , которая следует за хи квадратное распределение.

ИСТОРИЯ

Одна из первых таблиц хи-квадрата была опубликована в 1902 году Элдертоном. Это содержит значения функции распределения , которые заданы до шести знаков после запятой.

В 1922 году Пирсон и Карл опубликовали таблицу значения для неполной гамма-функции, до семи знаков после запятой.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X будет случайной величиной , которая следует распределение хи-квадрат с v степенями свобода. Функция плотности случайной величины X имеет вид:

$$\phi(\tau) = \frac{\tau^{\frac{в}{2}} \quad 1 \text{ опыт } \quad 2^{\frac{\tau}{}}}{2^{\frac{в}{2}} \quad \frac{в}{2}}, \quad \tau \quad 0 ,$$

где представляет собой гамма-функцию (см. гамма-распределение).
Функция распределения случайных переменная X определяется:

$$F(x) = P(X \quad x) = \int_0^{икс} f(t)\,dt.$$

Таблица хи-квадрат дает значения функция распределения F(x) для различных значений v.

Мы часто используем таблицу хи-квадрат обратным способом, чтобы найти значение x , соответствующее данной вероятности. Обычно мы обозначаем через $\chi 2_{v,\alpha}$ стоимость случайная величина X, для которой

$$P(X \quad \chi 2\ v,\alpha) = 1 \quad \alpha .$$

Примечание: обозначение χ2 читается как «хи-квадрат».

Machine Translated by Google

ПРИМЕРЫ

См. Приложение F.

Таблица хи-квадрат позволяет нам при заданном число степеней свободы v, чтобы определить:

1. Значение функции распределения F(x), учитывая x.
2. Значение $\chi^2_{v,\alpha}$, учитывая вероятность P(X $\chi^2_{v,\alpha}$).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Распределение хи-квадрат
Хи-квадрат критерия согласия
критерий хи-квадрат
Критерий независимости хи-квадрат
Статистическая таблица

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Элдертон, В.П.: Таблицы для проверки соответствия теории наблюдениям. Биометрика 1, 155–163 (1902).

Пирсон, К.: Таблицы неполных - функция. Офис канцелярских принадлежностей Ее Величества (издательство Кэмбриджского университета, Кембридж с 1934 г.), Лондон (1922 г.)

Хи-квадрат Тест

Существует ряд тестов хи-квадрат, все которые включают сравнение результатов теста с значения из распределения хи-квадрат. Наиболее известными из этих тестов являются вводные ниже:

- Критерий независимости хи- квадрат используется для определения того, являются ли две качественные категориальные переменные, связанные с выборка независимы .
- Критерий согласия хи- квадрат используется для определения того, является ли распределение наблюдаемое для образца может быть аппроксимировано теоретическим распределением. Мы может захотеть узнать, например, распределение, наблюдаемое для sam соответствует определенному распределению вероятностей (нормальному распределению, распределение Пуассона и др.).
- Тест хи-квадрат для неизвестной переменной ance используется, когда мы хотим проверить, эта дисперсия принимает определенную постоянную ценить.
- Критерий хи-квадрат используется для проверки однородность рассчитанных дисперсий для многих образцов, взятых из нормального распределенное население.

ИСТОРИЯ

В 1937 г. Бартлет М.С. предложил метод проверка однородности дисперсии для многие выборки взяты из нормально распределенной популяции.
См. Также критерий независимости хи-квадрат. и критерий согласия хи-квадрат.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Математические аспекты хи-квадрата тест независимости и критерий согласия хи-квадрат рассматриваются в соответствующие им записи.
Тест хи-квадрат используется для проверки того, неизвестная дисперсия принимает конкретное постоянное значение :
Пусть (x1,...,xn) — случайная выборка из нормально распределенной совокупности неизвестного среднего μ и неизвестной дисперсии σ2.

У нас есть веские основания полагать, что дисперсия населения равна предполагаемой значение σ^2_0. Гипотезы для каждого случая описано ниже.

A: Двусторонний чехол:

$$H0: \quad \sigma 2 = \sigma 2_0$$
$$H1: \quad \sigma 2 = \sigma 2_0$$

Machine Translated by Google

B: Односторонний тест:

$$H0: \quad \sigma 2 \quad \sigma 2_0$$

$$H1: \quad \sigma 2 > \sigma 2_0$$

C: Односторонний тест:

$$H0: \quad \sigma 2 \quad \sigma 2_0$$

$$H1: \quad \sigma 2 < \sigma 2_0$$

Затем мы определяем статистику данного критерий хи-квадрат с использованием:

$$x2 = \frac{\sum_{i=1}^{n} (xi \quad \bar{x})^2}{\sigma 2_0}.$$

Эта статистика при H0 распределена хи-квадрат с n 1 степенями свободы. В Другими словами, ищем значение $\chi 2_{n \quad 1,\alpha}$ в таблице хи-квадрат, и затем мы сравниваем это значение с вычисленным значением χ2. Правила принятия решений зависят от случая, и следующие.

Случай А

Если $\chi 2 \quad \chi 2_{n \quad 1,\alpha 1}$ или если $\chi 2 \quad \chi 2_{n \quad 1,1 \quad \alpha 2}$ мы отвергнуть нулевую гипотезу H0 для альтернативной гипотезы H1, где мы разделили уровень значимости α на α1 и α2 такой что α1+α2 = α. В противном случае мы не отвергаем нулевая гипотеза H0.

Случай Б

Если $x2 < x2_{n \quad 1,\alpha}$ мы не отвергаем ноль гипотеза H0. Если $\chi 2 \quad \chi 2_{n \quad 1,\alpha}$ мы отвергаем нулевая гипотеза H0 для альтернативной гипотезы H1.

Случай С

Если $x2 > x2_{n-1,1-a}$ мы не отвергаем нулевая гипотеза H0. Если $\chi 2 \quad \chi 2_{n-1,1-a}$ мы отвергнуть нулевую гипотезу H0 в пользу альтернативной гипотезы H1.
Другие тесты хи-квадрат предлагаются в работа Остле, Б. (1963).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

χ2 (хи-квадрат) статистика должна быть рассчитана используя абсолютные частоты, а не относительные те.

Обратите внимание, что критерий хи-квадрат может быть ненадежным. для небольших выборок, особенно когда некоторые из предполагаемые частоты малы (< 5).
Эту проблему часто можно решить путем группировки категорий вместе, если такая группировка разумный.

ПРИМЕРЫ

Рассмотрим партию изделий, произведенных машина. Их можно разделить на классов в зависимости от их диаметра (в мм), как в следующей таблице:

Диаметр (мм) xi	Количество предметов ни
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
Общий	N = 100

У нас есть случайная выборка из нормально распределенная популяция, где среднее значение и дисперсия неизвестны. Продавец этих товаров хотел бы, чтобы дисперсия σ2 быть меньше или равно 0,05. Мы тестируем

Machine Translated by Google

следующие гипотезы:

нулевая гипотеза H0 : σ2 0,05

альтернативная гипотеза H1 : σ2 > 0,05 .

В этом случае мы используем одностороннюю гипотезу тест.

Начнем с вычисления среднего для выборки:

$$x^{-} = \frac{\sum_{i=1}^{11} n_i \cdot x_i}{H} \text{ "=" } \frac{5995}{100} = 59{,}95 .$$

Затем мы можем вычислить статистику χ2 :

$$x2 = \frac{\sum_{i=1}^{11} n_i \cdot (x_i - \bar{x})2}{\sigma_0^2}$$

$$\text{"="} \frac{2 \cdot (\ 0{,}45)2 + \ldots + 2 \cdot (0{,}55)2}{0{,}05}$$

$$\text{"="} \frac{3{,}97}{0{,}05} = 79{,}4 .$$

Используя уровень значимости α = 5%, мы затем найти значение $\chi^2_{99,0,05}$ (= 123,2) в таблица хи-квадрат.

Поскольку $\chi2 = 79{,}4 < \chi^2_{99,0,05,}$ мы не отвергаем нулевая гипотеза, означающая, что продавец должен быть рад продать эти предметы так как они существенно не отличаются диаметр.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение хи-квадрат
- Хи-квадрат критерия согласия
- таблица хи-квадрат
- Критерий независимости хи-квадрат

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бартлетт, М.С.: Некоторые примеры статистических методов исследования в сельском хозяйстве и прикладная биология. Дж. Рой. Стат. соц. (Прил.) 4, 137–183 (1937)

Остле, Б.: Статистика в исследованиях: основные концепции и методы для научных работников. Издательство Государственного колледжа Айовы, Эймс, Айова (1954)

C

Хи-квадрат Тест Независимости

Критерий независимости хи-квадрат направлен на определить, связаны ли две переменные с выборкой независимы или нет.

изучаемые переменные являются категориальными качественными переменные.

Критерий независимости хи-квадрат выполняется с использованием таблицы непредвиденных обстоятельств.

ИСТОРИЯ

Первые таблицы сопряженности использовались только для перечисления. Однако, поощряемый работа Кетле, Адольф (1849 г.), статистики начали интересоваться связи между переменными, используемыми в столы. Например, Пирсон, Карл (1900) выполнил фундаментальную работу по таблицам непредвиденных обстоятельств.

Юл, Джордж Удни (1900) предложил несколько иной подход к изучению таблицы непредвиденных обстоятельств к Пирсону, которые приводят к разногласиям между ними. Пирсон также спорил с Фишером, Рональд Эйлмер о количестве степеней свободы использовать в тесте независимости хи-квадрат. Все использовали разные числа, пока Фишер, Р.А. (1922) в конце концов не доказал, что правильный.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим две качественные категориальные переменные X и Y. У нас есть выборка , содержащая n наблюдений этих переменных.

Machine Translated by Google

Эти наблюдения можно представить в

таблица непредвиденных обстоятельств.

Обозначим наблюдаемую частоту
категория i переменной X и категория j
переменной Y как nij.

Категории переменной Y

Категории из переменная X	Y1 ...	Yc	Итого
X1	n11 ...	n1c	n1.
...	...	... ...	...
Xr	nr1 ...	nrc	Всего n.1 ... nc
			номер н..

Гипотезы, подлежащие проверке, следующие:

Нулевой хайп. H0 : две переменные независимы,

Альтернативный хайп. H1 : две переменные не являются независимыми.

Шаги, связанные с тестом

1. Вычислите ожидаемые частоты,
   обозначается eij, для каждого случая в таблице непредвиденных обстоятельств при независимости гипотеза:

$$\text{эйдж} = \frac{\text{ни.} \cdot \text{нью-йорк}}{\text{н..}},$$

$$\text{ни.} = \sum_{k=1}^{c} \text{ник и ндж} = \sum_{k=1}^{p} \text{нкдж},$$

   где с представляет количество столбцов (или количество категорий переменной X в непредвиденных обстоятельств) и количество строк (или количество категорий переменной Y).

2. Рассчитайте значение χ2 (хи-квадрат).
   статистика, которая на самом деле является мерой отклонение наблюдаемых частот nij из ожидаемых частот eij:

$$x2 = \sum_{i=1}^{c} \sum_{j=1}^{p} \frac{(nij - eij)2}{\text{Эйдж}}.$$

3. Выберите уровень значимости, который будет использоваться.
   в тесте и сравнить рассчитанное значение χ2 со значением, полученным из таблица хи-квадрат, $\chi^2_{v,\alpha.}$ Номер степеней свободы соответствуют количество случаев в таблице, которые могут принимать произвольные значения; значения, принятые другие дела возлагаются на них итоги строк и столбцов. Итак, количество Степени свободы задаются:

$$v \text{ знак равно } (r - 1) (c - 1) .$$

4. Если расчетное χ2 меньше, то $\chi^2_{v,\alpha}$ из таблицы, нуль не отбрасываем гипотеза. Две переменные можно считать независимыми.
   Однако, если вычисленное значение χ2 больше тогда $\chi^2_{v,\alpha}$ из таблицы, мы отвергаем нулевая гипотеза для альтернативной гипотезы. Тогда мы можем сделать вывод, что оба переменные не являются независимыми.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Для того, чтобы
чтобы уметь применять хи-квадрат критерия независимости:

1. Выборка, содержащая n наблюдений, должна быть случайной;
2. Каждое отдельное наблюдение может только появляются в одной категории для каждой переменной. Другими словами, каждое отдельное наблюдение может появляться только в одном строка и один столбец непредвиденных обстоятельств стол.

Обратите внимание, что критерий независимости хи-квадрат не очень надежен для небольших выборок, т.к. особенно когда расчетные частоты малы (< 5). Чтобы избежать этой проблемы, мы можем группировать категории вместе, но только тогда, когда это полученные группы разумны.

Machine Translated by Google

ПРИМЕРЫ

Мы хотим определить, не зависит ли доля курильщиков от пола.

Две изучаемые переменные являются категориальными и качественными и содержат две категории каждая:

- Переменная «пол:» M или Ж;
- Переменная «статус курения:» «курит» или «не курит».

Тогда гипотезы:

H0 : шанс стать курильщиком независимо от пола

H1 : шанс быть курильщиком невелик независимо от пола.

Таблица непредвиденных обстоятельств , полученная из выборка 100 человек (n = 100) показано ниже:

Статус курения

	«курит»	«делает нет дым»	Общий
Пол M 21		44	65
Ф	10	25	35
Всего 31		69	100

Теперь обозначим наблюдаемые частоты как nij (i = 1, 2, j = 1, 2).

Затем мы оцениваем все частоты в таблица, основанная на предположении , что два переменные не зависят друг от друга. Мы обозначим эти предполагаемые частоты через eij:

$$\text{эйдж} = \frac{\text{ни.} \cdot \text{нью-йорк}}{\text{н..}} .$$

Таким образом, мы получаем:

$$e11 = \frac{65 \cdot 31}{100} = 20{,}15$$

$$e12 = \frac{65 \cdot 69}{100} = 44{,}85$$

$$e21 = \frac{35 \cdot 31}{100} = 10{,}85$$

$$e12 = \frac{35 \cdot 69}{100} = 24{,}15 .$$

Приведена расчетная таблица частот ниже:

Статус курения

	«курит»	«делает нет дым»	Общий
Пол M 20.15		44,85	65
Ф	10,85	24.15	35
Всего 31		69	100

Если нулевая гипотеза H0 верна, статистика

$$x2 = \sum_{я=1}^{2} \sum_{j=1}^{2} \frac{(nij - eij)2}{\text{Эйдж}}$$

хи-квадрат распределен с (r - 1) (c - 1) = (2 - 1) (2 - 1) = 1 градус свободы и

$$\chi 2 = 0{,}036 + 0{,}016 + 0{,}066 + 0{,}030 = 0{,}148 .$$

Если выбран уровень значимости 5%, значение $\chi 2_{1,0,05}$ 3,84, из хи-квадрата стол.

Поскольку расчетное значение χ2 меньше затем значение, найденное в таблице хи-квадрат, мы не отвергаем нулевую гипотезу и сделать вывод, что две изучаемые переменные независимый.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение хи-квадрат
- таблица хи-квадрат
- Таблица сопряженности
- Тест на независимость

Machine Translated by Google

ССЫЛКА

Фишер, Р.А.: Об интерпретации χ2
из таблиц непредвиденных обстоятельств и
расчета PJRoy. Стат. соц. сер. A85, 87–94
(1922)

Пирсон, К.: По критерию того, что данный
система отклонений от вероятного в
случай коррелированной системы переменных
таков, что можно разумно предположить, что
она возникла в результате случайной выборки.
В: Ранние статистические статьи Карла Пирсона.
Издательство Кембриджского университета,
стр. 339–357. Впервые опубликовано в 1900 г.
Филос.маг. (5-я серия) 50, 157–175 (1948)

Кетле, А.: Письма, адресованные Е.Р.Х.
Великий герцог Саксен-Кобург-Готский, на
Теория вероятностей в применении к
Моральные и политические науки. (Французский
перевод Даунса, Олинтуса Грегори).
Чарльз и Эдвин Лейтон, Лондон (1849 г.)

Юле, Г.У.: Об ассоциации атрибутов в
статистика: с иллюстрацией от помощника
риал общества детства.Philos.Trans.
Рой. соц. Лонд. сер. A 194, 257–319
(1900)

Классификация

Классификация – это объединение
подобные объекты. Если каждый объект
характеризуется p переменными, классификация может
осуществляться по рациональным критериям.
В зависимости от используемых критериев объект
потенциально могут принадлежать к нескольким классам.

ИСТОРИЯ

Классификация жителей населенного пункта или
страну в зависимости от их пола и других физических
характеристик - это деятельность, которая
восходит к античным временам. Индусы, т.
древние греки и римляне все развили
открыл несколько типологий человека.
Самый старый из них происходит от Галена (129–199 гг.).
ОБЪЯВЛЕНИЕ).

Позже понятие классификации распространилось
в области биологии и зоологии; в
в этой связи следует упомянуть работы Линнея
(1707–1778).
Первые представления о реальных методах
кластерный анализ приписывают Адансону
(восемнадцатый век). Зубин (1938), Трион
(1939) и Торндайк (1953) также пытались
разработать некоторые методы, но истинное
развитие методов классификации совпадает
с появлением компьютера.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Методы классификации можно разделить на
две большие категории, одна основана на
вероятностных способностях , а другая нет.

К первой категории относится, например,
дискриминационный анализ. Вторая категория
можно еще разделить на две группы.
В первую группу входят так называемые
оптимальные методы классификации. В секунду
группы, мы можем различать несколько
подвиды метода классификации:

- Методы разделения, состоящие из распределения
  n объектов среди g групп в таких
  далеко, что каждый объект принадлежит исключительно
  только к одной группе. Количество групп г
  фиксируется заранее, а раздел
  применяется наиболее точно удовлетворяет
  критериям классификации.
- Методы разделения, включающие классы,
  нарушающие авторские права, когда объекту разрешено
  относятся одновременно к нескольким группам.
- Методы иерархической классификации,
  где структура данных в dif
  учитываются различные уровни классификации
  счет. Затем мы можем принять во внимание
  отношения, существующие между различными

Machine Translated by Google

другие группы, созданные во время раздела.
Два разных вида иерархической технологии существуют niques: методы агломерации и методы разделения.

Методы агломерации начинаются с отдельных объектов, что означает, что n объектов изначально распределяются по n группам. Две группы объединяются на каждом последующем шаге, пока не останется только одна группа.

Напротив, методы разделения начинаются со всех объектов, сгруппированных вместе, что означает, что все n объектов изначально находятся в одной группе. Новые группы создаются на каждом последующем шаге, пока не будут н группы.

- Геометрическая классификация, при которой объекты изображаются на точечной диаграмме , а затем группируются в соответствии с положением на диаграмме. В графическом представлении близость объектов друг к другу на графике соответствует сходству между объектами.

Первые три типа классификации обычно объединяются под термином «кластерный анализ». Общим для них является тот факт, что классифицируемые объекты должны иметь определенную структуру, которая позволяет нам измерить степень сходства между объектами.

Каждый тип классификации содержит множество методов, позволяющих создавать классы. подобных объектов.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Классификацию можно использовать в двух случаях:

- Описание дел; • Случаи предсказания.

В первом случае классификация производится на основе некоторых общепринятых стандартных признаков. Например, профессии можно классифицировать на фриланс, менеджеры, рабочие и т. д., и можно рассчитать среднюю заработную плату, среднюю частоту проблем со здоровьем и т. д. для каждого класса.

Во втором случае классификация приведет к предсказанию, а затем к действию. Например, если лисы в определенном регионе проявляют апатичное поведение и повышенное слюноотделение, можно сделать вывод о наличии нового заболевания бешенством. эпидемия. Затем это должно вызвать кампанию вакцинации.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Кластерный анализ
- Полный метод связи
- Анализ данных

ССЫЛКИ

Everitt, BS: Кластерный анализ. Холстед, Лондон (1974)

Гордон, AD: Классификация. Методы исследовательского анализа многомерных данных. Чепмен и Холл, Лондон (1981)

Кауфман Л., Руссо П.Дж.: Поиск групп в данных: введение в кластерный анализ. Уайли, Нью-Йорк (1990)

Торндайк, Р.Л.: Кто принадлежит семье? Психометрика 18, 267–276 (1953).

Tryon, RC: Кластерный анализ. Макгроу Хилл, Нью-Йорк (1939)

Зубин, Дж.: Метод измерения единомыслия. Дж. Ненормальный. Социальная психология. 33, 508–516 (1938)

Кластерный анализ

Кластеризация — это разделение набора данных на подмножества или кластеры таким образом, чтобы степень связи была сильной между членами одного и того же кластера и слабой между членами одного кластера.

Machine Translated by Google

различные кластеры в соответствии с некоторой определенной мерой расстояния.

Существует несколько методов проведения кластерного анализа : • Частичная кластеризация • Иерархическая кластеризация.

ИСТОРИЯ

См. классификацию и анализ данных.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Чтобы провести кластерный анализ набора из n объектов, нам нужно определить расстояние между объектами (или, в более общем случае, меру сходства между объектами), которые необходимо классифицировать. Существование какой-либо структуры внутри

предполагается совокупность объектов.

Для проведения иерархической классификации множества E объектов {x1, x2,...,xn} необходимо определить расстояние, связанное с E , по которому можно получить таблицу расстояний

между объектами E. Точно так же расстояние должно быть определено для любых подмножеств E.

Один из подходов к иерархической кластеризации заключается в использовании метода агломерации. Его можно свести к следующему алгоритму:

1. Найдите пару объектов (xi, xj) , расстояние между которыми наименьшее.

   другой.

2. Объединить пару объектов (xi, xj) в один элемент α и восстановить новую таблицу расстояний. Это достигается удалением строк и столбцов, связанных с xi и xj , и заменой их на линию и столбец, связанные с α.

   новая таблица расстояний будет иметь строку и на столбец меньше, чем в предыдущей таблице.

3. Повторяйте эти две операции до тех пор, пока не будет получено нужное количество классов или пока все объекты не будут собраны в один класс.

Обратите внимание, что расстояние между группой, образованной из агрегированных элементов, и другими элементами может быть определено по-разному.

пути, ведущие к различным методам. Примеры включают метод одиночной связи и метод полной связи.

Однозвенный метод — это метод иерархической классификации, в котором используется евклидова

расстояние для создания таблицы расстояний, а

расстояние между двумя классами определяется евклидовым расстоянием между двумя близкими

est элементов (минимальное расстояние).

В методе полной связи расстояние между двумя классами определяется евклидовым расстоянием между двумя элементами .

самые дальние (максимальное расстояние).

Учитывая, что единственная разница между этими двумя методами заключается в том, что расстояние между двумя

классы - это либо минимум, либо максимум

максимальное расстояние, здесь будет рассматриваться только метод с одним звеном.

Тогда для множества E = {X1, X2,...,Xn} устанавливается таблица расстояний между элементами множества E. любил.

Поскольку эта таблица симметрична и пуста по диагонали, рассматривается только одна половина таблицы:

d(X1, X2) d(X1, X3) ... d(X1, Xn) d(X2, X3) ... d(X2, Xn)

... ...

г(Хп-1, Хп)

где d(Xi, Xj) — евклидово расстояние между Xi и Xj для i < j, где значения i и j лежат между 1 и n.

Алгоритм работы с одним звеном следующий:

• Поиск минимума d(Xi, Xj) для

  я < j;

Machine Translated by Google

- Элементы Xi и Xj объединяются в
  новая группа Ck = Xi Xj;
- Затем множество E разбивается на

  {X1},...,{Xi 1},{Xi, Xj},{Xi+1},...,

  {Xj 1},{Xj+1},...,{Xn} ;
- Затем таблица расстояний создается заново
  без ассоциаций строк и столбцов

  с Xi и Xj, и с линией
  и столбец, представляющий расстояния
  между Xm и Ck, m = 1, 2,...,nm =
  i и m = j, определяемые как:

  d(Ck, Xm) = min{d(Xi, Xm); d(Xj, Xm)}.

Алгоритм повторяется до тех пор, пока не будет получено желаемое.
количество групп достигнуто или пока не будет
только одна группа, содержащая все элементы
менты.

В общем случае расстояние между двумя
группы дает:

d(Ck, Cm) = min{d(Xi, Xj) с Xi

принадлежащие Ck , а Xj — Cm} ,

Приведенная выше формула применима к
частный случай, когда группы состоят соответственно
из двух элементов и одного
одиночный элемент.

Этот ряд агломераций может быть представлен
дендрограммой , где по оси абсцисс показано
расстояние, разделяющее объекты.

Обратите внимание, что мы могли найти более одной пары
когда мы ищем пару ближайших элементов. При
этом пара, выбранная для агрегации на первом
шаге, не
влиять на последующие шаги (при условии, что алгоритм
не заканчивается на этом шаге), так как другая пара
ближайших элементов будет агрегирована на
следующем шаге. Порядок агрегации на
дендрограмме не показан
потому что он сообщает расстояние, которое разделяет
два сгруппированных объекта.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Выбор расстояния между

группа, состоящая из агрегированных элементов и
другие элементы могут работать в нескольких
способами, в зависимости от используемого метода,
как, например, в методе одной ссылки и
в методе полной связи.

ПРИМЕРЫ

Давайте проиллюстрируем, как метод одной ссылки
кластерный анализ может быть применен к
экзаменационные оценки, полученные пятью студентами
каждый изучает четыре курса: английский, французский,
математика и физика.
Мы хотим разделить этих пятерых студентов на двоих.
группы по методу одиночного звена.
Оценки, полученные на экзаменах,
которые варьируются от 1 до 6, суммируются в
следующую таблицу:

	английский	французский	математика	физика
Ален 5.0	3,5	4.0	4,5	
Джин 5.5	4.0	5,0	4,5	
4.5 марта	4,5	4.0	3,5	
Пол 4.0	5,5	3,5	4.0	
Пьер 4.0	4,5	3.0	3,5	

Затем мы вычисляем евклидовы расстояния
между учащимися и использовать их для создания
таблица расстояний:

	Ален	Жан	Марк	Поль	Пьер
Ален 0		1,22	1,5		2,35 2
Жан 1,22 0			1,8		2,65 2,74
1,5 марта		1,8 0			1,32 1,12
Пол 2,35 2,65 1,32 0					1,22
Пьер 2		2,74 1,12	1,22 0		

Рассматривая только верхнюю часть этого
симметричная таблица, получаем:

Machine Translated by Google

	Жан	Марк	Поль	Пьер
Ален	1,22	1,5	2,35	2
Жан		1,8	2,65	2,74
Марк			1,32	1.12
Павел				1,22

Минимальное расстояние 1,12, между Марк и Пьер; поэтому мы формируем первую группу из этих двух студентов. Затем вычисляем новые расстояния.

Например, мы вычисляем новое расстояние между Марком и Пьером с одной стороны.

и Алена, с другой, взяв минимальное расстояние между Марком и Аленом и минимальное расстояние между Пьером и Ален:

d({Марк,Пьер},Ален)
= min{d(Марк,Ален);d(Пьер,Ален)}
= мин{1,5; 2} = 1,5 ,

также

d({Марк,Пьер},Жан)
= min{d(Марк,Жан);d(Пьер,Жан)}
= мин{1,8; 2,74} = 1,8 ,

и

d({Марк,Пьер},Поль)
= min{d(Марк,Поль);d(Пьер,Поль)}
= мин {1,32; 1,22} = 1,22 .

Новая таблица расстояний принимает следующее форма:

	Жан	Марк и Поль Пьер	
Ален	1,22	1,5	2,35
Жан		1,8	2,65
Марк и Пьер			1,22

Минимальное расстояние теперь составляет 1,22, между Ален и Жан, а также между группой Марка и Пьера с одной стороны и Пола с другой стороны (другими словами, две пары показать минимальное расстояние); давайте выберем сначала перегруппировать Алена и Жана. Другой пара будет объединена на следующем шаге. Мы перестроить таблицу расстояний и получить:

d({Ален,Жан}, {Марк,Пьер})
= min{d(Ален,{Марк,Пьер}) ,
d(Жан,{Марк,Пьер})}
= мин{1,5; 1,8} = 1,5

а также:

d({Ален,Жан},Поль)
= min{d(Ален,Поль);d(Жан,Поль)}
= мин {2,35; 2,65} = 2,35 .

Это дает следующую таблицу расстояний:

	Марк и Пьер	Поль
Ален и Жан	1,5	2,35
Марк и Пьер		1,22

Обратите внимание, что Пол теперь должен быть интегрирован в группа сформировалась из Марка и Пьера, и новое расстояние будет:

d({(Марк,Пьер),Поль},{Ален,Жан})
= min{d({Марк,Пьер},{Ален,Жан}),
d(Поль,{Ален,Жан})}
= мин{1,5; 2,35} = 1,5

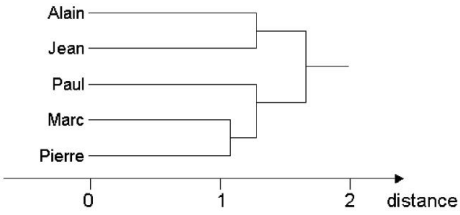
что дает следующую таблицу расстояний:

	Ален и Жан
Марк, Пьер и Поль	1,5

В итоге получаем две группы: {Ален и Жан} и {Марк, Пьер и Paul} , которые разделены расстоянием 1.5.

Machine Translated by Google

Следующая дендрограмма иллюстрирует последовательные агрегаты:



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Классификация
- Полный метод связи
- Дендрограмма
- Расстояние
- Таблица расстояний

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Селё, Г., Дидай, Э., Говерт, Г., Лешеваль Лиер, Ю., Раламбондрани, Х.: Автоматическая классификация понятий — аспекты статистика и информатика. Дуно, Париж (1989)

Эверитт, Б.С.: Кластерный анализ. Холстед, Лондон (1974)

Гордон, AD: Классификация. Методы для Исследовательский анализ многомерных Данные. Чепмен и Холл, Лондон (1981)

Джамбу, М., Лебо, Миссури: Классификация автоматический анализ данных. Дюно, Париж (1978)

Кауфман, Л., Руссеу, П.Дж.: Поиски Группы в данных: введение в кластерный анализ. Уайли, Нью-Йорк (1990)

Лерман, Л.С.: Классификация и анализ порядкового номера. Дюно, Париж (1981)

Томассон Р., Даудин Дж. Дж., Данзарт М., Массон, Дж. П.: Дискриминация и классификация. Массон, Париж (1988)

Выборочное обследование

В кластерной выборке первым шагом является разделение население на подмножества , называемые кластерами. Каждый кластер состоит из особей, должен быть репрезентативным для населения.

Затем кластерная выборка включает выбор случайную выборку кластеров, а затем наблюдение за всеми индивидуумами, принадлежащими к каждому из них.

ИСТОРИЯ

См. выборку.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Кластерная выборка — это процесс случайного извлечения репрезентативных наборов (известный как кластеры) из большей совокупности единиц а затем применить анкету ко всем единицы в кластерах. Кластеры часто состоят из географических единиц, таких как городские районы. В данном случае метод предполагает деление города на районы, а затем выбор районы для включения в выборку. Окончательно, все люди или домохозяйства в выбранном района допрашивают.

Есть две основные причины выполнять выборочное обследование. Во многих запросах есть отсутствие полного и надежного списка единиц совокупности, на которых основывается выборка, или может оказаться, что создание такого списка слишком дорого. Например, во многих странах, в том числе промышленно развитых, редко иметь полные и актуальные списки всех члены населения, домохозяйства или сельские поместья. В этой ситуации выборка может быть достигнута географическим образом: каждый городской район делится на районы и каждый сельский район в сельские поместья. районы и сельскохозяйственные районы считаются кластерами, и мы используем

Machine Translated by Google

полный список кластеров, потому что у нас нет
полный и актуальный список всех единиц населения.
Поэтому мы выбираем необходимое количество
кластеров из списка, а затем
опросить всех юнитов в выбранном класе
тер.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Преимущество кластерной выборки в том, что она
нет необходимости иметь полный и актуальный список
всех единиц совокупности
выполнить анализ.
Например, во многих странах нет
обновленные списки людей или жилья. Расходы
создания таких списков часто являются запретительными.
Поэтому легче анализировать подмножества
населения (так называемые кластеры).
В целом кластерная выборка дает оценки , которые не
столь точны, как простые
случайная выборка, но это падение точности
легко компенсируется гораздо более низкой стоимостью кластера
выборка.
Для максимально эффективной кластерной выборки :

- Кластеры не должны быть слишком большими, и
  должно быть достаточно большое количество
  кластеры,
- размеры кластеров должны быть максимально однородными
  возможный;
- Индивидуумы, принадлежащие к каждому кластеру,
  должны быть как можно более гетерогенными.
  относительно наблюдаемого параметра.

Еще одна причина для использования кластерной выборки:
расходы. Даже при полной и актуальной
список всех единиц населения существует, он может быть
предпочтительно использовать кластерную выборку из
с экономической точки зрения, так как выполняется
быстрее, требует меньшего количества рабочих и
минимизирует транспортные расходы. Поэтому более
целесообразно использовать кластерную выборку, если
деньги, сэкономленные на этом, гораздо важнее
меньше, чем увеличение выборочной дисперсии
что получится.

ПРИМЕРЫ

Рассмотрим N , размер населения
город X. Мы хотим изучить распределение
«Возраст» для города X без проведения переписи.
Население разделено на G частей.
Выполняется простая случайная выборка
среди этих G частей, и мы получаем g частей.
Финальный образец будет состоять из всех
лица в выбранных частях g .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Выборка
- Простая случайная выборка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Хансен, М. Х., Гурвиц, В. Н., Мэдоу,
MG: Методы выборочного обследования и теория.
Том. I. Методы и приложения.
Том. II. Теория. Чепмен и Холл, Лондон (1953)

Коэффициент решимости

Коэффициент детерминации, обозначаемый
R2 — отношение объясняемой вариации (суммы
квадратов из-за регрессии) к
общая вариация (общая сумма квадратов тотальной CC
(TSS)) в модели простой или множественной линейной
регрессии:

$$P2 = \frac{\text{Объясненная вариация}}{\text{Общая вариация}}.$$

Он равен квадрату коэффициента корреляции и может
принимать значения от 0 до 1.
Часто выражается в процентах.

ИСТОРИЯ

См. коэффициент корреляции.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим следующую модель множественной регрессии:

$$Yi = \beta 0 + \beta 1Xi1 + \cdots + \beta pXip + \varepsilon i$$

для i = 1,...,n, где

Yi — зависимые переменные, Xij (i = 1,...,n, j = 1,...,p) — независимые переменные, εi — случайные ненаблюдаемые члены ошибки , βj (j = 1,.. .,p) — оцениваемые параметры.

Оценка параметров β0, β1,..., βp дает оценку

$$\hat{Y}i = \hat{\beta} 0 + \hat{\beta} 1Xi1 + \cdots + \hat{\beta} pXip .$$

Коэффициент детерминации позволяет измерить качество соответствия уравнения регрессии измеренным значениям. Чтобы определить качество подгонки уравнения регрессии , рассмотрим разрыв между наблюдаемым значением и оценкой. сопряженное значение для каждого наблюдения образец. Этот зазор (или остаток) также может быть выражен следующим образом:

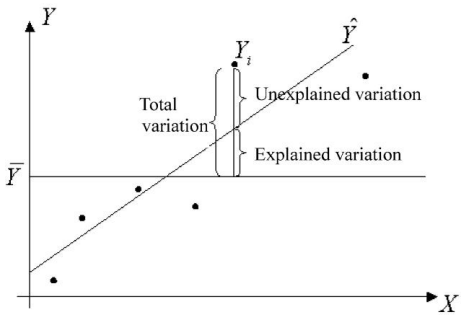
$$\sum_{i=1}^{n} (Yi - \bar{Y})^2 = \sum_{i=1}^{n} (Yi - \hat{Y}i)^2 + \sum_{i=1}^{n} (\hat{Y}i - \bar{Y})^2$$

$$TCC = RSS + REGSS$$

где

TSS — это общая сумма квадратов, RSS — остаточная сумма квадратов, а REGSS — сумма квадратов регрессии.

Эти понятия и отношения между ними представлены на следующем графике:



Используя эти понятия, мы можем определить R2, что является коэффициентом детерминации. Это измеряет долю вариации переменной Y, которая описывается уравнением регрессии следующим образом:

$$P2 = \frac{\text{РЕГСС}}{\text{ТСС}} = \frac{\sum_{i=1}^{n} (\hat{Y}i - \bar{Y})^2}{\sum_{i=1}^{n} (Yi - \bar{Y})^2}$$

Если функция регрессии должна использоваться для прогнозирования последующих наблюдений, предпочтительно иметь высокое значение R2 , поскольку чем выше значение R2, тем меньше необъяснимое изменение.

ПРИМЕРЫ

В следующей таблице приведены значения валового национального продукта (ВНП) и спроса на отечественную продукцию за период 1969–1980 годов для конкретной страны.

Год	ВНП Икс	Спрос на дом эстетические продукты Д
1969 г.	50	6
1970 г.	52	8
1971 г.	55	9
1972 г.	59	10

Machine Translated by Google

Год	ВНП Икс	Спрос на домашние эстетические продукты Д
1973 г.	57	8
1974 г.	58	10
1975 г.	62	12
1976 г.	65	9
1977 г.	68	11
1978 г.	69	10
1979 г.	70	11
1980 г.	72	14

Попытаемся оценить спрос на небольшие товары в зависимости от ВНП согласно модели

$$Yi = a + b \cdot Xi + \varepsilon i , \qquad я = 1,..., 12 \quad .$$

Оценивая параметры a и b по методу наименьших квадратов дает следующее оценщики:

$$б\hat{} = \frac{\sum_{я=1}^{12} (Xi - \bar{X})(Yi - \bar{Y})}{\sum_{я=1}^{12} (Xi - \bar{X})^2} = 0,226 \quad ,$$

$$a\hat{} = \bar{Y} - b\hat{} \cdot \bar{X} = 4,047 .$$

Расчетная строка написана

$$Y\hat{} = 4,047 + 0,226 \cdot X \quad .$$

Качество подгонки измеренных точек линии регрессии задается коэффициентом детерминации:

$$P2 = \frac{РЕГСС}{ТСС} \text{ "=" } \frac{\sum_{я=1}^{12} (Y\hat{}i - \bar{Y})^2}{\sum_{я=1}^{12} (Yi - \bar{Y})^2} .$$

Мы можем вычислить среднее значение , используя:

$$\bar{Y} = \frac{\sum_{я=1}^{12} И}{н} = 9,833 .$$

И	$Y_я$	$(И - Д_2)$	$(Y_я - Д)^2$
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
Общий		47,667	30.489

Таким образом, мы получаем:

$$R2 = = \frac{30.489}{47,667} 0,6396$$

или, в процентах:

$$R2 = 63,96\% .$$

Таким образом, мы можем сделать вывод, что согласно выбранной модели , 63,96 % вариации в спрос на мелкие товары объясняется изменение ВНП.

Очевидно, что значение R2 не может превышать 100%. В то время как 63,96% является относительно высоким, это недостаточно близко к 100%, чтобы исключить попытку для дальнейшей модификации модели .

Этот анализ также показывает, что другие переменные помимо ВНП следует учитывать учет при определении функции, соответствующей спросу на мелкие товары, поскольку ВНП лишь частично объясняет вариация.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коэффициент корреляции

Множественная линейная регрессия

Регрессивный анализ

Простая линейная регрессия

Коэффициент эксцесса

Коэффициент эксцесса используется для измерения уверен пикообразность или плоскостность кривой. Это основывается на моментах распределения. Этот коэффициент является одним из показателей эксцесс.

ИСТОРИЯ

См. коэффициент асимметрии.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Коэффициент эксцесса (β2) основан по центрированному моменту четвертого порядка распределение равно :

$$\mu 4 = E\,(X \quad \mu)4\,.$$

Чтобы получить коэффициент эксцесса не зависящее от единиц измерения, момент четвертого порядка делится стандартным отклонением совокупности σ в четвертой степени. Тогда коэффициент эксцесса становится равным:

$$\beta 2 = \frac{\text{мк4}}{\sigma 4}\,.$$

Для выборки (x1, x2, ... ,xn) оценка этого коэффициента обозначается через b2. Это равно к:

$$б2 = \frac{\text{м4}}{S4}\,,$$

где m4 — центрированный четвертый порядок момент образца, определяемый по формуле:

$$\text{м4} = \frac{1}{\text{н}} \cdot \sum_{\text{я}=1}^{\text{н}} (\text{xi} \quad \bar{}\text{x})^{4}\,,$$

где $\bar{x}$ — среднее арифметическое, n — сумма число наблюдений , S4 – стандартное отклонение выборки, возведенной в четвертая власть.

Для случая, когда случайная величина способный X принимает значения xi с частотами fi, i = 1, 2,...,h, центрированный четвертый порядок момент образца определяется формулой :

$$\text{м4} = \frac{1}{\text{н}} \cdot \sum_{\text{я}=1}^{\text{час}} \text{fi} \cdot (\text{xi} \quad \bar{}\text{x})^{4}\,.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Для нормального распределения коэффициент эксцесс равен 3. Поэтому кривая будет называться платикуртической (имеется в виду более плоская, чем нормальное распределение), если он имеет коэффициент эксцесса меньше 3. Он будет лептокуртичным (то есть более резким, чем нормальное распределение), если β2 больше 3.

Докажем теперь, что коэффициент эксцесса равен 3 для нормального распределения.

Мы знаем, что $\beta 2 = \sigma 4$, а $\frac{\text{мк4}}{\text{это}}$ означает, что центрированный момент четвертого порядка делится на стандартное отклонение возведено в четвертую власть. Можно доказать, что центрированное sth момент порядка, обозначаемый мкс, удовлетворяет следующему соотношению для нормального распределения:

$$\mu s = (s \quad 1)\sigma 2 \cdot \mu s \quad 2\,.$$

Эта формула является рекурсивной формулой, которая выражает моменты высших порядков как функцию моментов низших порядков.

Учитывая, что μ0 = 1 (момент нулевого порядка любой случайной величины равно 1, так как ожидаемое значение этой переменной , увеличенное в нулевой степени) и что μ1 = 0 (центрированный момент первого порядка равен нулю для любой случайной величины), имеем:

$$\mu 2 = \sigma 2$$

$$\text{м3} = 0$$

Machine Translated by Google

$$\mu 4 = 3\sigma 4$$

и т.п.

Тогда коэффициент эксцесса равен:

$$\beta 2 = \frac{\text{мк4}}{\sigma 4} \text{ "=" } \frac{3\sigma 4}{\sigma 4} = 3 .$$

ПРИМЕРЫ

Мы хотим рассчитать эксцесс распределения суточного оборота на 75 пекарен. Позволять рассчитаем коэффициент эксцесса β2 используя следующие данные:

Таблица с классификацией дневных оборотов 75 пекарни

Частота оборота	
215–235	4
235–255	6
255–275	13
275–295	22
295–315	15
315–335	6
335–355	5
355–375	4

Момент четвертого порядка выборки равен предоставлено:

$$\text{м4} = \frac{1}{\text{н}} \sum_{\text{я=1}}^{\text{час}} \text{fi}(\text{xi} - \bar{\text{x}})^4 ,$$

где n = 75, $\bar{x}$ = 290,60 , xi — центр интервала классов i. Мы можем обобщить расчеты в следующей таблице:

xi	xi Икс	я	фи (си Икс)4
225	65,60	4	74075629.16
245	45,60	6	25942428.06
265	25,60	13	5583457,48
285	5,60	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

Поскольку S = 33,88, коэффициент эксцесса равно:

$$\beta 2 = \frac{\frac{1}{75}(281020067,84)}{(33,88)4} = 2,84 .$$

Поскольку β2 меньше 3, мы можем заключить что распределение дневного оборота в 75 пекарен — это платикуртик, то есть более плоское, чем нормальное распределение.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Мера эксцесса
- Мера формы

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

См. коэффициент асимметрии β1 де Пирсона.

Коэффициент асимметрии

Коэффициент асимметрии измеряет асимметрия распределения. Он основан на понятие момента распределения . Этот коэффициент является одним из показателей асимметрия.

ИСТОРИЯ

Между концом девятнадцатого века ри и к началу двадцатого века Пирсон, Карл изучил большие наборы данные , которые иногда значительно отклонялись от нормы и демонстрировали значительную асимметрию.

Он впервые использовал следующий коэффициент как мера асимметрии:

$$\text{асимметрия} = \frac{\bar{\text{x}} - \text{мода}}{\text{С}} ,$$

где $\bar{x}$ представляет собой среднее арифметическое и S стандартное отклонение.

Эта мера равна нулю, если данные распределены симметрично.

Machine Translated by Google

Эмпирически он обнаружил, что при умеренно асимметричном распределении (гамма распределение):

$$\text{Mo} \quad \bar{x} \quad 3 \cdot (\text{Md} \quad \bar{x}) ,$$

где Mo и Md обозначают моду, а медиана набора данных. Заменив это выражение в предыдущий коэффициент, получается следующая альтернативная формула:

$$\text{асимметрия} = \frac{3 \cdot (\bar{x} \quad \text{Md})}{\text{C}} .$$

После этого Пирсон К. (1894, 1895) введен коэффициент асимметрии, известный как коэффициент $\beta 1$, на основании расчетов центральных моментов. Этот коэффициент вычислить сложнее, но больше описательным и лучше приспособленным к большому числу наблюдений.

Пирсон, К. также создал коэффициент эксцесс ($\beta 2$), который используется для измерения сплющенность кривой. Этот коэффициент также исходя из моментов распределения изучается.

Таблицы предельных значений коэффициентов $\beta 1$ и $\beta 2$ можно найти в работах Пирсона и Хартли (1966, 1972). Если выборка оценивает падение за предел для $\beta 1$, $\beta 2$, заключаем, что популяция значительно изогнуты или перекошены.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Коэффициент асимметрии основан на cen ный момент третьего порядка распределения в вопросе, который равен:

$$\mu 3 = E (X \quad \mu)3 .$$

Чтобы получить коэффициент асимметрии, равный независимо от единицы измерения, момент третьего порядка делится на стандартный отклонение населенности σ , возведенное в третья власть. Полученный коэффициент, обозначенный $\overline{\beta 1}$, равен:

$$\overline{\beta 1} = \frac{\text{мк3}}{\sigma 3} .$$

Оценка этого коэффициента, рассчитанная для выборки (x1, x2, ... xn), обозначается $\overline{b1}$. Он равен:

$$\overline{б1} = \frac{\text{м3}}{\text{S3}} ,$$

где m3 — центрированный элемент третьего порядка момент образца, определяемый по формуле:

$$\text{м3} = \frac{1}{\text{н}} \cdot \sum_{\text{я}=1}^{\text{н}} (\text{xi} \quad \bar{x})^3 .$$

Здесь $\bar{x}$ — среднее арифметическое, n — сумма число наблюдений , S3 — стандартное отклонение в третьей степени.

Для случая, когда случайная величина X принимает значения xi с частотами fi, i = 1, 2,...,h, центральный момент третьего порядка выборки находится по формуле :

$$\text{м3} = \frac{1}{\text{н}} \cdot \sum_{\text{я}=1}^{\text{час}} \text{fi} \cdot (\text{xi} \quad \bar{x})^3 .$$

Если коэффициент положительный, распределение распространяется вправо. Если он отрицательный, то раздача расширяется влево. Ифитис рядом нулю, распределение приблизительно симметрично.

Если выборка берется из нормальной совокупности, статистика $\overline{b1}$ примерно соответствует нормальное распределение со средним значением 0 и стандартное отклонение $\frac{6}{\text{н}}$. Если размер выборка n больше 150, нормальная таблица может быть использована для проверки асимметрии гипотеза.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Этот коэффициент (как и другие меры асимметрии) представляет интерес только в том случае, если он

Machine Translated by Google

можно использовать для сравнения форм двух или больше раздач.

ПРИМЕРЫ

Предположим, что мы хотим сравнить формы полученных распределений дневного оборота для 75 пекарен за два разных года. Мы затем рассчитайте коэффициент асимметрии в оба случая.

Данные классифицированы в таблице ниже :

Частота оборота	на 1 год	Частоты на 2 год
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

Момент третьего порядка образца равен предоставлено:

$$м3 = \frac{1}{н} \cdot \sum_{я=1}^{час} fi \cdot (xi - \bar{x})^3 .$$

Для года 1 n = 75, $\bar{x}$ = 290,60 и xi центр каждого интервала класса i. Расчет сведены в следующую таблицу:

xi	xi − Икс	я	фи (си − Икс)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

Поскольку S = 33,88, коэффициент асимметрии равно:

$$\overline{б1} = \frac{\frac{1}{75}(821222,41)}{(33,88)3} = \frac{10949,632}{38889.307}$$

$$= 0,282 .$$

Для года 2 n = 75 и $\bar{x}$ = 265,27. расчеты резюмируются в следующем стол:

xi	xi − Икс	я	фи (си − Икс)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

Поскольку S = 42,01, коэффициент асимметрии равно:

$$\overline{б1} = \frac{\frac{1}{75}(4750082,83)}{(42.01)3} = \frac{63334.438}{74140.933}$$

$$= 0,854 .$$

Коэффициент асимметрии для года 1 составляет близко к нулю ($\overline{b1}$ = 0,282), поэтому дневной распределение оборота для 75 пекарен на 1-й год очень близок к симметричному распределение. Для 2-го года коэффициент асимметрии выше; это означает, что распределение распространяется вправо.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Мера формы
- Мера асимметрии

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Пирсон, Э.С., Хартли, Х.О.: Биометрика
Таблицы для статистиков, тт. I и II.
Издательство Кембриджского университета, Кембридж
(1966,1972)

Пирсон, К.: Вклад в математическую теорию
эволюции. И. В: Карл
Ранние статистические статьи Пирсона.
Издательство Кэмбриджского университета, Кембридж,
стр. 1–40 (1948). Впервые опубликовано в 1894 г.
как: О рассечении асимметричных частотных
кривых. Филос. Транс. Рой. соц.
Лонд. сер. А 185, 71–110

Пирсон, К.: Вклад в математическую теорию
эволюции. II: перекос
изменение однородного материала. В:
Ранние статистические статьи Карла Пирсона.
Издательство Кембриджского университета, Кембридж,
стр. 41–112 (1948). Впервые опубликовано в
1895 г. в Филосе. Транс. Рой. соц. Лонд.
сер. А 186, 343–414

Коэффициент вариации

Коэффициент вариации является мерой
относительная дисперсия. Он описывает стандартное
отклонение в процентах от среднего арифметического.

Этот коэффициент можно использовать для сравнения
дисперсии количественных переменных , которые
не выражаются в одних и тех же единицах (например,
при сравнении заработной платы в разных странах,
приведенной в разных валютах),
или дисперсии переменных, которые имеют очень
разные средства.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Коэффициент вариации CV определяется как
отношение стандартного отклонения к
среднее арифметическое для набора наблюдений;
другими словами:

$$\text{Резюме} = \frac{C}{\bar{x}} \cdot 100$$

для образца, где:

S - стандартное отклонение выборки,
и

$\bar{x}$ среднее арифметическое выборки,

или:

$$\text{Резюме} = \frac{\sigma}{\mu} \cdot 100$$

для населения, где

σ – стандартное отклонение генеральной
совокупности,

μ - среднее значение населения.

Этот коэффициент не зависит от единицы
измерение, используемое для переменной.

ПРИМЕРЫ

Изучим распределение заработной платы на двоих.
компаний из двух разных стран.
Согласно опросу , арифметика
средние значения и стандартные отклонения
заработная плата следующая:

Компания А:

$\bar{x} = 2500$ швейцарских франков ,

$S = 200$ швейцарских франков ,

$$CV = \frac{200}{2500} \cdot 100 = 8\%$$

Компания Б:

$\bar{x} = 1000$ швейцарских франков ,

$S = 50$ швейцарских франков ,

$$CV = \frac{50}{1000} \cdot 100 = 5\%$$

Стандартное отклонение составляет 8% от
среднее арифметическое для компании А, и
5% для компании В. Распределение заработной платы
немного более однородна для компании В
чем для компании А.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Среднее арифметическое

Мера рассеивания

Среднеквадратичное отклонение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Джонсон, Н.Л., Леоне, Ф.К.: Статистика и планирование экспериментов в инженерных и физических науках. Уайли, Нью-Йорк (1964)

Коллинеарность

Известно, что переменные математически коллинеарны, если одна из них является линейной величиной.

бинирование остальных переменных. Они известны как статистически коллинеарные, если одна из них представляет собой приблизительно линейную комбинацию других переменных. В случае регрессионной модели, в которой объясняющие переменные сильно коррелируют друг с другом, мы говорим, что существует коллинеарность (или мультиколлинеарность) между объясняющими переменными. В первом случае определить оценки методом наименьших квадратов просто невозможно, а во втором случае эти оценки могут иметь значительную дисперсию.

ИСТОРИЯ

Термин «коллинеарность» впервые был использован в математике в начале двадцатого века в связи с переоткрытием теоремы Паппа Александрийского (математик третьего века). Пусть A, B, C — три точки на прямой, а A', B', C' — три точки на другой прямой. Если мы свяжем пары, используя AB'.A'B, CA'.AC' и BC'.B'C, их пересечения произойдут в линии; другими словами, три точки пересечения будут лежать на одной прямой.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В случае матрицы объясняющих переменных X коллинеарность означает, что один из столбцов X является (приблизительно) линейной комбинацией других столбцов. Этот

следует, что XX почти сингулярна. Следовательно, оценка методом наименьших квадратов 1 β = XXXY получается обращением почти вырожденной матрицы, что приводит к неустойчивости ее компонент. Техника гребневой регрессии была создана для решения этих проблем коллинеарности.

Коллинеарная связь между более чем двумя переменные не всегда будут результатом наблюдения парных корреляций между переменными. Лучшее указание на наличие проблемы коллинеарности дают коэффициенты инфляции дисперсии, VIF. Коэффициент инфляции дисперсии независимой переменной Xj определяется следующим образом:

$$VIFj = \frac{1}{1 - R2\,j},$$

где R2 – коэффициент детерминации модели _

$$Xj = \beta 0 + \sum_{k=j} \beta k Xk + \varepsilon .$$

Коэффициент VIF принимает значения от 1 и . Если Xj математически коллинеарны другим переменным, мы получаем R2 = 1 и VIFj = . С другой стороны, если Xj взаимно независимы, мы имеем R2 = 0 и VIFj = 1. На практике мы считаем, что существует реальная проблема с коллинеарностью, когда VIFj больше 100, что соответствует R2 больше, чем 0,99. Дж

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Обращение сингулярной матрицы, подобное обращению 0, недопустимая операция. Используя

Machine Translated by Google

тот же принцип, обращение почти сингулярной матрицы аналогично обращению очень маленькой числа. Некоторые элементы матрицы поэтому должен быть очень большим. Следовательно, когда объясняющие переменные коллинеарны, некоторые элементы матрицы (X X) 1 из β вероятно, будет очень большим. Вот почему коллинеарность приводит к неустойчивым регрессионным оценкам. Помимо этой проблемы, коллинеарность также приводит к проблеме вычислений; это трудно точно вычислить обратную почти вырожденная матрица.

ПРИМЕРЫ

Исследовано тринадцать порций цемента. в следующем примере. Каждая порция содержит четыре ингредиента, как описано в стол. Цель эксперимента – определить, как величины X1, X2, X3 и X4, соответствующие количествам этих четырех ингредиенты, влияют на количество тепла Y , выделяемого при затвердевании цемента.

Yi количество теплоты, отданное за время закалка i-й порции (в джоулях);

Xi1 количество ингредиента 1 (трехкальциевого алюмината) в i-й порции;

Xi2 количество ингредиента 2 (трехкальциевого силиката) в i-й порции;

Xi3 количество ингредиента 3 (тетракальцийалюмоферрит) в i-й порции;

Xi4 количество ингредиента 4 (двухкальциевый силикат) в i-й порции.

Таблица: Тепло, отдаваемое цементными порциями во время твердения

Порция	Энгр клиент	Ингредиент	Энгр клиент	Нагревать
я	1 Икс1	2 Икс2	3 Икс3	4 Икс4	Д
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

Источник: Биркес и Додж (1993).

Начнем с простой линейной регрессии. Модель, используемая для линейной регрессии

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + e .$$

Получаем следующие результаты:

Переменная	β^я	SD	тк
Константа 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	0,1441	0,7091	0,20

Мы видим, что только коэффициент X1 значительно больше нуля; в остальных случаях tc < t(α/2,n 2) (значение взято из таблица) для уровня значимости α = 0,1. При этом стандартные отклонения расчетных коэффициентов β 2, β 3 и β 4 больше затем сами коэффициенты. Когда

Machine Translated by Google

переменные сильно коррелированы, известно что эффект одного может маскировать эффект другой. Поэтому коэффициенты могут по-видимому, незначительно отличается от нуль.

Проверить наличие мультиколлинеарности для пары переменных мы вычисляем матрицу корреляции.

	Икс1	Икс2	Икс3	Икс4
X1	1	0,229	0,824	0,245
X2	0,229	1	0,139	0,973
X3	-0,824	-0,139	1	0,030
X4	0,245	0,973	0,030	1

Отметим, что сильная отрицательная корреляция (-0,973) существует между X2 и X4. Глядя на данные, мы можем увидеть причину что. За исключением частей с 1 по 5, общая количество силикатов (X2 + X4) почти постоянен по порциям и составляет примерно 77; следовательно, X4 приблизительно 77 - X2. эта ситуация не позволяет нам различать отдельные эффекты X2 и X4 . Например, мы видим что четыре наибольших значения X4 (60, 52, 47 и 44) соответствуют значениям Y меньшим тогда средний жар 95,4. Поэтому сначала На первый взгляд кажется, что большое количество ингредиента 4 приведет к небольшому количеству тепла.

Однако отметим также, что четыре крупнейших значения X4 соответствуют четырем наименьшим значениям X2, что дает отрицательную корреляцию между X2 и X4. Это говорит о том, что ингредиент 4, взятый отдельно, имеет небольшой эффект на переменную Y, и что малые количества 2, взятые сами по себе, могут объяснить малое количество излучаемого тепла. Следовательно, линейная зависимость между двумя объясняющими переменными (X2 и X4) усложняет определение влияния каждой переменной в отдельности на переменная отклика Y.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Множественная линейная регрессия
- Регрессия хребта

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Биркес Д., Додж Ю.: Альтернативные методы регрессии. Уайли, Нью-Йорк (1993)

Бок, Р.Д.: Многомерные статистические методы в поведенческих исследованиях. Макгроу Хилл, Нью-Йорк (1975)

Комбинация

Комбинация - это неупорядоченный набор уникальных элементов или объектов.
k -комбинация — это подмножество из k элементов.
Количество k-комбинаций из набора n элементов - количество аранжировок.

ИСТОРИЯ

См. комбинаторный анализ.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

1. Комбинация без повторения

   Комбинация без повторения описания ситуация, когда каждый нарисованный объект не помещается обратно для следующего рисунка.Каждый поэтому объект может встречаться только один раз в каждая группа.

   Количество бесповторных комбинаций k предметов из n определяется по формуле:

$$C_n^k \text{ "=" } \binom{n}{k} \text{ "=" } \frac{n!}{k! \cdot (n-k)!}.$$

2. Комбинация с повторениями

   Сочетание с повторениями (или с денежный перевод) используются, когда каждый нарисованный объект помещается обратно для следующего рисования. Затем каждый объект может появляться несколько раз в каждой группы, r = 0, 1,...,k.

Machine Translated by Google

Количество комбинаций с повторением k объектов среди n определяется по формуле:

$$K_{k_n} = \binom{n + k - 1}{k} = \frac{(n + k - 1)!}{k! \cdot (n - 1)!}.$$

ПРИМЕРЫ

1. Комбинации без повторений

Рассмотрим ситуацию, когда мы должны выбрать комиссию из трех человек из собрание из восьми человек. Сколько различные комитеты потенциально могут быть выбирается из этой сборки, если каждый человек можно выбрать только один раз в каждой группе? Здесь нам нужно вычислить число возможных комбинаций трех человек из восьми:

$$C_{k_n} = \frac{n!}{k! \cdot (n - k)!} = \frac{8!}{3! \cdot (8 - 3)!} = \frac{40320}{6 \cdot 120} = 56$$

Следовательно, из собрания из восьми человек можно сформировать 56 различных комитетов по три человека.

2. Комбинации с повторениями

урна с шестью пронумерованными шарами. Мы выполнить четыре последовательных розыгрыша, и поместите вытащенный шар обратно в урну после каждый рисунок. Сколько различных комбинаций может получиться из этого рисунка? В этом случае мы хотим найти число комбинаций с повторением (поскольку каждый выпавший шар возвращается в урну перед следующим рисунком). Мы получаем

$$K_{k_n} = \frac{(n + k - 1)!}{k! \cdot (n - 1)!} = \frac{9!}{4! \cdot (6 - 1)!} = \frac{362880}{24 \cdot 120} = 126$$

различные комбинации.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Договоренность
Биномиальное распределение
Комбинационный анализ
Перестановка

C

Комбинационный анализ

Комбинационный анализ относится к группе методики, которые можно использовать для определения количество элементов в конкретном наборе без необходимости подсчитывать их один за другим. Рассматриваемые элементы могут быть результатом из научного эксперимента или другого Возможные последствия случайного события. Три конкретных понятия важны в комбинаторный анализ:

- Перестановки;
- Комбинации;
- Договоренности.

ИСТОРИЯ

Комбинационный анализ веками интересовал математиков. По словам Такача (1982), такой анализ восходит к древним Греция. Однако индусы, персы (включая поэта и математика Хайям, Омар) и (особенно) китайцы также изучали такие проблемы. 3000 летняя китайская книга «И Цзин» описывает возможные расположения множества n элементов, где n 6. В 1303 г. Чу Ши Цзе опубликовал труд под названием «Су Юань Юй Цзянь» (Драгоценное зеркало четырех элементы). На обложке книги изображен треугольник, который показывает комбинации k элементов, взятых из набора размера n , где 0 k n. Этот арифметический треугольник был также исследовали несколько европейских математиков, таких как Штифель, Тарталья и Эригон, и особенно Паскаль, написавший «Трактат».

Machine Translated by Google

te du треугольник arithmétique» (Трактат о арифметический треугольник) в 1654 году (хотя это было не публиковался до его смерти в 1665 г.). Другой документ о комбинациях был опубликовано в 1617 году Путеаном, Эрикием под названием «Erycii Puteani Pretatis Thaumata в Bernardi Bauhusii и Societate Jesu Pro teum Parthenium». Тем не менее, комбинаторный анализ только показал свою истинную силу с работы Ферма (1601–1665) и Паскаля (1623–1662). Термин «комбинаторный анализ» был введен Лейбницем (1646–1646 гг.). 1716) в 1666 г. В своем труде «Dissertatio de Arte Combinatoria», он систематически изучал проблемы, связанные с аранжировками, перестановками и комбинациями.

Следует упомянуть и другие работы в этой области здесь, такие как Уоллис, Дж. (1616–1703), сообщается в «Доктрине перестановок». и комбинации» (существенная и фундаментальная часть «Учения о шансах»), ортоз Бернулли Ж., Муавра А. де, Кардано Г. (1501–1576) и Галилея (1564–1576). 1642 г.).

Во второй половине девятнадцатого века Ри, Кейли (1829–1895) решил некоторые проблемы, связанные с этим типом анализа, с помощью графики, которую он назвал «деревьями». Наконец, мы должны упомянем также важную работу Мак-Магона (1854–1929) «Комбинационный анализ». (1915–1916).

ПРИМЕРЫ

См. расположение, комбинацию и мутацию.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Арифметический треугольник
Договоренность
Биномиальный
Биномиальное распределение
Комбинация
Перестановка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

МакМахон, Пенсильвания: комбинаторный анализ, тт. I и II. Издательство Кембриджского университета, Кембридж (1915–1916)

Стиглер, С.: История статистики, Измерение неопределенности до 1900. Белкнап, Лондон (1986).

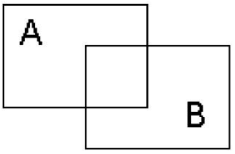
Takács, L .: Комбинаторика. В: Коц, С., Джонсон, Н.Л. (ред.) Энциклопедия статистических наук, том. 2. Уайли, Нью-Йорк (1982)

Совместимость

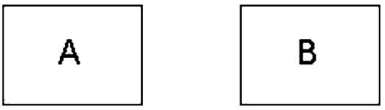
Два события называются совместимыми, если наступление первого события не препятствует наступлению второго (или иного слова, если пересечение между двумя события не нулевые):

$$P(A \quad B) = 0 .$$

Мы можем представить два совместимых события A и B схематично следующим образом:



Два события A и B несовместимы (или взаимоисключающие), если появление A предотвращает появление Б или наоборот. Мы можем представить это следующим образом:



Machine Translated by Google

Это означает, что вероятность того, что эти два события происходят одновременно, равен нулю:

$$A \quad B = \varphi \qquad P(A \quad B) = 0 .$$

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если два события A и B совместимы, то вероятность того, что хотя бы одно из событий происходит, можно получить, используя следующие формула:

$$P(A \quad B) = P(A) + P(B) \quad P(A \quad B) .$$

С другой стороны, если два события A и B несовместимы, вероятность того, что два события A и B происходят одновременно равен нулю:

$$P(A \quad B) = 0 .$$

Вероятность того, что хотя бы одно из событий можно получить, просто добавив индивидуальные вероятности A и B:

$$P(A \quad B) = P(A) + P(B) .$$

ПРИМЕРЫ

Рассмотрим случайный эксперимент , который включает в себя вытягивание карты из колоды 52 карты. Нас интересуют три мычащие события:

A = «нарисовать сердце»

B = «нарисуй королеву»

C = «нарисовать трефу» .

Вероятности , связанные с каждым из эти события :

$$P(A) = 52 \quad \frac{13}{}$$

$$P(B) = 52 \quad \frac{4}{}$$

$$P(C) = 52 \quad \frac{13}{} .$$

События A и B совместимы, так как можно нарисовать и сердце и королеву одновременно (червовая дама). Следовательно, пересечение между точками A и B является Королева сердец. Вероятность этого события дан кем-то:

$$P(A \quad B) = 52 \quad \frac{1}{} .$$

Вероятность союза двух события A и Б (нарисовать либо сердце, либо ферзь) тогда равно:

$$P(A \quad B) = P(A) + P(B) \quad P(A \quad B)$$

$$\text{"="} \frac{13}{52} + \frac{4}{52} \quad \frac{1}{52}$$

$$\text{"="} \frac{}{13} .$$

С другой стороны, события A и C несовместимы, потому что карта не может быть и сердце и клуб! Перекресток _ между A и C — пустое множество.

$$A \quad C \text{ знак равно } \varphi .$$

Вероятность союза двух события A и C (вытягивание сердца или трефы) просто дается суммой вероятностей каждого события:

$$P(A \quad C) = P(A) + P(C)$$

$$\text{"="} \frac{13}{52} + \frac{13}{52}$$

$$\text{"="} \frac{1}{2} .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Событие

Независимость

Вероятность

Machine Translated by Google

Дополнительный

Рассмотрим выборочное пространство для случайного эксперимента.

Для любого события A элемент определяет , мы можем новое событие B , содержащее все

элементы выборочного пространства , не входящие в A.

Это событие B называется дополнением к A по отношению к и получается отрицанием A.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим событие A, которое является элементом выборочного пространства .

Комплемент A относительно обозначенного A^{-}.

Это дается отрицанием A:

$$A^{-} = - A$$

$$= \{w \quad ; w \quad / A\} .$$

ПРИМЕРЫ

Рассмотрим случайный эксперимент , состоящий в трехкратном подбрасывании монеты.

Выборочное пространство этого эксперимента

$$= \{TTT, TTH, TTT, TTX,$$

$$HTT, HTH, HHT, HHH\} .$$

Рассмотрим событие

$$A = \text{«Голова (H) встречается дважды»}$$

$$= \{THH, HTH, HHT\} .$$

Дополнение A относительно равного

$$A^{-} = \{TTT, TTH, TTT, XTT, ЧЧЧ\}$$

$$= \text{«Голова (H) не встречается дважды»} .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Событие
- Случайный эксперимент
- Образец пространства

Полный метод связывания

Метод полной связи — это метод иерархической классификации , в котором расстояние

между двумя классами определяется как наибольшее расстояние, которое может быть получено, если мы выберем по одному элементу из каждого класса и меры

расстояние между этими элементами. В др.

Другими словами, это расстояние между наиболее удаленными элементами каждого класса.

Например, расстояние , используемое для построения таблицы расстояний, является евклидовым расстоянием.

Используя метод полной связи, расстояние между двумя классами определяется

Евклидово расстояние между самыми дальними элементов (максимальное расстояние).:

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

См. кластерный анализ.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Классификация
- Кластерный анализ
- Дендрограмма
- Расстояние
- Таблица расстояний

Полностью

Рандомизированный дизайн

Полностью рандомизированный план — это тип экспериментального плана, в котором экспериментальные единицы случайным образом распределяются по разным лечения.

Он используется, когда экспериментальные единицы считаются «однородными»; то есть, когда в эксперименте нет неконтролируемого фактора .

ИСТОРИЯ

См. план экспериментов.

Machine Translated by Google

ПРИМЕРЫ

Мы хотим протестировать пять различных препаратов на основе аспирина. Для этого мы случайным образом распределяем пять видов лекарств среди 40 пациентов. Обозначая пять лекарств через A, B, C, D и E, мы получаем следующее случайное распределение:

A приписывается 10 людям;

B приписывается 12 лицам;

C приписывается 4 человекам;

D приписывается 7 лицам;

E приписывается 7 людям.

Таким образом, мы имеем полностью рандомизированный план, в котором методы лечения (лекарства) случайным образом распределяются между экспериментальными единицами (пациентами), и каждый пациент получает только одно лечение. Мы также предполагаем, что пациенты «однородны»: нет различий между ними. Более того, мы предполагаем, что не существует неконтролируемого фактора , влияющего на вены во время лечения.

В этом примере полностью рандомизированный план представляет собой факторный эксперимент , в котором используется только один фактор: аспирин. Пять видов аспирина представляют собой разные уровни фактора.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- План экспериментов
- Эксперимент

Композитный Индекс

Составные индексы позволяют измерять с помощью одного числа относительные изменения внутри группы переменных при переходе от одной ситуации к другой.
индекс потребительских цен, индекс оптовых цен, индекс занятости и Индекс Доу-Джонса — все это примеры составных индексов.

Цель использования составных индексов состоит в том, чтобы суммировать все простые индексы, содержащиеся в комплексном числе (значение, образованное из набора простых значений), в одном индексе.

Наиболее часто используемые составные индексы:

- индекс Ласпейреса ;
- индекс Пааше ;
- Индекс Фишера.

ИСТОРИЯ

См. номер индекса.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Существует несколько методов создания составных индексов.

Чтобы проиллюстрировать эти методы, воспользуемся сценарием, в котором индекс цен определяется для текущего периода n по отношению к базовому периоду 0.

1. Порядковый номер среднего арифметического (метод суммы):

$$B/0 = \frac{Пн}{P0} \cdot 100 \quad ,$$

где Pn — сумма цен товаров в текущем периоде, а P0 — сумма цен товаров в отчетном периоде.

2. Среднее арифметическое простого индекса num берс:

$$B/0 = \frac{1}{H} \cdot \frac{Пн}{P0} \cdot 100 \quad ,$$

где N — количество рассматриваемых товаров, а Pn — простой порядковый номер P0 каждой позиции.

В этих двух методах каждый элемент имеет одинаковое значение. Это ситуация

Machine Translated by Google

что часто не соответствует действительности.

3. Номер индекса взвешенной арифметики означает (метод взвешенной суммы):
Общая формула для порядкового номера рассчитано методом взвешенной суммы следующее:

$$B/0 = \frac{Пн \cdot Q}{P0 \cdot Q} \cdot 100 .$$

Выбор количества Q для каждого товара здесь может оказаться проблематичным: Q должен быть одинаковым как для числителя, так и для знаменателя при расчете индекс цен.
В индексе Ласпейреса значение Q соответствующий отчетному году использовал. В индексе Пааше значение Q используется текущий год. Другие статистики предлагали использовать значение Q на данный год.

ПРИМЕРЫ

Рассмотрим следующую таблицу, указывающую (фиктивные) цены трех потребительских товаров в отчетном году (1970) и их текущие цены.

	Цена (франков)	
Товары	1970 (0)	Сейчас ($Пн$)
Молоко	0,20	1,20
Хлеб	0,15	1.10
Масло	0,50	2.00

Используя эти числа, мы теперь исследуем три основных метода построения составных индексов.

1. Индекс среднего арифметического метод суммы):

$$B/0 = \frac{Пн}{P0} \cdot 100$$

$$"=" \frac{4.30}{0,85} \cdot 100 = 505,9 .$$

По этому методу индекс цен увеличилась на 405,9% (505,9 100) между отчетным годом и текущим год.

2. Среднее арифметическое простого индекса num берс:

$$B/0 = \frac{1}{H} \cdot \frac{Пн}{P0} \cdot 100$$

$$"=" \frac{1}{3} \cdot \frac{1,20}{0,20} + \frac{1.10}{0,15} + \frac{2.00}{0,50} \cdot 100$$

$$= 577,8 .$$

Этот метод дает несколько иное результат предыдущего, поскольку мы получить увеличение на 477,8% (577,8 - 100) в индексе цен между отчетным и текущим годами.

3. Номер индекса взвешенной арифметики означает (метод взвешенной суммы):

$$B/0 = \frac{Пн \cdot Q}{P0 \cdot Q} \cdot 100 .$$

Этот метод используется совместно с индекс Ласпейреса или индекс Пааше индекс.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Индекс Фишера
- Индекс
- Индекс Ласпейреса
- Индекс Пааше
- Простой порядковый номер

Условная возможность

Вероятность события при условии, что другое известно, что событие произошло.

Условная вероятность обозначается $P(A|B)$, который читается как «вероятность A, обусловленного B ».

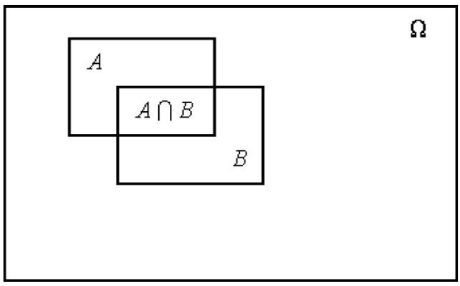
Machine Translated by Google

ИСТОРИЯ

Доминировала концепция независимости. теории вероятностей до начала двадцатый век. В 1933 году Колмогоров, Андрей Николаевич ввел понятие условной вероятности; эта концепция сейчас играет важную роль в теоретической и прикладная вероятность и статистика.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим случайный эксперимент , для которого мы знаем пространство выборки . Рассмотрим два события A и B из этого пространства. Вероятность A , P(A) зависит от множества возможных события в эксперименте ().



Теперь считаем, что у нас есть дополнительный информация об эксперименте: произошло событие Б. Вероятность происходящее событие A будет функцией пространства Брат, чем функция. Вероятность обусловленного B рассчитывается следующим образом:

$$P(A|B) = \frac{P(A \quad B)}{П(Б)} .$$

Если A и B — два несовместимых события, пересечение между A и B является

пустое место. Тогда мы будем иметь P(A|B) = P(B|A) = 0.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Понятие условной вероятности является одним из из наиболее важных по вероятности ори. Эта важность в основном обусловлена следующими моментами:

1. Нас часто интересует вычисление вероятность события , когда некоторая информация о результате уже известна. В этом случае искомая вероятность равна условная вероятность.
2. Даже при частичной информации о результат неизвестен, условные вероятности могут быть полезны при расчете нужные вероятности.

ПРИМЕРЫ

Рассмотрим группу из 100 автомобилей, распределенных по двум критериям комфорт и скорость.
Мы будем делать следующие различия:

машина может быть быстрый / медленный ,

машина может быть комфортный / неудобный .

Разделение 100 автомобилей на основе этих критериев представлено в следующей таблице:

	быстрый	медленный	общий
комфортный	40	10	50
неудобный	20	30	50
общий	60	40	100

Рассмотрим следующие события:

A = «выбрана быстрая машина»

и B = «выбран комфортабельный автомобиль».

Вероятность этих двух событий:

P(A) = 0,6 ,

P(B) = 0,5 .

Тогда вероятность выбора быстрой машины равна 0,6, а выбор комфортного автомобиля составляет 0,5.

Machine Translated by Google

Теперь представьте, что нам дали дополнительную информацию: была выбрана быстрая машина. Какова тогда вероятность того, что этот автомобиль тоже удобно?

Мы вычисляем вероятность того, что B знает что произошло A, или условная вероятность B, зависящая от A:
Находим в таблице, что

$$P(A \cap B) = 0{,}4 .$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0{,}4}{0{,}6} = 0{,}667 .$$

Вероятность того, что автомобиль комфортный, учитывая, что мы знаем, что это быстро, есть ли 2 перед 3.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Событие
- Вероятность
- Случайный эксперимент
- Образец пространства

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Колмогоров, А.Н.: Grundbegriffe der Wahrscheinlichkeitsrechnung. Спрингер, Берлин Гейдельберг Нью-Йорк (1933)

Колмогоров А.Н. Основы теории вероятностей. Издательство Челси Компания, Нью-Йорк (1956)

Доверительный интервал

Доверительный интервал — это любой интервал , построенный вокруг оценки , которая с определенной вероятностью содержит истинное значение. значение соответствующего параметра _ население .

ИСТОРИЯ

Согласно Desrosières, A. (1988), Боули, А. был одним из первых, кто заинтересовался понятие доверительного интервала. Боули представил свои первые достоверные интервальные расчеты Королевскому статистическому обществу в 1906 г.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Для построения доверительного интервала содержащее истинное значение параметра θ с заданной вероятностью, уравнение необходимо решить в следующем виде:

$$P(Li \le \theta \le Ls) = 1 - \alpha ,$$

где

θ оцениваемый параметр,
Li нижняя граница интервала ,
Ls— верхняя граница интервала, а
1 − α – заданная вероятность, называемая уровень достоверности интервала.

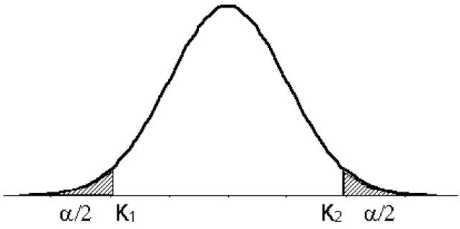
Вероятностьα измеряет риск ошибки интервал, означающий вероятность того, что интервал не содержит истинного значения параметр θ.
Для решения этого уравнения используется функция f(t, θ) должна быть определена, где t — оценка θ, для которой известно распределение вероятностей .

Определение этого интервала для f(t, θ) включает написание уравнения:

$$P(k1 \le f(t, \theta) \le k2) = 1 - \alpha ,$$

где константы k1 и k2 задаются формулой распределение вероятности функции f(t, θ). В общем случае риск ошибки α делится на две равные части при распределении по $\frac{\alpha}{2}$ каждой стороне распределения f(t, θ). Если, для Например, функция f(t, θ) следует центрированному и редуцированному нормальному распределению , константы sk1 и k2 будут симметричными и могут быть представлено $z_{\frac{\alpha}{2}}$ и $+z_{\frac{\alpha}{2}}$, как показано на следующем рисунке.

Machine Translated by Google



После того как константы k1 и k2 найдены, необходимо выделить параметр θ в приведенном выше уравнении. Таким образом находится доверительный интервал θ для доверительной вероятности 1 α.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Следует быть очень осторожным при интерпретации доверительного интервала. Если для доверительной вероятности 95 % найти доверительный интервал для среднего значения μ, где нижняя и верхняя границы равны k1 и k2 соответственно, то можно сделать следующий вывод (например): «На основании изученных выборки, мы можем утверждать, что среднее значение генеральной совокупности может быть найдено в установленном интервале ». Было бы неверным делать вывод о том, что это 95% шанс найти среднее значение населения в интервале. Действительно, поскольку μ и пределы sk1 и k2 интервала являются константами, интервал может содержать или не содержать μ. Однако если статистик имеет возможность повторить эксперимент (состоящий из взятия выборки из генеральной совокупности) несколько раз, то 95% полученных интервалов будут содержать истинное значение μ.

ПРИМЕРЫ

Предприятие, производящее лампочки, хочет проверить средний срок службы своих лампочек. Распределение случайной величины X, которая представляет продолжительность жизни в часах, представляет собой нормальное распределение со средним значением μ и стандартным отклонением σ = 30.

Чтобы оценить μ, предприятие сжигает n = 25 лампочек.

Он получает средний срок службы $\bar{x} = 860$ часов. Он хочет установить доверие интервал вокруг оценки $\bar{x}$ при конф. уровень плотности 0,95. Поэтому первым шагом является получение функции $f(t, \theta) = f(x, \bar{\mu})$ для известного распределения. Здесь мы используем:

$$f(t, \theta) = f(x, \bar{\mu}) = \frac{\bar{x} \quad \mu}{\frac{}{\sigma n}},$$

который следует за центрированным и редуцированным ни неправильное распределение. Уравнение P(k1 f(t, θ) k2) = 1 α принимает вид:

$$P \quad z0{,}025 \quad \frac{\bar{x} \quad \mu}{\frac{}{\sigma n}} \quad z0{,}025 = 0{,}95,$$

потому что риск ошибки α был разделен на две равные части в $\frac{\alpha}{2} = 0{,}025$.
Стол для центрированных и уменьшенных ни mal распределение, нормальная таблица, дает z0,025 = 1,96. Поэтому:

$$P \ -1{,}96 \quad \frac{\bar{x} \quad \mu}{\frac{}{\sigma n}} \quad 1{,}96 = 0{,}95.$$

Чтобы получить доверительный интервал для μ при доверительном уровне 0,95, μ необходимо выделить в приведенном выше уравнении:

$$P \ -1{,}96 \quad \frac{\bar{x} \quad \mu}{\frac{}{\sigma n}} \quad 1{,}96 = 0{,}95$$

$$P \quad 1{,}96 \quad \frac{o}{\bar{x} = \mu} \quad 1{,}96 \quad \frac{o}{n \quad n} = 0{,}95$$

$$P\bar{x} \quad 1{,}96 \quad \frac{o}{\mu = \bar{x}} + 1{,}96 \quad \frac{o}{n \quad n} = 0{,}95.$$

Machine Translated by Google

Заменив $\bar{x}$, σ и n их соответствующими значениями, мы получим:

$$P\ 860 \quad 1{,}96 \quad \mu \quad \frac{30}{860} \quad 25$$

$$+ 1{,}96 \quad \frac{30}{25} \quad = 0{,}95$$

$$P(848{,}24 \quad \mu \quad 871{,}76) = 0{,}95 .$$

Таким образом, доверительный интервал для μ при доверительном уровне 0,95 составляет:

[848,24, 871,76].

Это означает, что мы можем с вероятностью 95% утверждать , что в этом интервале находится истинное значение параметра μ, соответствующее среднему сроку службы лампочек .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Уровень достоверности
- Оценка
- Оценщик

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Боули, Алабама: Обращение президента к экономическому отделу Британской Ассоциации ссылка Дж. Рой. Стат. соц. 69, 540–558 (1906)

Desrosières, A. La partie pour le tout: com ment généraliser? La préhistoire delacon trainte de représentativité. Оценка и зондаж. Cinq взносы à l'histoire de la statistique. Экономика, Париж (1988)

Уровень достоверности

Доверительный уровень – это вероятность того, что доверительный интервал , построенный вокруг оценщик содержит истинное значение _ соответствующий параметр населения .

Обозначим уровень достоверности (1 α), где α соответствует риску ошибки; то есть к вероятности того, что доверительный интервал не содержит истинного значения

параметр.

ИСТОРИЯ

Первый пример доверительного интервала появляется в работе Лапласа (1812 г.).
Согласно Desrosières, A. (1988), Bowley, AL, был одним из первых, кто заинтересовался концепцией доверительного интервала.

См. проверку гипотез.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть θ — параметр, связанный с популяцией. θ подлежит оценке, а T – его оценка.

матор из случайной выборки. Мы оцениваем точность T как оценки θ путем построения доверительного интервала вокруг оценки, который часто интерпретируется как допустимая погрешность.

Чтобы построить этот доверительный интервал, мы обычно поступаем следующим образом. По распределению оценки T определяется интервал, который, вероятно, содержит истинное значение параметра. Обозначим этот интервал через (T ε, T + ε), а вероятность истинного значения параметра нахождения в этом интервале как (1 α). Тогда мы можем сказать, что предел погрешности ε связан с α вероятностью:

$$P\ (T - \varepsilon \quad \theta \quad T + \varepsilon) = 1 - \alpha .$$

Уровень вероятности, связанный с интервалом оценки, называется достоверностью.

уровень или степень достоверности.
Интервал T ε θ T + ε называется доверительным интервалом для θ при доверительном интервале.

уровень 1 α. Возьмем, например, α = 5%,

Machine Translated by Google

что даст доверительный интервал параметра θ на уровне вероятности 95%.
Это означает, что если использовать T в качестве оценки θ, то указанный интервал в среднем будет содержать истинное значение параметра 95 раз из 100 отсчетов, а 5 раз не будет содержать его.

Величина ε доверительного интервала соответствует половине длины интервала.
Таким образом, этот параметр дает нам представление о погрешности оценки. Для заданного уровня достоверности 1-α, чем меньше доверительный интервал, тем эффективнее оценка.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Наиболее часто используемые уровни достоверности составляют 90%, 95% и 99%. Однако при необходимости вместо них можно использовать другие уровни.
Хотя мы хотели бы использовать самый высокий уровень достоверности, чтобы максимизировать вероятность того, что доверительный интервал содержит истинное значение параметра , но если мы увеличиваем доверительный интервал, то увеличивается и интервал . Поэтому то, что мы выигрываем с точки зрения уверенности, теряется с точки зрения точности, поэтому приходится искать компромисс.

ПРИМЕРЫ

Компания, производящая лампочки, хочет изучить средний срок службы своих ламп. Распределение случайной величины X , представляющей время жизни в часах, представляет собой нормальное распределение среднего значения μ и стандартного отклонения σ = 30.

Для оценки μ компания сжигает случайную выборку из n = 25 лампочек.

Компания получает средний срок службы лампы $\bar{x} = 860$ часов. Затем он хочет построить 95% доверительный интервал для μ вокруг оценки $\bar{x}$.

Стандартное отклонение σ населения известно; значение ε равно $z\alpha/2 \cdot \sigma\bar{X}$. Значение $z\alpha/2$ получается из нормальной таблицы и зависит от вероятности, приписываемой параметру α. Затем мы выводим доверительный интервал оценки μ на уровне вероятности 1 α:

$$\bar{X} \quad z\alpha/2\sigma\bar{X} \quad \mu \quad \bar{X} + z\alpha/2\sigma\bar{X}.$$

Из гипотезы о том , что срок службы лампы X следует нормальному распределению со средним значением μ и стандартным отклонением σ = 30, мы делаем вывод, что выражение

$$\frac{\bar{x} \quad \mu}{\sigma_n}$$

следует стандартному нормальному распределению.
Отсюда получаем:

$$P \quad z0{,}025 \quad \frac{\bar{x} \quad \mu}{\sigma_n} = 0{,}95 \quad ,$$

где риск ошибки α делится на два равные части $\frac{\alpha}{2} = 0{,}025$.
Таблица стандартного нормального распределения (обычная таблица) дает $z_{0{,}025} = 1{,}96$. Тогда мы имеем:

$$P \ -1{,}96 \quad \frac{\bar{x} \quad \mu}{\sigma_n} \ 1{,}96 = 0{,}95.$$

Чтобы получить доверительный интервал для μ, мы должны выделить μ в следующем уравнении с помощью следующих преобразований:

$$P \quad 1{,}96 \quad \frac{\bar{x} \quad \mu}{\sigma_n} \quad 1{,}96 = 0{,}95 \quad ,$$

$$P\bar{x} \quad 1{,}96 \quad \mu \quad \bar{x} + 1{,}96 \quad \frac{\sigma}{n} = 0{,}95 .$$

Machine Translated by Google

Подставив $\bar{x}$, σ и n вместо соответствующих значений, оцениваем доверительный интервал:

$$860 \quad 1{,}96 \cdot \frac{30}{\sqrt{25}}$$

$$\mu \quad 860 + 1{,}96 \cdot \frac{30}{\sqrt{25}}$$

$$848{,}24 \quad \mu \quad 871{,}76 .$$

Мы можем подтвердить с достоверностью 95% что этот интервал содержит истинное значение параметр μ , который соответствует средний срок службы лампы.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Доверительный интервал
- Оценка
- Оценщик
- Проверка гипотезы

Таблица сопряженности

Таблица непредвиденных обстоятельств представляет собой перекрестную таблицу, содержащую различные атрибуты совокупности или наблюдаемой выборки. Анализ непредвиденных обстоятельств состоит из обнаружения и изучения отношения (если они существуют) между этими атрибуты.

Таблица непредвиденных обстоятельств может быть двумерной таблицей с r строками и c столбцами, связанными друг с другом. к двум качественным категориальным переменным обладающие, соответственно, r и c категориями. Он также может быть многомерным, когда количество качественных переменных больше то два: если, например, элементы популяция или выборка характеризуются _ по трем атрибутам соответствующая контингентная таблица имеет размеры I×J×K, где I представляет количество категорий, определяющих первый атрибут, J количество категорий второго атрибута и K количество категории третьего атрибута.

ИСТОРИЯ

Термин «непредвиденные обстоятельства», используемый в отношении скрещенная таблица категориальных данных, кажется возникли у Пирсона, Карла (1904 г.), который использовал термин «случайность» для обозначения мера общего отклонения относительно независимость .

См. также критерий независимости хи-квадрат.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если мы рассмотрим двумерную таблицу, содержащую записи для двух качественных категориальных переменных X и Y , которые имеют соответствен r и c , таблица непредвиденных обстоятельств:

	Категории переменной Y			
	Y1	…	Yc	Итого
Категории X1	n11	…	n1c	n1.
…	…	…	…	…
переменная X Xr	nr1	…	nrc	номер
Итого №1		…	нк	н..

где

nij представляет наблюдаемую частоту для категория i переменной X и категория j переменной Y;

ни. представляет собой сумму частот наблюдается для категории i переменной X,

nj представляет собой сумму наблюдаемых частот для категории j переменной Y, а

n.. – общее количество наблюдений.

В случае многомерной таблицы элементы таблицы обозначены nijk, представляющими наблюдаемую частоту для категории i переменной X, категории j переменной Y и категория k переменной Z.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Независимость двух категориальных качественных переменных, представленных в контингентно-

Machine Translated by Google

Циркулярную таблицу можно оценить, выполнив тест независимости хи-квадрат.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Критерий независимости хи-квадрат
- Распределение частоты

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Fienberg, SE: Анализ кросс-классифицированных категориальных данных, 2-е изд. MIT Press, Кембридж, Массачусетс (1980)

Пирсон, К.: К теории непредвиденных обстоятельств. и его связь с ассоциацией и нормальным корреляция. Исследования компании Drapers' Воспоминания, биометрическая сер. I., стр. 1–35. (1904)

Непрерывное распределение Функция

Функция распределения непрерывного случайная величина определяется как вероятность того, что случайная величина примет значение меньше или равно действительному числу.

ИСТОРИЯ

См. вероятность.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Функция, определяемая

$$F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x)dx .$$

называется функцией распределения con незначительная случайная величина.

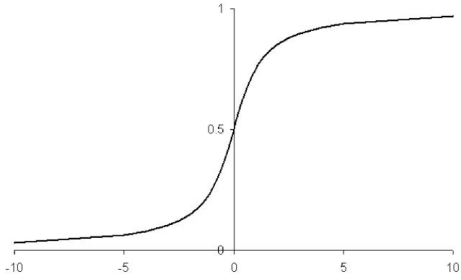
Другими словами, функция плотности f равна производная непрерывного распределения функция.

Свойства непрерывного распределения Функция

1. F(x) — непрерывно возрастающая функция. для всех x;
2. F принимает свои значения в интервале [0, 1];
3. $\lim_{b \to -\infty} F(б) = 0$;
4. $\lim_{б \to \infty} F(б) = 1$;
5. F(x) — непрерывная и дифференцируемая функция.

Эта функция распределения может быть графически представлена в системе координат. Различные значения случайной величины X равны

по оси абсцисс, а соответствующие значения F(x) по оси ординат.



ОБЛАСТИ И ОГРАНИЧЕНИЯ

Вероятность того , что непрерывная случайная величина X примет значение в интервале

val]a, b] для всех a < b, что означает, что P(a < X ≤ b), равно F(b) − F(a) , где F функция распределения случайной величины способный X.

Демонстрация : событие {X ≤ b} может быть записывается как объединение двух взаимоисключающих событий: {X ≤ a} и {a < X ≤ b}:

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\} .$$

Найдя вероятность с каждой стороны уравнение, получаем:

$$P(X \leq b) = P(\{X \leq a\} \cup \{a < X \leq b\})$$
$$= P(X \leq a) + P(a < X \leq b) .$$

Machine Translated by Google

Сумма вероятностей является результатом того факта, что два события исключают друг друга.

Вычитая P(X a) с каждой стороны, мы имеем:

$$P(a < X \quad b) = P(X \quad b) \quad P(X \quad a) .$$

Наконец, из определения функции распределения получаем:

$$P(a < X \quad b) = F(b) \quad F(a) .$$

ПРИМЕРЫ

Рассмотрим непрерывную случайную величину X для которого функция плотности определяется выражением:

$$e(x) = \begin{cases} 1, \text{если } 0 < x < 1 \\ 0 \text{ если нет} \end{cases} .$$

Вероятность того , что X примет значение в интервале [a, b], где 0 < a и b < 1, следующая:

$$P(a \quad X \quad б) = \int_{a}^{б} f(x)dx$$

$$\text{"="} \int_{a}^{б} 1 \, дх$$

$$= x \Big|_{a}^{б}$$

знак равно б - а .

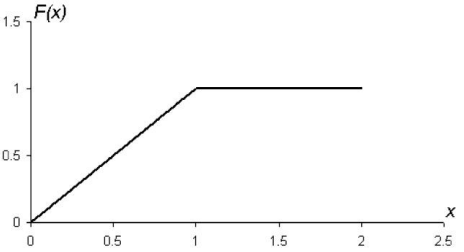
Следовательно, для 0 < x < 1 функция распределения:

$$F(x) = P(X \quad x)$$

$$= P(0 \quad X \quad x)$$

$$= x .$$

Эта функция представлена на следующем рисунке:



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция плотности
- Вероятность
- Случайный эксперимент
- Случайная переменная
- Ценить

Непрерывная вероятность Распределение

Каждая случайная величина имеет соответствующее частотное распределение. Для непрерывной случайной величины это распределение также непрерывно.

Непрерывное распределение вероятностей — это модель , которая представляет частотное распределение непрерывной переменной в лучший путь.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Распределение вероятностей непрерывной случайной величины X задается ее функцией плотности f(x) или ее функцией распределения F(x).

Обычно его можно охарактеризовать ожидаемым значением:

$$E[X] = \int_{Д} x \cdot f(x) \, dx = \mu$$

и его дисперсия:

$$Вар(X) = \int_{Д} (x - м)^2 \cdot f(x)dx$$

$$= E(X - \mu)^2$$

$$= E X2 - E[X]^2 ,$$

Machine Translated by Google

где D представляет собой интервал , охватывающий диапазон значений , которые может принимать X. Одним из существенных свойств непрерывной случайной величины является то, что вероятность того, что она примет конкретное числовое значение , равна нулю, в то время как вероятность того, что она примет значение в интервале (конечном или бесконечном), обычно отлична от нуля.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Наиболее известным непрерывным распределением вероятностей является нормальное распределение.

Непрерывные распределения вероятностей часто используются для аппроксимации дискретных распределений вероятностей. Они используются при построении моделей точно так же, как и при применении статистических методов.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Бета-распределение
- Распределение Коши
- Распределение хи-квадрат
- Функция непрерывного распределения
- Функция плотности
- Дискретное распределение вероятностей
- Ожидаемое значение
- Экспоненциальное распределение
- Распределение Фишера
- Гамма-распределение
- Распределение Лапласа
- Логнормальное распределение
- Нормальное распределение
- Вероятность
- Распределение вероятностей
- Случайная переменная
- Студенческое распределение
- Равномерное распределение
- Дисперсия случайной величины

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Джонсон, Н.Л., Коц, С.: Распределения в статистике: непрерывное одномерное распределение буции, тт. 1 и 2. Уайли, Нью-Йорк (1970).

Контраст

C

В дисперсионном анализе контраст представляет собой линейную комбинацию наблюдений или факторов.

уровни или обработки в факторном эксперименте, где сумма коэффициентов равна
нуль.

ИСТОРИЯ

Согласно Шеффе Х. (1953), Тьюки Дж. В. (1949 и 1951) первыми предложили метод одновременной оценки всех контрастов.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим T1, T2,...,Tk, являющиеся суммами n1, n2,...,nk наблюдений. Линейная функция

$$cj = c1j \cdot T1 + c2j \cdot T2 + \cdots + ckj \cdot Tk$$

является контрастом тогда и только тогда, когда

$$\sum_{i=1}^{k} ni \cdot cij = 0 .$$

Если каждое ni = n, что означает, что если Ti является суммой одного и того же числа наблюдений, условие сводится к следующему:

$$\sum_{i=1}^{k} cij = 0 .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В большинстве экспериментов , включающих несколько вариантов лечения, экспериментатору интересно провести сравнение между различными вариантами лечения. Статистик использует

контрасты для проведения этого типа сравнения
сын.

Machine Translated by Google

ПРИМЕРЫ

Когда проводится дисперсионный анализ для трехуровневого фактора некоторые контрасты проценты:

$c1 = T1 - T2$

$c2 = T1 - T3$

$c3 = T2 - T3$

c4 знак равно $T1 - 2 \cdot T2 + T3$.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Эксперимент
- Факторный эксперимент

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Остле, Б.: Статистика в исследованиях: основные концепции и методы для научных работников. Издательство Государственного колледжа Айовы, Эймс, Айова (1954)

Шеффе, Х.: Метод оценки всех контрастов в дисперсионном анализе. Биометрия 40, 87–104 (1953).

Тьюки, Дж. В.: Сравнение индивидуальных средних дисперсионный анализ. Биометрия5, 99–114 (1949)

Тьюки, Дж. В.: Быстрые и грязные методы в Статистика. Часть II: Простые анализы для Стандартные конструкции. Документы конференции по контролю качества, 1951 г. Американское общество. по контролю качества, Нью-Йорк, стр. 189–197 (1951)

Конвергенция

В статистике термин «конвергенция» связан с теорией вероятностей . Эту статистическую сходимость часто называют стохастической сходимостью, чтобы отличить ее от классической. конвергенция.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Различные типы стохастической сходимости можно определить. Пусть $\{xn\}n$ N — множество случайных величин. Наиболее важные виды стохастическая сходимость:

1. $\{Xn\}n$ сходится по распределению к случайная величина X , если

$$\lim_{n} FXn(z) = FX(z) \quad z \quad ,$$

где FXn и FX — распределение функции Xn и X соответственно. Эта сходимость есть просто точечная сходимость (хорошо известная в математике) множество функций распределения Xн .

2. $\{Xn\}n$ сходится по вероятности к случайной величине X , если:

$$\lim_{n} P(|Xn \quad X| > \varepsilon) = 0 \quad ,$$

для каждого $\varepsilon > 0$.

3. $\{Xn\}n$ почти наверное сходится к случайной величине X , если:

$$П\ ш \mid \lim_{n} Xn(ш) = X(ш) = 1 \quad .$$

4. Предположим, что все элементы Xn имеют конечное ожидание. Множество $\{Xn\}n$ сходится в среднем квадратичном к X , если:

$$\lim_{n} E(Xn - X)^2 = 0 .$$

Обратите внимание, что:

- Почти наверняка сходимость и среднее квадратная сходимость подразумевает сходимость по вероятности;
- Сходимость по вероятности (слабая сходимость) подразумевает сходимость по распределению.

Machine Translated by Google

ПРИМЕРЫ

Пусть Xi — независимые случайные величины, равномерно распределенные на [0, 1].
Определим следующий набор случайных величин из Xi:

$$Zn = n \cdot \min_{i=1,\ldots,n} Кси .$$

Мы можем показать, что множество {Zn}n сходится по распределению к экспоненциальному распределению Z с параметром, равным 1, следующим образом:

$$1 - FZn(t) = P(Zn > t)$$

$$= P \text{ мин } Xi > i=1,\ldots,n \frac{т}{н}$$

$$= P\ X1 > \frac{т}{н} \text{ и } X2 > \frac{т}{—} \text{ и}$$

$$\ldots Xп > \frac{нет}{н}$$

$$\text{инд.} \prod_{я=1}^{н} Си > \frac{т}{н} = 1 - \frac{т}{н}^{н} .$$

Теперь для $\lim_{п} , 1 , \frac{т}{н}^{н} = \exp(-т)$ поскольку:

$$\lim_{п} ФЗн = \lim_{п} P(Zn \quad т)$$

$$= 1 - \exp(-t) = FZ .$$

Наконец, пусть Sn будет числом успехов, полученных в ходе n испытаний Бернулли с вероятностью успеха p. Теорема Бернулли говорит нам, что $Sn_{н}$ сходится по вероятности к «случайной» переменной, которая принимает значение p с вероятностью 1.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Теорема Бернулли
- Центральная предельная теорема
- Теорема сходимости
- Теорема де Муавра – Лапласа.
- Закон больших чисел
- Вероятность
- Случайная переменная
- Стохастический процесс

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Ле Кам, Л.М., Ян, К.Л.: Асимптотика в статистике: некоторые основные понятия. Springer, Берлин Гейдельберг, Нью-Йорк (1990)

Staudte, RG, Sheater, SJ: Надежная оценка и тестирование. Уайли, Нью-Йорк (1990)

Теорема сходимости

Теорема о сходимости приводит к важнейшим теоретическим результатам теории вероятностей. Среди них мы находим закон больших чисел и центральный предел
орем.

ПРИМЕРЫ

Центральная предельная теорема и закон большие числа сходятся
руды.

Закон больших чисел утверждает, что среднее значение суммы одинаково распределенных случайных величин сходится к их общему математическому ожиданию.
С другой стороны, центральная предельная теорема утверждает, что распределение суммы суф достаточно большое число случайных величин стремится приблизиться к нормальному распределению.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Центральная предельная теорема
- Закон больших чисел

Коэффициент корреляции

Простой коэффициент корреляции является мерой силы линейной связи между двумя случайными величинами.

Machine Translated by Google

Коэффициент корреляции может принимать значения
события , происходящие в интервале [1; 1].
два крайних значения этого интервала представляют
совершенно линейная связь между переменными,
«положительная» в первом случае и «отрицательная»
в другом. Значение 0 (ноль) подразумевает
отсутствие линейной зависимости.

Представленный здесь коэффициент корреляции равен
также называемая корреляцией Браве – Пирсона.
коэффициент.

ИСТОРИЯ

Понятие корреляции возникло в
1880-х годов с работами Гальтона, Ф. В его
автобиография «Воспоминания о моей жизни» (1890 г.),
он пишет, что придумал эту концепцию во время
прогулки по территории замка Наворт.
когда ливень заставил его найти убежище.

Согласно Stigler, SM (1989), Porter, TM (1986), проводил
исторический
исследование, когда он нашел забытую статью
написанное Гальтониным в 1890 г. в The North
American Review под заголовком «Родство и
корреляция». В этой статье, которую он опубликовал
сразу после ее открытия, Гальтон (1908 г.)
объяснял сущность и значение
понятие корреляции.
Это открытие было связано с предыдущими работами
математика, особенно тех, кто
наследственность и линейная регрессия. Гальтон имел
интересовался этой областью исследований с тех пор, как
1860. Он опубликовал работу под названием «Природная
наследство» (1889 г.), положившее начало
точка для его мыслей о корреляции.
В 1888 г. в статье, направленной в Королевское
статистическое общество, озаглавленной «Соотношения и их
измерение в основном по антропометрическим
данные», Гальтон использовал термин «корреляция» для
в первый раз, хотя он все еще чередовал термины
«соотношение» и
«корреляция», и он говорил о «соотношении
индекс." С другой стороны, он ссылается на кон
за исключением отрицательной корреляции. В соответствии с
Stigler (1989), Гальтон, по-видимому, только
предположил, что корреляция является положительной
связью.
Пирсона, Карл писал в 1920 г., что корреляция была
открыта Гальтоном, чей
работа «Естественное наследство» (1889) подтолкнула
ему изучить и эту концепцию вместе с двумя
другие исследователи, Уэлдон и Эджворт.
Затем Пирсон и Эджворт разработали
теория корреляции.
Уэлдон считал, что коэффициент корреляции
следует назвать «функцией Гальтона». Как
когда-либо Эджворт заменил термин Гальтона «индекс
взаимосвязи» и термин Уэлдона «индекс отношения Гальтона».
функция» термином «коэффициент корреляции».

Согласно Mudholkar (1982), Pearson, K. систематизировал
анализ корреляции и создал теорию корреляции.

для трех переменных. Исследователи из Uni
версити-колледж, в первую очередь его помощник
Yule, GU, также были заинтересованы в развитии
множественной корреляции.
Спирмен опубликовал первое исследование о рангах
Корреляция в 1904 году.

Среди работ, выполненных
в этой области стоит выделить те
Йоля, который в статье под названием «Почему
мы иногда получаем бессмысленную корреляцию
между временными рядами» (1926) обсуждал
проблема интерпретации корреляционного анализа.
Наконец, устойчивость корреляции была
исследовали Мостеллер и Тьюки (1977).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Простой коэффициент линейной корреляции

Простая линейная корреляция — это термин, используемый для
описывают линейную зависимость между двумя
количественные переменные X и Y (см. простой
линейная регрессия).

Machine Translated by Google

Если X и Y — случайные величины , следующие низкое неизвестное совместное распределение, то коэффициент простой линейной корреляции равен ковариации между X и Y , деленной ed произведением их стандартных отклонений:

$$= \frac{\text{Cov}(X, Y)\ p}{\sigma X \sigma Y} .$$

Здесь Cov(X, Y) — измеренная ковариация между X и Y; σX и σY – стандартные отклонения X и Y соответственно .

Для выборки размера n, (X1, Y1), (X2, Y2),..., (Xn, Yn) из совместного распределения величина

$$p = \frac{\sum_{i=1}^{n} (Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (Xi - \bar{X})^2 \sum_{i=1}^{n} (Yi - \bar{Y})^2}}$$

является оценкой ρ; это корреляция выборки .

Если мы обозначим $(Xi - \bar{X})$ через xi и $(Yi - \bar{Y})$ через yi, мы можем записать это уравнение как:

$$p = \frac{\sum_{i=1}^{n} xiyi}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} .$$

Проверка гипотезы

Чтобы проверить нулевую гипотезу

$$H0 : p = 0$$

против альтернативной гипотезы

$$H1 : p = 0 \quad ,$$

вычисляем статистику t:

$$\tau = \frac{p}{\text{старший}} ,$$

где Sr — стандартное отклонение оценки r :

$$Cp = \sqrt{\frac{1 - r2}{n - 2}} .$$

При H0 статистика t следует распределению Стьюдента с n 2 степенями свободы.

Для заданного уровня значимости α H0 отбрасывается, если $|t| \quad t\alpha, n \quad 2$; значение $t\alpha, n \quad 2$ — критическое значение теста, приведенного в таблице Стьюдента.

Множественный коэффициент корреляции

Известный как коэффициент детерминации

Обозначается R2, определяет, правильно ли гиперплоскость, оцененная по множественной линейной регрессии, соответствует точкам данных.

Значение коэффициента множественной детерминации R2 равно:

$$P2 = \frac{\text{Объясненная вариация}}{\text{Общая вариация}}$$

$$"=" \frac{\sum_{i=1}^{n} (Y\hat{}i - \bar{Y})^2}{\sum_{i=1}^{n} (Yi - \bar{Y})^2} .$$

Он соответствует квадрату множественного коэффициента корреляции. Заметить, что

$$0 \quad R2 \quad 1 \quad .$$

В случае простой линейной регрессии можно вывести следующее соотношение:

$$r = \text{знак}(\beta\hat{}\ 1) \sqrt{R2} \quad ,$$

Machine Translated by Google

где βˆ 1 – оценка регрессии
коэффициент β1, и определяется как:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Если существует линейная зависимость между двумя переменными
коэффициент корреляции равен
1 или 1.

Положительное отношение (+) означает, что два переменные изменяются в одном направлении. Если люди получают высокие баллы в первом переменная (например, независимая переменная), они будут иметь тенденцию к получить высокие баллы по второй переменной (зависимая переменная). Противоположное также верно.

Отрицательное отношение ( ) означает, что лица, набравшие высокие баллы в первом переменная будет иметь тенденцию к снижению баллы во втором и наоборот.

Обратите внимание, что если переменные независимы, коэффициент корреляции равен нулю.
обратный вывод не обязательно верен.
Тот факт, что две или более переменных связаны статистического анализа недостаточно, чтобы заключить, что существует причинно-следственная связь. Наличие статистической корреляции
не доказательство причинно-следственной связи.

Статистика дает многочисленные корреляции коэффициенты. Выбор, что использовать для тот или иной набор данных зависит от различных факторы, такие как:

- Тип шкалы, используемый для выражения способный;
- Характер основного распределения (непрерывный или дискретный);
- Характеристики распределения оценки (линейные или нелинейные).

ПРИМЕРЫ

Данные для двух переменных X и Y _ показано в таблице ниже:

Нет из заказ	Икс	Д	Икс	Икс Д	у "=" Д	xy	x^2	y^2
1	174	64	1,5	1,3	1,95	2,25		1,69
2	175	59	0,5	6,3	3,14	0,25	36,69	
3	180	64	4,5	1,3	5,85	20,25		1,69
4	168	62	7,5	3,3	24,75	56,25	10,89	
5	175	51	0,5	14,3	7,15	0,25	204,49	
6	170	60	5,5	5,3	29,15	30,25	28,09	
7	170	68	5,5	2,7	14,85	30,25		7,29
8	178	63	2,5	2,3	5,75	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
Всего	1755	653				358,5	284,5	1034,1

$\bar{X} = 175,5$ и $\bar{Y} = 65,3$.

Теперь произведем необходимые расчеты чтобы получить коэффициент корреляции между две переменные. Применение формулы дает:

$$p = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \text{ "=" } \frac{358,5}{\sqrt{284,5 \cdot 1034,1}} = 0,66 .$$

Проверка гипотезы

Мы можем рассчитать сметный стандарт отклонение г :

$$Ср = \sqrt{\frac{1 - r^2}{n - 2}} \text{ "=" } \sqrt{\frac{0,56}{10 \quad 2}} = 0,266 .$$

Machine Translated by Google

Вычисление статистики t дает:

$$t = \frac{p - 0}{\text{старший}} \text{ "=" } \frac{0,66}{0,266} = 2,485 .$$

Если мы выберем уровень значимости α, равный 5%, значение из таблицы Стьюдента, t0,025,8, равно равно 2,306.

Поскольку $|t| = 2,485 > t_{0,025,8} = 2,306$, нулевая гипотеза

$$H0 : p = 0$$

отвергается.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент детерминации
- Ковариация
- Зависимость
- Коэффициент ранговой корреляции Кендалла
- Множественная линейная регрессия
- Регрессивный анализ
- Простая линейная регрессия
- Коэффициент ранговой корреляции Спирмена

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Гальтон, Ф.: Соотношения и их измерение, в основном из антропологических данные. проц. Рой. соц. Лонд. 45, 135–145 (1888)

Гальтон, Ф.: Естественное наследование. Макмиллан, Лондон (1889 г.)

Гальтон, Ф.: Родство и корреляция. Север Являюсь. Откр. 150, 419–431 (1890 г.)

Гальтон, Ф.: Воспоминания о моей жизни. Метуэн, Лондон (1908 г.)

Мостеллер, Ф., Тьюки, Дж. В.: Анализ данных и регрессия: второй курс Статистика. Эддисон-Уэсли, Рединг, Массачусетс (1977)

Mudholkar, GS: Множественный коэффициент корреляции. В: Коц, С., Джонсон, Н.Л. (ред.) Энциклопедия статистических наук, об. 5. Уайли, Нью-Йорк (1982)

Пирсон, К.: Исследования по истории статистики и вероятностей. Биометрика 13, 25–45 (1920). Перепечатано в: Pearson, ES, Кендалл М.Г. (ред.) Исследования по истории статистики и вероятностей, том. I. Гриффин, Лондон

Портер, Т. М.: Расцвет статистического мышления, 1820–1900 гг. Университет Принстон Пресса, Принстон, Нью-Джерси (1986)

Стиглер, С.: Рассказ Фрэнсиса Гальтона о изобретение корреляции. Стат. науч. 4, 73–79 (1989)

Юле, Г.Ю.: К теории корреляции. Дж. Рой. Стат. соц. 60, 812–854 (1897 г.)

Юл Г.Ю. (1926) Почему мы иногда получить нонсенс-корреляции между временными рядами? Изучение отбора проб и природы временных рядов. Дж. Рой. Стат. соц. (2) 89, 1–64

Анализ корреспонденции

Анализ корреспонденции – это анализ данных метод, который используется для описания таблиц непредвиденных обстоятельств (или перекрестных таблиц). Этот анализ принимает форму графического представления ассоциаций и «соответствия» между строками и столбцами.

ИСТОРИЯ

Теоретические принципы переписки анализ восходит к работам Хартли, Х.О. (1935) (опубликованных под его первоначальным именем Хиршфельд) и Фишера , Р.А. (1940) по таблицам непредвиденных обстоятельств. Они были впервые представлено в рамках логической статистики.

Термин «анализ корреспонденции» впервые появился осенью 1962 г., а первый презентация этого метода, которая относится к

Machine Translated by Google

этот термин был дан Benzécri, JP в
зима 1963 года. В 1976 году работы Бена
zecri, JP, в которой прослеживаются двенадцать лет его
лабораторные работы, были опубликованы, и с тех пор
тогда алгебраические и геометрические свойства
этого описательного аналитического инструмента имеют
становятся более известными и используемыми.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим таблицу непредвиденных обстоятельств, относящуюся к двум
категориальные качественные переменные X и Y
которые имеют, соответственно, категории r и c :

	Д 1	Д 2	...	Ус	Общий
X1	n11	n12	...	n1c	n1.
X2	n21	n22	...	n2c	n2.
...	...	...	...	...	...
Xr	№1	№2	...	нрк	номер
Общий	п.1	п.2	...	нк	н..

где

nij представляет частоту , с которой категория i
переменной X и категории j переменной
Y наблюдается,

ни. представляет собой сумму наблюдаемых частот
для категории i переменной X,

nj представляет собой сумму наблюдаемых частот для
категории j переменной Y,

n.. представляет собой общее количество наблюдений.

Будем считать, что r с; если не возьмем
транспонирование исходной таблицы и использование
это транспонировать как новую таблицу непредвиденных обстоятельств.
Анализ соответствия таблицы непредвиденных
обстоятельств, в которой строк больше, чем столбцов,
выполняется следующим образом:

1. Строятся таблицы профилей рядов XI и профилей
столбцов XJ .
Для фиксированной линии (столбца) профилем
линии (столбца) является линия (столбец), полученная
путем деления каждого элемента в этой строке
(столбце) на сумму элементов в строке
(столбец).
Профиль линии строки i получается с помощью
делением каждого члена строки i на ni., что
представляет собой сумму наблюдаемых частот в
ряд.
Построена таблица профилей рядов
путем замены каждой строки непредвиденных обстоятельств
таблица со своим профилем:

	Д	Д 2	...	Ус	Общий
X1	1 n11 / n1.	n12 / n1.	...	n1c / n1.	1
X2	n21 / n2.	n22 / n2.	...	н2с / н2.	1
...	...	...	...	...	...
Xr	№1 / номер	№2 / номер	...	нрк / номер	1
Всего n	.1	н .2	... н	.c	р

Также принято умножать каждый элемент таблицы
на 100, чтобы преобразовать
слагаемые в проценты и сделать
сумма слагаемых в каждой строке 100%.
Матрица профиля колонны построена
аналогично, но на этот раз каждый столбец
таблицы сопряженности заменяется
с его профилем: профиль столбца столбца j
получается путем деления каждого члена
столбца j на nj, которое представляет собой сумму
частот , наблюдаемых для категории,
соответствующей этому столбцу.

	Д	Д 2	...	Ус	Общий
X1	1 n11 / п.1	n12 / п.2	...	n1c / nc	н 1.
X2	n21 / п.1	n22 / п.2	...	n2c / nc	н 2.
...	...	...	...	...	...
Xr	№1 / №1	№2 / №2	...	нрк / нк	н р.
Всего 1		1	. . .	1	с

Таблицы профилей строк и профилей столбцов
соответствуют преобразованию
таблица непредвиденных обстоятельств , которая используется для
строки и столбцы сопоставимы.

Machine Translated by Google

2. Определить матрицу инерции V.

Это делается следующим образом: •

Средневзвешенное значение столбца r координаты вычисляются:

$$\text{гдж} = \sum_{i=1}^{p} \frac{\text{ни.}}{\text{н..}} \cdot \frac{\text{ний}}{\text{ни.}} \text{"="} \frac{nj_i}{} \quad j = 1,\ldots,c.\ \text{н..}$$

• Полученные с значения gj записываются r раз в строки матрицы G; • Диагональная матрица DI строится из диагональных элементов ni. ; $\overline{\text{н..}}$

• Наконец, матрица инерции V рассчитывается по следующей формуле:

$$V = (XI - G) \cdot DI \cdot (XI - G)\ .$$

3. Используя матрицу M, состоящую из n.. $\overline{\text{н..}}$ члены на ее диагонали и нулевые члены в остальных местах определяем матрицу C:

$$C = \overline{M} \cdot V \cdot \overline{M}\ .$$

4. Найдите собственные значения (обозначенные kl) и собственные векторы (обозначенные vl) этого матрица C.

Собственные значения kc , kc 1,...,k1 (записанные в порядке убывания) являются инерцией. Соответствующие собственные векторы называются факториальными осями (или осями инерции).

Для каждого собственного значения мы вычисляем соответствующую инерцию, объясняемую факториальной осью. Например, первая факториальная ось объясняет:

$$\frac{100 \cdot \text{лиц1}}{\text{с}} \quad \text{(в \%) инерции.}$$

$$\text{кл}$$

$$\text{л=1}$$

Таким же образом две первые факториальные оси объясняют:

$$\frac{100 \cdot (k1 + k2)}{\text{с}} \quad \text{(в \%) инерции.}$$

$$\text{кл}$$

$$\text{л=1}$$

Если мы хотим узнать, например, число собственных значений и, следовательно, факториальные оси, которые объясняют не менее 3/4 общей инерции, мы суммируем объясненную инерцию каждого из собственных значений, пока не получим 75%.

Затем вычисляем главные оси инерции от этих факториальных осей.

5. Главные оси инерции, обозначенные ul, равны затем дается:

$$ul = M \quad 1 \cdot \overline{vl}\ ,$$

это означает, что его j-я компонента:

$$\text{уджл} = \frac{\overline{nj}}{\text{н..}} \cdot vjl\ .$$

6. Затем вычисляем главные компоненты, обозначаемые yk , которые являются ортогональными проекциями координат строки на главные оси инерции: i- я координата

l -я главная компонента принимает следующее значение:

$$yil = xi \cdot M \cdot ul\ ,$$

означающий, что

$$\text{йил} = \sum_{j=1}^{c} \frac{\text{ний}}{\text{ни.}} \cdot \frac{\overline{\text{н..}}}{nj} \cdot vjl$$

— координата строки i на оси l.

7. После вычисления основных компонент yl (координат строки) определяем главные компоненты координат столбца (обозначается zl) с помощью yl, благодаря формулам транзакций:

$$zjl = \frac{1}{\overline{kl}} \sum_{i=1}^{p} \frac{nij}{} \cdot yil\ ,\ nj$$

для j = 1, 2,...,c;

$$\text{йил} = \frac{1}{\overline{\text{кл}}} \sum_{i=1}^{c} \overline{\text{ний}} \cdot zjl\ ,$$

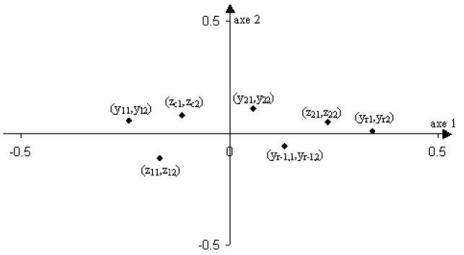
$$\text{ни.}$$

для i = 1, 2,...,r.

Machine Translated by Google

zjl — координата столбца j на
l -я факториальная ось.

8. Одновременное представительство
координаты строки и координаты столбца
на точечной диаграмме с двумя факторными
осями, дает значение оси
в зависимости от точек, с которыми он связан.
Качество представления связано
к доле инерции, объясненной
по двум используемым основным осям. Чем ближе
объясненная инерция к 1, тем лучше
качество.



9. Может быть полезно вставить дополнительные
строки точек или координаты столбца (иллюстративный
переменные) вместе с активными переменными
используется для анализа корреспонденции.
Рассмотрим дополнительную координату строки:

(ns1, ns2, ns3,..., nsc)

профильного ряда

$$\frac{ns1}{нс.}, \frac{нс2}{нс.}, \ldots, \frac{нск}{нс.}.$$

l -й главный компонент точки дополнительной
строки задается второй формулой
пункт 7):

$$йил = \frac{1}{кл} \sum_{j=1}^{c} \frac{нсж}{нс.} \cdot ЗЖЛ,$$

для i = 1, 2,...,r.

Аналогично поступаем со столбцом
координировать, но применить первую формулу
вместо пункта 7.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

При одновременном представлении строки
координата и координата столбца, это
не рекомендуется интерпретировать возможную
близость, пересекаемую между строками и столбцами.
потому что две точки не совпадают
начальное пространство. С другой стороны,
интересно интерпретировать положение координаты
строки, сравнивая ее с набором координат столбца.
координаты (и наоборот).

ПРИМЕРЫ

Рассмотрим пример компании, которая
хочет узнать, насколько здоров ее персонал. Один
из вопросов, затронутых в вопроснике, касается
количества посещений врача
(включая стоматологов) в год. Три возможных
предлагаются ответы:

- От 0 до 6 посещений;
- От 7 до 12 посещений;
- Более 12 посещений в год.

Опрошенные сотрудники распределены по пяти
категории:

- Управленческий персонал старше 40 лет;
- Управленческий персонал моложе 40 лет;
- Сотрудники старше 40 лет;
- Сотрудники моложе 40 лет;
- Офисный персонал.

Результаты представлены в следующей таблице
непредвиденных обстоятельств, в которой указаны поля.
добавлен:

	Количество посещений в год			
	от 0 до 6	от 7 до 12	> 12	Всего
Управленческий штатные сотрудники > 40	5	7	3	15
Управленческий штатные сотрудники < 40	5	5	2	12

Machine Translated by Google

	Количество посещений в год			
	от 0 до 6	от 7 до 12	> 12	Всего
Сотрудники > 40	2	12	6	20
Сотрудники < 40	24	18	12	54
Офис персонал	12	8	4	24
Общий	48	50	27	125

Используя эти данные , мы опишем восемь основные шаги, которые необходимо выполнить для получения графическое изображение сотрудника здоровье с помощью анализа соответствия.

1. Сначала определяем таблицу линейных профилей матрицаXI, разделив каждый элемент на сумма элементов прямой, в которой он расположен:

$$\begin{matrix} 0{,}333 & 0{,}467 & 0{,}2 & 1 \\ 0{,}417 & 0{,}417 & 0{,}167 & 1 \\ 0{,}1 & 0{,}6 & 0{,}3 & 1 \\ 0{,}444 & 0{,}333 & 0{,}222 & 1 \\ 0{,}5 & 0{,}333 & 0{,}167 & 1 \\ 1{,}794 & 2{,}15 & 1{,}056 & 5 \end{matrix}$$

и таблицу профилей столбцов путем деления каждого элемента на сумму элементов в соответствующем столбце:

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. Затем вычисляем матрицу инерции V для частот (определяемых $\frac{ни.}{н..}$ для вал ues i от 1 до 5), действуя в следующим образом:

- Начнем с расчета взвешенного среднее значение пяти точек:

$$\text{гдж} = \sum_{я=1}^{5} \frac{ни.}{н..} \cdot \frac{ний}{ни.} \text{ "=" } \frac{\quad}{н..}$$

для j = 1, 2 и 3;

- Затем мы записываем эти три значения пятью раз в матрице G, как показано ниже:

$$г = \begin{matrix} 0{,}384 & 0{,}400 & 0{,}216 \\ 0{,}384 & 0{,}400 & 0{,}216 \\ 0{,}384 & 0{,}400 & 0{,}216 \\ 0{,}384 & 0{,}400 & 0{,}216 \\ 0{,}384 & 0{,}400 & 0{,}216 \end{matrix} .$$

- Затем мы строим диагональную матрицу DI, содержащий $\frac{ни.}{н..}$;
- Наконец, мы вычисляем матрицу инерция V, определяемая следующим формула:

$$V = (XI \quad G) \cdot DI \cdot (XI \quad G),$$

что в данном случае дает:

$$B = \begin{matrix} 0{,}0175 & -0{,}0127 & -0{,}0048 \\ 0{,}0127 & 0{,}0097 & 0{,}0029 \\ 0{,}0048 & 0{,}0029 & 0{,}0019 \end{matrix} .$$

3. Определим квадратную матрицу третьего порядка M который содержит $\frac{ни.}{н..}$ члены на его диагонали и нулевые члены везде:

$$M = \begin{matrix} 2{,}604 & 0 & 0 \\ 0 & 2{,}5 & 0 \\ 0 & 04.630 & \end{matrix} .$$

Квадратный корень из M, обозначаемый $\sqrt{M}$, равен получается путем извлечения квадратного корня из каждого диагональный элемент M. Используя этот новый матрица, определяем

$$C = \sqrt{M} \cdot V \cdot \sqrt{M}$$

$$C = \begin{matrix} 0{,}0455 & -0{,}0323 & -0{,}0167 \\ 0{,}0323 & 0{,}0243 & 0{,}0100 \\ 0{,}0167 & 0{,}0100 & 0{,}0087 \end{matrix} .$$

Machine Translated by Google

4. Собственные значения C получаются путем диагонализации матрицы . Упорядочивание этих значений в порядке убывания дает:

$$k1 = 0{,}0746$$
$$к2 = 0{,}0039$$
$$к3 = 0 .$$

Объясняемая инерция определяется для каждого из этих значений. Например:

$$k1 : \frac{0{,}0746}{= 95{,}03\% \; 0{,}0746 + 0{,}0039} ,$$

$$k2 : \frac{0{,}0039}{= 4{,}97\% \; . \; 0{,}0746 + 0{,}0039}$$

Для последнего, k3, объясненная инерция равна нулю.

Первые две оси факториала, связанные с собственными значениями k1 и k2, объясняют всю инерцию. Поскольку третье собственное значение равно нулю, нет необходимости вычислять связанный с ним собственный вектор . Затем мы сосредоточимся на вычислении первых двух нормированных собственных векторов:

$$v1 = \begin{matrix} 0{,}7807 \\ 0{,}5576 \\ 0{,}2821 \end{matrix} \quad и$$

$$v2 = \begin{matrix} 0{,}0807 \\ 0{,}5377 \\ 0{,}8393 \end{matrix} .$$

5. Затем вычисляем главные оси инерции к:

$$ui = M \overline{\overline{1}} \cdot vi , \quad \text{для я} = 1 \text{ и } 2 ,$$

где $M \overline{\overline{1}}$ получается инвертированием диагональных элементов $\overline{M}$.

Мы нашли:

$$и1 = \begin{matrix} 0{,}4838 \\ 0{,}3527 \\ 0{,}1311 \end{matrix} \quad и$$

$$и2 = \begin{matrix} 0{,}0500 \\ 0{,}3401 \\ 0{,}3901 \end{matrix} .$$

6. Затем мы вычисляем основные компоненты, проецируя строки на главные оси инерции. Построив вспомогательную матрицу U из двух векторов u1 и u2, определим:

$$Y = XI \cdot M \cdot U$$

$$\text{"="} \begin{matrix} 0{,}1129 & 0{,}0790 \\ 0{,}0564 & 0{,}1075 \\ 0{,}5851 & 0{,}0186 \\ 0{,}1311 & -0{,}0600 \\ 0{,}2349 & 0{,}0475 \end{matrix} .$$

Мы можем видетькоординаты пяти рядов записаны горизонтально в этой матрице Y; например, в первом столбце указаны компоненты, относящиеся к первой факториальной оси, а во втором — компоненты, относящиеся ко второй оси.

7. Затем мы используем координаты строк для найдите эти столбцы с помощью следующих формул перехода, записанных в виде матриц:

$$Z = K \cdot Y \cdot XJ ,$$

где:

K — диагональная матрица второго порядка , которая имеет $1/\overline{\overline{ki}}$ членов на диагонали и нулевые члены в других местах;

Y — транспонирование матрицы, содержащей координаты строк, а;

XJ — матрица профиля колонны.

Machine Translated by Google

Получаем матрицу:

$$Z = \begin{matrix} 0{,}3442 & -0{,}2409 & -0{,}1658 \\ 0{,}0081 & 0{,}0531 & -0{,}1128 \end{matrix},$$

где каждый столбец содержит компоненты одного из трех столбцов

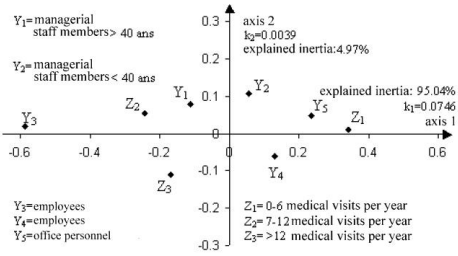
обедает; например, первая строка соответствует каждой координате на первой факториальной оси , а вторая строка – каждой координаты по второй оси.

Мы можем проверить формулу перехода, которая дает Y из Z:

$$Y = XI \cdot Z \cdot K$$

используя те же обозначения, что и ранее.

8. Теперь мы можем представить пять категорий опрошенных людей и три категории ответов, предложенных на одном и том же факторный график:



Мы можем изучать этот факторный график на трех различных уровнях анализа в зависимости от:

- набор категорий опрашиваемых лиц;
- Набор модальностей для медицинской визы его;
- Оба одновременно.

В первом случае непосредственная близость между две строки (между двумя категориями персонала) означают схожие профили визитов к врачу. На факториальном графике это имеет место для сотрудники моложе 40 лет (Y4) и офисный персонал (Y5). Мы можем проверить из таблицы линейные профили, что проценты для этих две категории действительно очень похожи. Точно так же близость между двумя столбцы (представляющие две категории , связанные с количеством посещений врача) указывают аналогичное распределение людей внутри Бизнес для этих категорий. Это может быть наблюдается для модальностей Z2 (от 7 до 12 визитов к врачу в год) и Z3 (более 12 посещений в год).

Если мы рассмотрим строки и столбцы одновременно (а не раздельно, как это делали мы ранее) становится возможным идентифицировать сходство между категориями для определенных модальностей. Например, сотрудники моложе 40 лет (Y4) и офисный персонал (Y5), похоже, ведет себя так же по отношению к здоровье: высокая доля из них (0,44 и 0,5 соответственно) обращаться к врачу реже 6 раз в год (Z1).

В заключение, ось 1 сталкивается на одном рядом с категориями, указывающими на среднее или большое количество посещений (Z2 и Z3) — сотрудников или руководящих работников старше 40 лет (Y1 и Y3) — и с другой стороны с модальностями, связанными с низким числом число посещений (Z1): Y2, Y4 и Y5. Тогда первый фактор можно интерпретировать как важность медицинского контроля в зависимости от возраста.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Таблица сопряженности

Анализ данных

собственное значение

собственный вектор

Факторная ось

Графическое представление

Матрица инерции

Матрица

диаграмма рассеяния

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бензекри, JP: L'Analyse des données. Том. 1: Таксиномия. Том. 2: L'Factoryelle des корреспонденции. Дюно, Париж (1976)

Бензекри, JP: Histoire et préhistoire de l'analyse des données, les cahiers de l'analyse des données 1, no. 1–4. Дюно, Париж (1976)

Фишер, Р.А.: Точность дискриминантных функций. Энн. Ойген. (Лондон) 10 422–429 (1940)

Гринакр, М.: Теория и приложения анализа соответствия. Академический, Лондон (1984)

Hirschfeld, HO: Связь между корреляцией и случайностью. проц. Камб. Филос. соц. 31, 520–524 (1935)

Лебар, Л., Салем, А.: Статистический анализ текстовых материалов Données. Дюно, Париж (1988)

Сапорта, Г.: Вероятность, анализ данных и статистики. Технип, Париж (1990)

Ковариация

Ковариация между двумя случайными переменными Способности X и Y являются мерой того, насколько две случайные величины изменяются вместе. Если X и Y — независимые случайные величины, ковариация X и Y равна нулю. Обратное, однако, неверно.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим X и Y, две случайные величины , определенные в одном и том же пространстве выборок . Ковариация X и Y, обозначаемая Cov(X, Y), определяется выражением

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] ,$$

где E[.] — ожидаемое значение.

Составление правой части уравнения дает:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[XY - XE[Y] \\ &\quad + E[X]E[Г]] \\ &= E[XУ] - E[X]E[У] \\ &\quad - E[X]E[У] + E[X]E[У] \\ &= E[XУ] - E[X]E[У] .\end{aligned}$$

Свойства ковариации

Рассмотрим X, Y и Z, которые являются случайными величинами , определенными в одном и том же , пространстве выборок , и a, b, c и d, которые являются константами Найди это:

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) = acCov (X, Y).

Следствия определения 1. Если X и Y — независимые случайные способный,

$$\mathrm{Cov}(X, Y) = 0 .$$

На самом деле E[XY] = E[X]E[Y], что означает, что:

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0 .$$

Обратное, как правило, неверно: Cov(X, Y) = 0 не обязательно означает, что X и Y независимы.

2. Cov(X, X) = Var(X) , где Var(X) представляет дисперсию X.

Фактически:

$$\begin{aligned}\mathrm{Cov}(X, X) &= E[XX] - E[X]E[X] \\ &= E[X^2] - (E[X])^2 \\ &= \mathrm{Вар}(X) .\end{aligned}$$

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Рассмотрим две случайные величины X и Y и их сумму X + Y. Тогда мы имеем:

$$E[X + Y] = E[X] + E[Y] \text{ и}$$

$$Вар(X + Y) = Вар(X) + Вар(Y) + 2Cov(X, Y) .$$

Теперь мы покажем эти результаты для дискретных переменных.

способен. Если Pji = P(X = xi, Y = yj), мы имеем:

$$E[X + У] = \sum_{я}\sum_{дж} (xi + yj)Pji .$$

$$= \sum_{я}\sum_{дж} xiPji + \sum_{я}\sum_{дж} yjPji$$

$$= \sum_{я} xi \sum_{j} Пджи + \sum_{дж} yj \sum_{я} Пджи$$

$$= ксиПи + yjPj$$

$$= E[X] + E[У] .$$

Более того:

$$\begin{aligned} Var(X + Y) &= E[(X + Y)^2] - (E[X + Y])^2 \\ &= E[X^2] + 2E[XУ] + E[У^2] \\ &\quad (E[X + Y])^2 \\ &= E[X^2] + 2E[XУ] + E[У^2] \\ &\quad (E[X] + E[У])^2 \\ &= E[X^2] + 2E[XУ] + E[У^2] \\ &\quad (Э[X])^2 \quad 2E[X]E[Y] \quad (E[Y])^2 \\ &= Var(X) + Var(Y) + 2(E[XY] \\ &\quad E[X]E[У]) \\ &= Var(X) + Var(Y) + 2Cov(X, Y) . \end{aligned}$$

Эти результаты можно обобщить для n случайных величин X1, X2,...,Xn, где xi с ожидаемым значением , равным E(Xi) , и дисперсией, равной Var(Xi). Тогда мы имеем:

$$\begin{aligned} &E[X1 + X2 + \cdots + Xn] \\ &= E[X1] + E[X2] + \cdots + E[Xn] \\ &= \sum_{я=1}^{н} Э[Си] . \end{aligned}$$

$$\begin{aligned} &Var(X1 + X2 + \cdots + Xn) \\ &= Var(X1) + Var(X2) + \cdots + Var(Xn) \\ &\quad + 2[Cov(X1, X2) + \cdots + Cov(X1, Xn) \\ &\quad + Cov(X2, X3) + \cdots + Cov(X2, Xn) \\ &\quad + \cdots + Cov(Xn \quad 1, Xn)] \\ &= \sum_{я=1}^{н} Вар(Кси) + 2 \sum_{я=1}^{n\ 1} \sum_{j>я} Cov(Xi, Xj) . \end{aligned}$$

ПРИМЕРЫ

Рассмотрим два психологических теста, проведенных последовательно. Каждый субъект получает оценку X от 0 до 3 за первый тест и оценку Y от 0 до 2 за второй тест. Учитывая, что вероятности того, что X будет равен 0, 1, 2 и 3, составляют соответственно 0,16, 0,3, 0,41 и 0,13, а вероятности того, что Y будет равен 0, 1 и 2, соответственно 0,55, 0,32 и 0,13, мы имеем:

$$\begin{aligned} E[X] &= 0 \cdot 0{,}16 + 1 \cdot 0{,}3 + 2 \cdot 0{,}41 \\ &\quad + 3 \cdot 0{,}13 \\ &= 1{,}51 \end{aligned}$$

$$\begin{aligned} E[Y] &= 0 \cdot 0{,}55 + 1 \cdot 0{,}32 + 2 \cdot 0{,}13 \\ &= 0{,}58 \end{aligned}$$

$$\begin{aligned} E[XУ] &= 0 \cdot 0 \cdot 0{,}16 \cdot 0{,}55 \\ &\quad + 0 \cdot 1 \cdot 0{,}16 \cdot 0{,}32 + \ldots \\ &\quad + 3 \cdot 2 \cdot 0{,}13 \cdot 0{,}13 \\ &= 0{,}88 . \end{aligned}$$

Machine Translated by Google

Затем мы можем вычислить ковариацию X и Y:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[XY] \quad E[X]E[Y] \\ &= 0{,}88 - (1{,}51 \cdot 0{,}58) \\ &= 0{,}00428 .\end{aligned}$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент корреляции
- Ожидаемое значение
- Случайная переменная
- Дисперсия случайной величины

Ковариационный анализ

Ковариационный анализ — это метод, используемый для оценки и проверки эффектов лечения. Это проверяет, есть ли существенная разница между средствами нескольких видов лечения принимая во внимание наблюдаемые значения переменной до обработки .

Ковариационный анализ — это точный способ проведения сравнений лечения , поскольку он включает в себя подгонку переменной отклика Y к сопутствующей переменной X, которая соответствует значениям , наблюдаемым до лечения.

ИСТОРИЯ

Ковариационный анализ восходит к 1930 году. Впервые он был разработан Фишером Р.А. (1932). После этого другие авторы применяли ковариационный анализ к сельскохозяйственным и медицинским задачам. Например, Бартлетт М.С. (1937) применил ковариационный анализ к своим исследованиям выращивания хлопка в Египте и удоев коров зимой.

Delurry, DB (1948) использовал ковариационный анализ для сравнения действия различных препаратов (атропина, хинидина, атрофина) на мышцы крыс.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Мы рассматриваем здесь ковариационный анализ полностью рандомизированного плана, подразумевающего один фактор.

Линейная модель , которую мы будем рассматривать, следующий:

$$Y_{ij} = \mu + \tau_i + \beta X_{ij} + \varepsilon_{ij} \ , \ i = 1, 2,\ldots,t, \ j = 1, 2,\ldots,n_i$$

где

Y_{ij} представляет наблюдение j, получающее лечение i, μ — общее среднее, общее для всех видов лечения,

τ_i — фактический эффект лечения i на th наблюдение,

X_{ij} – значение сопутствующей переменной, и

ε_{ij} – экспериментальная ошибка наблюдения Y_{ij}.

Расчеты

Чтобы рассчитать коэффициент F , который поможет нам определить, есть ли существенная разница между обработками, нам необходимо вычислить суммы квадратов и суммы произведений. Следовательно, если $\bar{X}_{i.}$ и $\bar{Y}_{i.}$ являются соответственно средними значениями X и Y для лечения i, а если $\bar{X}_{..}$ и $\bar{Y}_{..}$ являются соответственно средними значениями всех значений X и Y, мы получаем формулы, приведенные ниже.

1. Общая сумма квадратов для X:

$$SXX = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2 .$$

2. Сумма квадратов для Y (SYY) рассчитывается так же, как SXX, но вместо X используется Y.

Machine Translated by Google

3. Сумма произведений X и Y:

$$\text{ССЮ} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}..)(Y_{ij} - \bar{Y}..) .$$

4. Сумма квадратов обработок для икс:

$$\text{ТХХ} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{X}_i. - \bar{X}..)^2 .$$

5. Сумма квадратов лечения для Y (ТЫГ) рассчитывается так же, как ТХХ, но X сменяет Ю.

6. Сумма произведений обработок X и Y :

$$\text{ТХУ} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{X}_i. - \bar{X}..)(\bar{Y}_i. - \bar{Y}..) .$$

7. Сумма квадратов ошибок для X :

$$\text{EXX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i.)^2 .$$

8. Сумма квадратов ошибок для Y: рассчитывается так же, как EXX, но X сменяет Ю.

9. Сумма произведений ошибок X и Y:

$$\text{EXY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i.)(Y_{ij} - \bar{Y}_i.) .$$

Подстановка соответствующих значений и вычисление этих формул соответствует анализу дисперсии для каждого из X, Y и XY. Степени свободы , связанные с эти различные формулы выглядят следующим образом:

1. На общую сумму: $\sum_{i=1}^{t} n_i - 1.$
2. По сумме обработок : $t - 1.$
3. По сумме ошибок : $\sum_{i=1}^{t} n_i - t.$

Подгонка переменной Y к сопутствующей переменной X дает две новые суммы квадраты:

1. Скорректированная общая сумма квадратов:

$$\text{SSобщ} = \text{SYY} - \frac{\text{S2}_{XY}}{\text{CXX}} .$$

2. Скорректированная сумма квадратов ошибки:

$$\text{SSerr} = \text{EYY} - \frac{\text{E2}_{XY}}{\text{EXX}} .$$

Новые степени свободы для этих двух суммы:

1. $\sum_{i=1}^{t} n_i - 2;$
2. $\sum_{i=1}^{t} n_i - t - 1,$ где из-за настройки вычитается степень свободы .

Третья скорректированная сумма квадратов, скорректированная сумма квадратов обработок , равна предоставлено:

$$\text{SStr} = \text{SStot} - \text{SSerr} .$$

Оно имеет то же число степеней свободы t 1, что и прежде.

Таблица ковариационного анализа

Теперь у нас есть все элементы, необходимые для составить таблицу ковариационного анализа. сумма квадратов деленная на количество степеней свободы дает средства квадраты.

Источник вар иация	Степени бесплатно дом	Сумма квадратов и продуктов		
		$\sum_{i=1}^{t} x_i^2$	$\sum_{i=1}^{t} x_i y_i$	$\sum_{i=1}^{t} y_i^2$
Обращаться менты	$t - 1$	ТХХ	ТХУ	ТЫГ
Ошибки	$\sum_{i=1}^{t} n_i - t$	EXX	EXY	EYY
Общий	$\sum_{i=1}^{t} n_i - 1$	SXX	SXY	SYY

Machine Translated by Google

Примечание: числа в x2 я и y2 я столбцы не могут быть отрицательными; с другой рука, числа в столбце i=1 xiyi т может быть отрицательным.

Корректирование

Источник вар иация	Степени свобода	Сумма квадраты	Среднее значение квадраты
Обращаться менты	т - 1	ССтр	MCtr
Ошибки	т i=1 ni t 1	SSerr	MCerr
Общий	т i=1 ni - 2	TSS	

Ф Соотношение: тестирование методов лечения

Отношение F , используемое для проверки нулевой гипотезы что есть существенная разница между средства лечения после корректировки _ переменной X, определяется как:

$$\Phi = \frac{MCtr}{MCerr} .$$

Отношение соответствует распределению Фишера с t 1 и t i=1 ni t 1 степеней свободы. Нулевая гипотеза

$$H0 : \tau1 = \tau2 = \cdots = \tau t$$

будет отклонено на уровне значимости α, если коэффициент F выше или равен значению таблицы Фишера, другими словами, если:

$$F \quad Ft \quad 1, \tau \quad _{я=1} ni \quad t \quad 1,\alpha .$$

Понятно, что мы предполагаем, что коэффициент β отличен от нуля при проведении ковариационного анализа. Если это не так, достаточно простого дисперсионного анализа .

Тест относительно наклона β

Итак, мы хотели бы знать, существует ли значительный эффект сопутствующего вари способности до применения лечения.
Для проверки этой гипотезы предположим, что нулевая гипотеза

$$H0 : \beta = 0$$

и альтернативная гипотеза

$$H1 : \beta = 0 .$$

Коэффициент F может быть установлен:

$$\Phi = \frac{E2\ XY/EXX}{MCerr} .$$

Он следует распределению Фишера с 1 и т i=1 ni t 1 степень свободы.
нулевая гипотеза будет отвергнута на уровне значимости α, если отношение F превосходит или равно значению в таблице Фишера, в другими словами, если:

$$F \quad F1, \tau \quad _{я=1} ni \quad t \quad 1,\alpha .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Основные гипотезы , которые необходимо построить до начала ковариационного анализа, те же самые, что и для анализа. дисперсионного или регрессионного анализа. Эти являются гипотезы нормальности, однородности (гомоскедастичности), дисперсии и независимости.

В ковариационном анализе, как и в анализе дисперсия, нулевая гипотеза предполагает, что независимые выборки взяты из разных популяций с одинаковыми средними значениями.
Тем более, что всегда есть условия связанный с любым статистическим методом, те, которые применяются к ковариационному анализу, следует:

1. Распределение населения должно быть примерно нормально, если не полностью нормальный.

Machine Translated by Google

2. Популяции, из которых взяты образцы должны иметь одинаковую дисперсию σ2, что означает:

$$\sigma^2_1 = \sigma^2_2 \overset{"="}{} \dots = \sigma^2_k ,$$

где k — количество популяций, сравниваться.

3. Выборки должны быть выбраны случайным образом, и все выборки должны быть независимыми.

Мы также должны добавить основную гипотезу, специфичную для ковариационного анализа, которая состоит в том, что проведенное лечение не должно влиять на значения сопутствующих переменных способный X.

ПРИМЕРЫ

Рассмотрим эксперимент , состоящий в сравнении эффектов трех разных диет на поголовье коров . Данные представлены в виде таблицы , который включает в себя три различных диеты, каждая из которых было отведено пять подъездов. начальные веса обозначаются сопутствующей переменной X (в кг), а прирост веса (после обработки) обозначаются Y:

Диеты

123					
ХУ	ХУ	ХУ			
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

Сначала мы вычисляем различные суммы квадратов и продукты:

1. Общая сумма квадратов для X:

$$SXX = \sum_{i=1}^{3}\sum_{j=1}^{5} (Xij - \bar{X}..)^2$$

$$= (32 - 29,8667)^2 + \dots + (32 - 29,8667)^2$$

$$= 453,73 .$$

2. Сумма квадратов для Y:

$$СГГ = \sum_{i=1}^{3}\sum_{j=1}^{5} (Yij - \bar{Y}..)^2$$

$$= (167 - 167,4667)^2 + \dots + (162 - 167,4667)^2$$

$$= 3885,73 .$$

3. Сумма произведений X и Y:

$$ССЮ = \sum_{i=1}^{3}\sum_{j=1}^{5} (Xij - \bar{X}..)(Yij - \bar{Y}..)$$

$$= (32 - 29,8667) \cdot (167 - 167,4667) + \dots + (32 - 29,8667) \cdot (162 - 167,4667)$$

$$= 158,93 .$$

4. Сумма квадратов процедур для X:

$$ТХХ = \sum_{i=1}^{3}\sum_{j=1}^{5} (\bar{X}i. - \bar{X}..)^2$$

$$= 5 (28,2 - 29,8667)^2 + 5 (26,2 - 29,8667)^2 + 5 (35,2 - 29,8667)^2$$

$$= 223,33 .$$

5. Сумма квадратов процедур для Y:

$$ТЫГ = \sum_{i=1}^{3}\sum_{j=1}^{5} (\bar{Y}i. - \bar{Y}..)^2$$

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. Сумма произведений обработок X и Y:

$$\text{ТХУ} = \sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{X}_i. - \bar{X}_{..})(\bar{Y}_i. - \bar{Y}_{..})$$

$$= 5\,(28{,}2 - 29{,}8667)\cdot(159{,}6 - 167{,}4667) + 5\,(35{,}2 - 29{,}8667)\cdot(168{,}6 - 167{,}4667) = -27{,}67\ .$$

7. Сумма квадратов ошибок для X:

$$\text{EXX} = \sum_{i=1}^{3}\sum_{j=1}^{5}(X_{ij} - \bar{X}_i.)^2$$

$$= (32 - 28{,}2)^2 + \ldots + (32 - 35{,}2)^2 = 230{,}40\ .$$

8. Сумма квадратов ошибок для Y:

$$\text{ЭГГ} = \sum_{i=1}^{3}\sum_{j=1}^{5}(Y_{ij} - \bar{Y}_i.)^2$$

$$= (167 - 159{,}6)^2 + \ldots + (162 - 168{,}6)^2 = 3343{,}20\ .$$

9. Сумма произведений ошибок X и Y:

$$\text{EXY} = \sum_{i=1}^{3}\sum_{j=1}^{5}(X_{ij} - \bar{X}_i.)(Y_{ij} - \bar{Y}_i.)$$

$$= (32 - 28{,}2)(167 - 159{,}6) + \ldots + (32 - 35{,}2)(162 - 168{,}6) = 186{,}60\ .$$

Степени свободы , связанные с этими различными вычислениями, следующие:

1. По общей сумме:

$$\sum_{i=1}^{3} n_i - 1 = 15 - 1 = 14\ .$$

2. По суммам лечения:

$$т - 1 = 3 - 1 = 2\ .$$

3. Для сумм ошибок:

$$\sum_{i=1}^{3} n_i - т = 15 - 3 = 12\ .$$

Подгонка переменной Y к сопутствующей переменной X дает две новые суммы квадратов:

1. Общая скорректированная сумма квадратов:

$$\text{SSобщ} = \text{SYY} - \frac{\text{S2}_{XY}}{\text{SXX}} = 3885{,}73 - \frac{158{,}932}{453{,}73} = 3830{,}06\ .$$

2. Скорректированная сумма квадратов ошибки:

$$\text{SSerr} = \text{EYY} - \frac{\text{E2}_{XY}}{\text{EXX}} = 3343{,}20 - \frac{186.602}{230{,}40} = 3192{,}07\ .$$

Новые степени свободы для этих двух суммы:

1. $\sum_{i=1}^{3} n_i - 2 = 15 - 2 = 13;$
2. $\sum_{i=1}^{3} n_i - т - 1 = 15 - 3 - 1 = 11.$

Скорректированная сумма квадратов воздействий определяется по формуле:

$$\text{SStr} = \text{SSобщ} - \text{SSerr} = 3830{,}06 - 3192{,}07 = 637{,}99\ .$$

Machine Translated by Google

Он имеет то же число степеней свободы, что и прежде:

$$т - 1 = 3 - 1 = 2 \quad .$$

Теперь у нас есть все элементы, необходимые для Чтобы установить ковариационный анализ таблицы. Сумма квадратов деленная на градусы свободы дает средства квадратов.

Источник вар иация	Степени бесплатно дом	Сумма квадратов и продуктов		
		$\sum_{я=1}^{3} x_{я}^{2}$	$\sum_{я=1}^{3} xiyi$	$\sum_{я=1}^{3} ты_{я}^{2}$
Обращаться менты	2	223,33	27,67	543,53
Ошибки	12	230,40	186,60	3343,20
Всего	14	453,73	158,93	3885,73

Корректирование

Источник вариация	Степени бесплатно дом	Сумма квадраты	Среднее значение квадраты
Обращаться менты	2	637,99	318,995
Ошибки	11	3192,07	290,188
Общий	13	3830.06	

Отношение F , которое используется для проверки нуля гипотеза о том, что нет существенной разницы между средствами лечения

после корректировки переменной Y определяется как:

$$Ф = \frac{MCtr}{MCerr} \text{ "=" } \frac{318,995}{290,188} = 1,099 .$$

Если мы выберем уровень значимости моря α = 0,05, значение F в таблице Фишера равно:

$$F2,11,0,05 = 3,98 .$$

Поскольку F < F2,11,0,05, мы не можем отвергнуть нулевая гипотеза , и поэтому мы заключаем, что нет существенной разницы между реакции на три диеты после того, как переменная Y будет приведена к исходному весу X.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- План экспериментов
- Отсутствующий анализ данных
- Регрессивный анализ

C

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бартлетт, М.С.: Некоторые примеры статистических методов исследования в сельском хозяйстве и прикладная биология. Дж. Рой. Стат. соц. (Прил.) 4, 137–183 (1937)

Делури, Д.Б.: Ковариационный анализ. Биометрия 4, 153–170 (1948).

Фишер, Р.А.: Статистические методы для Исследовательские работники. Оливер и Бойд, Эдин Бург (1925)

Huitema, BE: Анализ ковариации и альтернативы. Уайли, Нью-Йорк (1980).

Вильдт, А.Р., Ахтола, О.: Анализ ковариации (Серия статей Sage University Papers по Количественные приложения в социальных сетях наук, статья 12). Мудрец, Тысяча Оукс, Калифорния (1978)

Ковариация

Часто бывает интересно, особенно в экономике, сравнить два временных ряда.
Поскольку мы хотим измерить уровень зависимости между двумя переменными, это чем-то напоминает концепцию корреляции. Однако в данном случае с момента
ряды связаны третьей переменной, временем, нахождение коэффициента корреляции ведут только к искусственной связи.
Действительно, если рассматривать два временных ряда , xt и yt, которые представляют собой совершенно независимые явления и являются линейными функциями

Machine Translated by Google

времени:

$$xt = a \cdot t + b \quad ,$$

$$yt = c \cdot t + d \quad ,$$

где a, b, c и d — константы, всегда можно устранить фактор времени между двумя уравнениями и получить функциональное соотношение типа y = e · Это соотношение утверждает, что существует lin x + f . уха зависимость между двумя временными рядами, что не так.

Поэтому измерение корреляции между эволюцией двух явлений во времени не предполагает наличия связи между ними. Термин ковариация поэтому используется вместо корреляции, и эта зависимость измеряется с помощью коэффициента ковариации. Мы можем различать: • коэффициент линейной ковариации; • Коэффициент ковариации тенденции.

ИСТОРИЯ

См. коэффициент корреляции и временной ряд.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Чтобы сравнить два временных ряда yt и xt, первым шагом является попытка представить их на одном графике.

Однако визуальное сравнение, как правило, затруднено. Производится следующая замена переменных :

$$Yt = \frac{yt - \bar{y}}{Sy} \quad \bar{x} \text{ и } Xt = \frac{xt}{x} \quad ,$$

которые являются центрированными и приведенными переменными

где Sy и Sx — стандартные отклонения соответствующих временных рядов.

Мы можем различать следующие коэффициенты ковариации:

• Коэффициент линейной ковариации

Форма этого выражения идентична форме для коэффициента корреляции r, но здесь расчеты не имеют одно и то же понимание, потому что цель состоит в том, чтобы обнаружить возможное существование отношения между вариации, которые сами связаны с время и измерять порядок величины

$$C = \frac{\sum_{t=1}^{n} (xt - \bar{x}) \cdot (yt - \bar{y})}{\sqrt{\sum_{t=1}^{n} (xt - \bar{x})^2 \cdot \sum_{t=1}^{n} (yt - \bar{y})^2}} \, .$$

Это дает значения от -1 до +1.

Если он близок к ±1, существует линейная зависимость между временными эволюциями двух переменных.

Заметить, что:

$$C = \frac{\sum_{t=1}^{n} Xt \cdot Yt}{n} \, .$$

Здесь n — количество наблюдений, а Yt и Xt — центрированный и редуцированный ряды, полученные заменой переменной соответственно . •

Коэффициент ковариации тенденции

Влияние, оказываемое средством, устраняется путем расчета:

$$K = \frac{\sum_{t=1}^{n} (xt - Txt) \cdot (yt - Tyt)}{\sqrt{\sum_{t=1}^{n} (xt - Txt)^2 \cdot \sum_{t=1}^{n} (yt - Tyt)^2}} \, .$$

Средние $\bar{x}$ и $\bar{y}$ просто заменены значениями вековых трендов Txt и Tyt каждого временного ряда.

Коэффициент ковариации тенденции K также принимает значения от -1 до +1, и чем ближе он приближается к ±1, тем сильнее ковариация между временными рядами.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Есть много примеров необходимости сравнить два временных ряда в экономике: для Например, если сравнить эволюцию цена продукта к эволюции количество товара на рынке, или эволюция национального дохода в Эволюция сделок с недвижимостью. Это важно знать, есть ли своего рода зависимость между двумя явлениями, которые развиваются во времени: это цель измерения ковариации.

Визуальное сравнение двух временных рядов важная операция, но часто это трудная задача, потому что:

- Сравниваемые данные могут происходят из очень разных доменов и представляют очень разные порядки величины, поэтому предпочтительнее изучать отклонения от среднего .
- Пики и впадины двух временных рядов могут иметь очень разные амплитуды; это тогда предпочтительнее гомогенизировать дисперсии, связав вариации обратно к стандартному отклонению времени ряд.

Визуальное сравнение упрощается, если мы рассмотрим центрированный и редуцированный вари возможности , полученные с помощью следующей переменной изменения:

$$Yt = \frac{yt - \bar{y}}{Си} \quad \text{и} \quad Xt = \frac{xt - \bar{x}}{x}.$$

Также аналогично соотношению коэффициент, могут существовать нелинейные отношения между двумя переменными, которые дают значение C что близко к нулю.

Поэтому важно соблюдать осторожность во время интерпретация.

Коэффициент ковариации тенденции предпочтительно используется, когда связь между временной ряд не является линейным.

ПРИМЕРЫ

Изучим ковариацию между двумя Временные ряды.

Переменная xt представляет годовое производство сельскохозяйственной продукции ; переменная yt – его среднегодовая цена за еди постоянные евро.

т	хт	йт
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

$$\sum_{т=1}^{8} xt = 2560 \quad , \qquad \sum_{т=1}^{8} yt = 24{,}8 \quad ,$$

что дает $\bar{x} = 320$ и $\bar{y} = 3{,}1$.

хт	(хт − $\bar{x}$)2	йт	(йт − $\bar{y}$)2
0	102,2	4,84	
340	115600	0,1	0,01
20	400	0,9	0,81
130	16900	0,3	0,09
0	0	0,4	0,16
80	6400	0,4	0,16
40	1600	1,1	1,21
150	22500	0,6	0,36

$$\sum_{т=1}^{8} (xt - \bar{x})^2 = 163400 \quad ,$$

$$\sum_{т=1}^{8} (yt - \bar{y})^2 = 7{,}64 \quad ,$$

что дает $\sigma x = 142{,}9$ и $\sigma y = 0{,}98$.

Центрированные и приведенные значения Xt и Yt затем рассчитываются.

Machine Translated by Google

Xt	Yt	Xt · Yt	Xt · Yt 1
0,00	2,25	0,00	5,36
2,38	0,10	0,24	0,01
0,14	0,92	0,13	0,84
0,91	0,31	0,28	0,00
0,00	0,41	0,00	0,23
0,56	0,41	0,23	0,11
0,28	1,13	0,32	1,18
1,05	0,61	0,64	

Затем рассчитывается коэффициент линейной ковариации рассчитано:

$$C = \frac{\sum_{t=1}^{8} Xt \cdot Yt}{8} = \frac{0,18}{8} = 0,0225 .$$

Если сравнить наблюдения Xt с Yt-1, что означает производство в этом году по сравнению с предыдущим годом, мы получаем:

$$C = \frac{\sum_{t=2}^{8} Xt \cdot Yt \quad 1}{8} = \frac{7,71}{8} = 0,964 .$$

Коэффициент линейной ковариации для сдвига одного года очень сильно. Существует сильный (положительная) ковариация со сдвигом в один год между двумя переменными.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент корреляции
- Скользящее среднее
- Светская тенденция
- Среднеквадратичное отклонение
- Временная последовательность

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кендалл, MG: временные ряды. Гриффин, Лондон (1973)

Ру, В .: Описательная статистика. Экономика, Париж (1987)

Кокс, Дэвид Р.

Кокс, Дэвид Р. родился в 1924 году в Бирмингеме в Англии. Он изучал математику в Кембриджском университете и получил докторскую степень по прикладной математике в Лидском университете в 1949 г. С 1966 по 1988 год он был профессором статистика в Имперском колледже Лондона и затем с 1988 по 1994 год преподавал в Наффилде. Колледж, Оксфорд.

Кокс, Дэвид — выдающийся статистик. Он был посвящен в рыцари королевой Елизаветой II в 1982 г. благодарность за вклад в статистическую науки, и многие университеты Англии присвоили ему звание доктора Honoris Causa. и в других местах. С 1981 по 1983 год был Президент Королевского статистического общества; он также был президентом Общества Бернулли. лет с 1972 по 1983 год и президентом Международный статистический институт с 1995 г. до 1997 года.

Из-за разнообразия предметов, которые он изучал и развивал, профессор Кокс глубокое влияние в его области. Его назвали Почетный доктор Университета г. Невшатель в 1992 году.

Кокс, сэр Дэвид является автором и соавтором более 250 статей и 16 книг, а также с 1966 по 1991 год он был редактором Биометрика.

Некоторые основные работы и статьи Кокса, Сэр Дэвид:

1964 (и Box, GEP) Анализ трансформации (с обсуждением) Дж. Рой. Стат. соц. сер. Б 26, 211–243.

1970 Анализ двоичных данных. Метуэн, Лондон.

Machine Translated by Google

1973 (и Хинкли, Д.В.) Теоретическая
Статистика. Чепмен и Холл, Лондон.

1974 (и Аткинсон, АС) Планирование
экспериментов по различению между
модели. Дж. Рой. Стат. соц. сер. Б 36,
321–348.

1978 (и Хинкли, Д.В.) Проблемы и
Решения в теоретической статистике.
Чепмен и Холл, Лондон.

1981 (и Снелл, Э.Дж.) Прикладная статистика:
Принципы и примеры. Чепмен
& Зал.

1981 Теория и общие принципы в
статистика. Дж. Рой. Стат. соц. сер. А
144, 289–297.

2000 (и Рид, Н.) Теория дизайна
Эксперименты. Чепмен и Холл, Лондон.

Кокс, Мэри Гертруда

Кокс, Гертруда Мэри родилась в 1900 г.
недалеко от Дейтона, штат Айова, США. Ее амбиции были
чтобы помогать людям, и поэтому сначала она
изучала курс социальных наук в течение двух лет.
Потом работала в детском доме для юных
мальчики в Монтане в течение двух лет. Чтобы
чтобы стать директором приюта, она
решила продолжить свое образование в Айове
Государственный колледж. Она закончила Айову
Государственный колледж в 1929 году. Чтобы
оплатить учебу, Кокс, Гертруда работала со Snedecor,
Джордж Уоддель, ее профессор, в своей
статистической лаборатории, что привело к тому,
что она заинтересовалась статистикой. После окончания,
она начала учиться на доктора психологии. В
1933 году, прежде чем закончить ее, доктор
Снедекор съел Джорджа, тогдашнего директора
Статистической лаборатории штата Айова убедил
ее стать его ассистентом, что
она согласилась, хотя и осталась в
области психологии, потому что она работала над
оценка статистических тестов в психологии и
анализ психологических
данные.

1 ноября 1940 года она стала
директор экспериментального отдела
Статистика штата Северная Каролина.

В 1945 году Совет по общему образованию дал
ее разрешение на создание института статистики
в Университете Северной Каролины с
кафедрой математической статистики в
Чапел-Хилл.
Она была одним из основателей журнала Journal of
Международное биометрическое общество в 1947 г.,
и она была его директором с 1947 по
1955 г. и его президент с 1968 по 1969 г.
В 1956 г. она была президентом Американской
статистической ассоциации (ASA).
1978 год.

Основные работы Кокса, М. Гертруды:

1957 г. (и Кокран, В.) Экспериментальный
Дизайны, 2-е изд. Уайли, Нью-Йорк

сп Критерий

Cpciterion — это цитерин по выбору модели.
в линейной регрессии. Для линейной регрессии
модели с параметрами p , включая любой
постоянный член, в модели эмпирическое правило
выбрать модель, в которой значение Cp равно
близко к количеству членов в модели.

ИСТОРИЯ

Представленный Маллоусом, Колином Л. в 1964 году,
использовался критерий выбора модели Cp
с тех пор в качестве критерия оценки
качество подгонки регрессионной модели.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть $Yi = \beta 0 + \sum_{j=1}^{p-1} Xji\beta j + \varepsilon i, \quad я = 1,\ldots,n$ — множественная линейная регрессия модель. Обозначим среднеквадратичную ошибку как СКО(уи). Критерий, введенный в этом разделе можно использовать для выбора модели с минимальная сумма среднеквадратических ошибок:

$$\begin{aligned} \text{СКО (yi)} &= \text{E (и - йи)}^2 \\ &\quad \sigma 2\,(1 \quad 2hii) \\ &= \text{E (уи - уи)}^2 \\ &\quad \sigma 2(1 \quad 2hii) \\ &= \text{E(RSS)} \quad \sigma 2(n \quad 2p)\,, \end{aligned}$$

где

n - количество наблюдений,

p - количество оцениваемых параметров,

hii — диагональные элементы шляпы матрица и

yi является оценкой yi.

Напомним следующее свойство hii :

$$\text{привет} = p\,.$$

Определите коэффициент

$$\begin{aligned} \text{п} &\overset{"="}{} \frac{\text{МСЭ (и)}}{\sigma 2} \\ &\overset{"="}{} \frac{\text{E (RSS)} \quad \sigma 2\,(n \quad 2p)}{\sigma 2} \\ &\overset{"="}{} \frac{\text{Э (РСС)}}{\sigma 2} \quad \text{п + 2р.} \end{aligned}$$

Если модель верна, мы должны иметь:

$$\text{E (RSS)} = (n \quad p)\,\sigma 2 \quad ,$$

что подразумевает

$$\text{п} = p\,.$$

На практике мы оцениваем $\text{п}_\text{к}$

$$\text{Cp} = \frac{\text{RSS}}{\sigma 2} - \text{п} + 2p,$$

где σ2 — оценка σ2. Здесь мы оцениваем σ2 , используя s2 полной модели. Для этой полной модели, мы фактически получаем Cp = p, что не интересный результат. Однако для всех остальных моделей, где мы используем только как подмножество объясняющих переменных, коэффициент Cp может иметь значения, отличные от p. Из моделей, включающих только подмножество объясняющих переменных, мы затем выбрать те, для которых значение Cp является ближайшим к п.

Если у нас есть k объясняющих переменных, мы можем также определите коэффициент Cp для модели, которая включает в себя подмножество X1,...,Xp 1, p k k объясняющих переменных в следующем манера:

$$\text{Cp} = \frac{(n \quad k)\,\text{RSS X1,...,Xp} \quad 1}{\text{RSS (X1,...,Xk)}} \quad \text{п + 2р,}$$

где RSS X1,...,Xp 1 — сумма квадраты невязок, относящихся к модели с p 1 объясняющими переменными, и RSS (X1,...,Xk) представляет собой сумму квадратов остатки, относящиеся к модели со всеми k объясняющие переменные. Из двух моделей которые включают p 1 объясняющих переменных, выбираем по критерию тот для которого значение коэффициента Cp равно ближайший к стр.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Критерий Cp используется при выборе переменных. При использовании с критерием R2 это Критерий может сказать нам о качестве соответствие выбранной модели. Основная идея такой процедуры заключается в следующем: вместо

Machine Translated by Google

пытаться объяснить одну переменную, используя все из доступных объясняющих переменных, это иногда возможно определить базовую модель с подмножеством этих переменных, и объяснительная сила этой модели почти такие же, как и у модели, содержат все объясняющие переменные. Другая причина этого заключается в том, что коллинеарность объясняющие переменные имеют тенденцию к уменьшению точность оценки, поэтому она может быть полезна удалить некоторые поверхностные переменные.

ПРИМЕРЫ

Рассмотрим некоторые данные, относящиеся к Чикагскому пожаров 1975 г. Обозначим переменную cor ответ на логарифм числа пожаров на 1000 домохозяйств на i- й район Чикаго в 1975 г. по Y, а переменные, соответствующие доле домохозяйств построенных до 1940 г., на количество краж и к среднему доходу по округу i на X1, X2 и X3 соответственно.

Поскольку этот набор содержит три объясняющие переменные, у нас есть 23 = 8 возможных моделей. Мы разделить восемь возможных уравнений на четыре наборы:

1. Множество A содержит единственное уравнение без объясняющие переменные:

$$Y = \beta 0 + \varepsilon .$$

2. Множество B содержит три уравнения с одним объясняющая переменная:

$$Y = \beta 0 + \beta 1X1 + \varepsilon$$

$$Y = \beta 0 + \beta 2X2 + \varepsilon$$

$$Y \text{ знак равно } \beta 0 + \beta 3X3 + \varepsilon .$$

3. Множество C содержит три уравнения с двумя объясняющие переменные:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \varepsilon$$

$$Y = \beta 0 + \beta 1X1 + \beta 3X3 + \varepsilon$$

$$Y \text{ знак равно } \beta 0 + \beta 2X2 + \beta 3X3 + \varepsilon .$$

4. Множество D содержит одно уравнение с тремя объясняющие переменные:

$$Y \text{ знак равно } \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \varepsilon .$$

Округ	Ико 1	Ико 2	Ико 3	Д
я	xi 1	xi 2	xi 3	йи
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

Округ я	ик1 xi 1	ик2 xi 2	ик3 xi 3	Д йи
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

Источник: Эндрюс и Герцберг (1985).

Обозначим получившиеся модели следующим образом: 1 для X1, 2 для X2, 3 для X3, 12 для X1 и X2, 13 для X1 и X3, 23 для X2 и X3 и 123 для полной модели. В следующей таблице представлены результаты, полученные для Критерий Ср для каждой модели:

Модель	сп
1	32.356
2	29.937
3	18.354
12	18.936
13	14.817
23	3,681
123	4.000

Если рассматривать модель из группы B (содержащую одну объясняющую переменную), то число оцениваемых параметров равно p = 2 и ни одно из значений Cp для трех моделей приближается к 2. Если мы теперь рассмотрим модель из группы C (с двумя объясняющими переменными), p равно 3 и Cp модели 23 подходит к этому. Наконец, для полного модели мы находим, что Cp = 4, т.е. также количество оцениваемых параметров, но это не интересный результат, как объяснялось ранее. Поэтому наиболее разумным выбор для модели выглядит следующим образом:

$$Y \text{ знак равно } \beta 0 + \beta 2X2 + \beta 3X3 + \varepsilon .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коэффициент детерминации

Коллинеарность

Матрица шляпы

Среднеквадратическая ошибка

Регрессивный анализ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Эндрюс Д.Ф., Герцберг, А.М.: Данные: Сборник задач от многих Поля для студентов и научных работников. Спрингер, Берлин Гейдельберг Нью-Йорк (1985)

Маллоус, К.Л. Выбор переменных в линейной регрессии: графическое пособие. Представлено на Центральном региональном совещании Институт математической статистики, Манхэттен, Канзас, 7–9 мая 1964 г. (1964 г.)

Маллоус, К.Л.: Некоторые комментарии к Ср. Технометрика 15, 661–675 (1973).

Крамер, Харальд

Крамер, Харальд (1893–1985) поступил в Стокгольмский университет в 1912 году в порядке изучать химию и математику; он стал учеником Леффлера, Миттаг и Рисс, Марсель. В 1919 году Крамер был назван доцент Стокгольмского университета. В то же время он работал в качестве актуарий страховой компании Svenska Страхование жизни, которое позволило ему учиться вероятность и статистика.

Machine Translated by Google

Его основная работа по актуарной математике
Теория коллективного риска. В 1929 г. его
попросили создать новое отделение в Стокгольме,
и он стал первым шведским профессором
актуарной и статистической математики.
матика. В конце Второй мировой войны
он написал свою основную работу «Математические
методы статистики», которая впервые была
опубликована в 1945 г. и недавно (в 1999 г.) была
переиздана.
Некоторые основные работы и статьи Крамера,
Харальда

1946 Математические методы статистики.
Издательство Принстонского университета, Принстон,
Нью-Джерси.

1946 Теория коллективного риска: обзор теории с
точки зрения теории случайных процессов.
Юбилейный том Skandia, Стокгольм.

Критерий Итого

Среднеквадратическая ошибка

Критерий общей среднеквадратичной ошибки — это способ сравнения оценок параметров смещенной или несмещенной модели.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Позволять

$$\beta = \beta\ 1,\ldots,\ \beta\ p\quad 1$$

bea-вектор оценок параметров регрессионной модели. Мы определяем общую среднеквадратичную ошибку, TMSE, вектора β оценок как сумму среднеквадратических ошибок (MSE) его компонентов.
Мы помним, что

$$\begin{aligned} \text{СКО}\ \beta\ j &= \text{E}\ \beta j - \beta j\ _ \quad ^2 \\ &= \text{V}\beta j + \text{E}\beta j \quad \beta j\ _\ _\ _\ _ \quad ^2, \end{aligned}$$

где E (.) и V (.) — обычные символы, используемые для обозначения ожидаемого значения и дисперсии. Определим общую среднеквадратичную ошибку как:

$$\begin{aligned} \text{TMSE}(\beta j\ ) &= \sum_{j=1}^{p-1} \text{СКО}\ \beta\ j \\ &\text{"="} \sum_{j=1}^{p-1} \text{E}\ \beta j \quad \beta j\ _ \quad ^2 \\ &\text{"="} \sum_{j=1}^{p-1} \text{Var}\ \beta\ j + \text{E}\ \beta\ j \quad \beta j \quad ^2 \\ &= (p \quad 1)\sigma 2 \cdot \text{След}\ (\text{V}) \\ &\quad + \sum_{j=1}^{p-1} \text{E}\ \beta j \quad \beta j\ _ \quad ^2 . \end{aligned}$$

где Vis матрица дисперсии-ковариации из β .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

К сожалению, когда мы хотим вычислить общую среднеквадратичную ошибку

$$\text{TMSE}\ \beta = \sum_{j=1}^{p-1} \text{E}\ \beta j \quad \beta j\ _ \quad ^2$$

вектора оценок

$$\beta = \beta\ 1,\ldots,\ \beta\ p\quad 1$$

для параметров модели

$$\text{Yi} = \beta 0 + \beta 1\text{Xs}\quad i1\ + \ldots + \beta p\quad 1\text{Xs}\quad ip\quad 1\ + \varepsilon i\ ,$$

нам нужно знать значения модели
параметры βj, заведомо неизвестные для всех реальных данных. Обозначение Xs означает, что j
данные для j-й объясняющей переменной были стандартизированы (см. стандартизированные данные). Следовательно, чтобы оценить эти TMSE , мы генерируем данные из структурного аналога.

Machine Translated by Google

модель. Использование генератора псевдослучайных числа, мы можем смоделировать все данные для искусственной модели проанализируйте ее с помощью различных моделей регрессии (например, простой регрессия или гребенчатая регрессия) и вычислить то, что мы называем общей квадратичной ошибкой, TSE векторов оценок β , относящихся к каждому методу:

$$TSE(\hat{\beta}) = \sum_{j=1}^{p-1} (\beta j \quad \beta j)^2 .$$

Мы повторяем эту операцию 100 раз, гарантируя, что 100 наборов данных являются псевдонезависимыми. Для каждой модели в среднем 100 TSE дает хорошую оценку TMCЭ. Обратите внимание, что некоторые статистики предпочитают модель, полученная путем выбора переменных благодаря своей простоте. С другой стороны, др. статистики предпочитают гребневой метод, потому что он использует всю доступную информацию.

ПРИМЕРЫ

Обычно мы можем сравнить следующие методы регрессии: линейная регрессия по средние квадраты, гребенчатая регрессия или метод выбора переменных.

В следующем примере тринадцать порций цемента были проверены. Каждая порция состоит из четырех ингредиентов, указанных в стол. Цель состоит в том, чтобы определить, как количества xi1, xi2, xi3 и xi4 этих четырех ингредиентов влияют на количество yi, отдаваемое тепло. из-за затвердевания цемента.

Тепло, отдаваемое цементом

Часть	Ингредиент				Нагревать
	1234				
я	xi 1	xi 2	xi 3	xi 4	йи
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 количество теплоты, отданное за счет закалка i-й порции (в джоулях);
xi1 количество ингредиента 1 (трехкальциевого алюмината) в i-й порции;
xi2 количество ингредиента 2 (трехкальциевого силиката) в i-й порции;
xi3 количество ингредиента 3 (тетракальций алюмоферрит) в i-м участке;
xi4 количество ингредиента 4 (силиката двухкальциевого) в i-й порции.

В этом параграфе мы сравним оценки, полученные методом наименьших квадратов (LS) регрессия с полученными гребнем регрессия (R) и полученные с помощью метод выбора переменных (SV) через общий среднеквадратическая ошибка TMSE. Три вектора оценок, полученные каждым методом являются:

$$YLS = 95{,}4 + 9{,}12X1^{c} + 7{,}94X2^{c} + 0{,}65X3^{c} \quad 2{,}41x4^{c}$$

$$YR = 95{,}4 + 7{,}64X1^{c} + 4{,}67X2^{c} \quad 0{,}91X3^{c} \quad 5{,}84x4^{c}$$

$$YSV = 95{,}4 + 8{,}64X1^{c} + 10{,}3x2^{c}$$

Отметим здесь, что все оценки были получены из стандартизированных объяснительных

Machine Translated by Google

гу переменные. Мы сравниваем эти три оценки, используя суммарные среднеквадратические ошибки трех векторов:

$$\hat{\beta} LS = \hat{\beta} \quad LS1 , \hat{\beta} LS2 , \hat{\beta} LS3 , \hat{\beta}_{ЛС4}$$

$$\hat{\beta} R = \hat{\beta} \quad R1 , \hat{\beta} R2 , \hat{\beta} R3 , \hat{\beta} R4$$

$$\hat{\beta} SV = \hat{\beta} \quad SV1 , \hat{\beta} SV2 , \hat{\beta} SV3 , \hat{\beta} CB4 .$$

Здесь индекс LS соответствует методу наименьших квадратов, R — гребневому методу, а SV — выбору переменной.

метод соответственно. Для этого последнего метода оценки коэффициентов uns выбранные переменные в модели рассматриваются быть нулем. В нашем случае имеем:

$$\hat{\beta} LS = (9,12,\ 7,94,\ 0,65,\quad 2,41)\ \hat{\beta} R =$$

$$(7,64,\ 4,67 ,\quad 0,91,\quad 5,84)\ \hat{\beta} SV =$$

$$(8,64,\ 10,3,\ 0,\ 0) .$$

Мы решили аппроксимировать лежащий в основе процесс, который приводит к получению данных по цементу, с помощью следующего уравнения наименьших квадратов:

$$YiMC = 95,4 + 9,12Xs + 7,94Xs\ i1\ i2 + 0,65Xs \quad 2,41Xs\ i3 \quad _{i4} + \varepsilon i .$$

Процедура состоит из генерации 13 случайных членов ошибки ε1,..., ε13 100 раз на основе нормального распределения с среднее значение 0 и стандартное отклонение 2,446 (напомним, что 2,446 является оценкой σ методом наименьших квадратов для данных по цементу). Затем мы вычисляем Y1LS,...,Y13LS, $_{ij}$ используя значения Xs в таблице данных. Таким образом, мы генерируем 100 отсчетов Y1LS,...,Y13LS из 100 отсчетов ε1,..., ε13 .

Три метода применяются к каждому из этих 100 образцов Y1LS,...,Y13LS (всегда используя те же значения для Xs ij), что дает 100 оценок $\hat{\beta}$LS , $\hat{\beta}$R и $\hat{\beta}$SV . Обратите внимание, что эти три метода применяются к этим 100 образцам без какого-либо влияния на результаты уравнений

$$YiLS = 95,4 + 9,12Xs \quad _{i1} + 7,94X c \quad _{i2} + 0,65X c \quad _{i3} \quad 2,41X c_{i4} ,$$

$$YiR = 95,4 + 7,64Xs \quad _{i1} + 4,67X c \quad _{i2} \quad 0,91Xs \quad 5,84Xs \text{ и } i3\ i4$$

$$YiSV = 95,4 + 8,64Xs \quad _{i1} + 10,3X c \quad _{i2}$$

получено для исходного образца. Несмотря на то, что метод выбора переменных выбрали переменные Xi1 и Xi2 в YiSV = 95,4 + 8,64Xs + 10,3Xs i1 i2, возможно, что для одной из этих 100 выборок метод выбрал лучшие переменные Xi2 и Xi3 , или только Xi3, или все подмножество четырех доступных переменных.

Точно так же, несмотря на то, что уравнение YiR = 95,4 + 7,64Xs + 4,67xXs 0,91Xs i2 i3 $_{i1}$ 5,84Xs $_{i4}$ было получено при k = 0,157 значение k пересчитывается для каждой из этих 100 выборок (подробнее см. пример гребневой регрессии).

Из этих 100 оценок $\hat{\beta}$LS , $\hat{\beta}$SV и $\hat{\beta}$R мы можем рассчитать 100 ETC для каждого метода, которые мы обозначаем как:

$$ETCLS = \hat{\beta} \quad LS1 \quad 9,122 \quad \hat{\beta} LS3LS2 \quad 7,942 + 2,412 \quad 0,652 + \hat{\beta} + \hat{\beta} \quad ЛС4 +$$

$$ETCR = \hat{\beta} R2 \quad R1 \quad 9,122 \quad 0,6527,942 + \hat{\beta} R3 + \hat{\beta} \quad + \beta R4 + 2,412$$

Machine Translated by Google

$$ETCSV = \hat{\beta} SV2 \; SV1 \quad 9{,}122$$

$$7{,}942 + \hat{\beta} SV3$$

$$+ \hat{\beta} 0{,}652 \; 2{,}412 .$$

$$+ \hat{\beta} \quad CB4 +$$

Средние значения 100 значений ETCLS, ETCR и ETCSV представляют собой оценки TMSELS, TMSEV и TMSER: оценки TMSE для трех рассматриваемых методов.

После проведения такого моделирования были получены следующие оценки:

$$TMSELS = 270 \quad ,$$

$$TMSER = 75 .$$

$$TMCЭСВ = 166 .$$

Они дают следующие различия:

$$TMCЭЛС — TMCЭСВ = 104 \quad ,$$

$$TMSELS — TMSER = 195 \quad ,$$

$$TMCЭСВ - TMCEP = 91 .$$

Поскольку стандартные отклонения 100 об.

обслуживаемых разностей составляют соответственно 350, 290 и 280, мы можем вычислить приблизительные 95% доверительные интервалы для разности связи между TMSE двух методов

$$TMSELS \quad TMSEV = 104 \pm \frac{2 \cdot 350}{100} ,$$

$$TMSELS \quad TMSER = 195 \pm 100 \; 2 \; \frac{2 \cdot 290}{\cdot 280} ,$$

$$TMSEV \quad TMSER = 91 \pm \frac{}{100} \cdot$$

Мы получаем

$$34 < TMSELS \quad TMSEV < 174 \quad ,$$

$$137 < TMSELS \quad TMSER < 253 \quad ,$$

$$35 < TMSEV \quad TMSER < 147 \quad .$$

Таким образом, мы можем заключить (по крайней мере, для конкретной модели, используемой для генерации смоделированных данных, и принимая во внимание нашу цель — получить малую TMSE), что лучшим из рассмотренных методов является гребневой метод с последующей процедурой выбора переменных.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Предвзятость

Ожидаемое значение

Матрица шляпы

Среднеквадратическая ошибка

Регрессия хребта

Стандартизированные данные

Дисперсия

Матрица дисперсии-ковариации

Взвешенный метод наименьших квадратов

ССЫЛКИ

Box,

GEP, Draper, NR: Основа для выбора дизайна поверхности отклика.
Варенье. Стат. доц. 54, 622–654 (1959)

Додж, Ю.: Анализ применения регрессии.
Дюно, Париж (1999)

Критическое значение

При проверке гипотез критическое значение — это предельное значение , при котором мы принимаем решение отвергнуть нулевую гипотезу H0 для данного уровня значимости.

ИСТОРИЯ

Понятие критического значения было введено Нейманом, Ежи и Пирсоном, Эгоном Шарпом в 1928 г.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Критическое значение зависит от типа используемого теста (двусторонний тест или односторонний тест).

Machine Translated by Google

тест справа или слева), вероятность
распределение и уровень значимости α.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Критическое значение определяется из
вероятностное распределение статистики
связанные с тестом. Это определяется
сверяясь со статистической таблицей , соответствующей
этому распределению вероятностей (нормальное
таблица, таблица студента, таблица Фишера, таблица
хи-квадрат и т. д.).

ПРИМЕРЫ

Компания производит стальные тросы. С использованием
выборке размера n = 100, он хочет проверить,
достаточно ли диаметры кабелей соответствуют
требуемому диаметру .
0,9 см в целом.
Стандартное отклонение σ популяции известно и
равно 0,05 см.
В этом случае проверка гипотезы включает
двусторонний тест. Гипотезы следующие:

нулевая гипотеза H0 : μ = 0,9

альтернативная гипотеза H1 : μ = 0,9 .

До уровня значимости α = 5%, глядя на нормальную
таблицу, мы находим, что критическое значение z α
$\frac{}{2}$ равняется 1,96.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Доверительный интервал
Проверка гипотезы
Уровень значимости
Статистическая таблица

ССЫЛКА

Нейман Дж., Пирсон Э.С.: Об использовании и
интерпретация определенных тестовых критериев для
для целей статистического вывода, части I
и II. Биометрика 20A, 175–240, 263–
294 (1928)

C

Циклические колебания

Циклические колебания – это термин,
описывают колебания, происходящие в течение длительного
периоды около вековой линии тренда или кривой
временного ряда.

ИСТОРИЯ

Смотрите временной ряд.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим Yt, временной ряд , заданный его
компонентами; Yt можно записать как:

- Yt = Tt ·St ·Ct ·It (мультипликативная модель),
  или;
- Yt = Tt + St + Ct + It (аддитивная модель).

где

Yt — данные в момент времени t;
Tt – вековой тренд в момент времени t;
St – сезонная вариация в момент времени t;
Ct — циклическое колебание в момент времени t, а;
Это нерегулярное изменение в момент времени t.

Первый шаг при исследовании временного ряда
всегда заключается в определении векового тренда Tt,
а затем определить сезонную изменчивость
St. После этого можно отрегулировать начальную
данные временного ряда Yt согласно этим
два компонента:

- $\frac{Yt}{Ст \cdot Тт} = Ct \cdot It$ (мультипликативная модель);
- Yt St Tt = Ct + It (аддитивная модель).

Чтобы избежать циклических колебаний, взвешенный
скользящая средняя устанавливается за несколько
только месяцы. Использование скользящих средних
позволяет использовать для сглаживания
неравномерных колебаний It за счет сохранения
циклических колебаний Ct.

Machine Translated by Google

Выбор взвешенной скользящей средней позволяет нам придать больший вес центральным значениям по сравнению с экстремальными значениями, чтобы более точно воспроизвести циклические колебания. веса будут присвоены центральным значениям и малых весов до крайних значений. Например, для скользящей средней , рассматриваемой на интервале в пять месяцев, веса -0,1, 0,3, 0,6, 0,3 и -0,1 могут быть использовал; так как их сумма равна 1, не будет потребность в нормализации.

Если значения Ct · It (вытекающие из корректировки, выполненные в отношении вековой тенденции и к сезонным вариациям ции) обозначаются через Xt, значение циклическое колебание за месяц t определяется:

$$Ct = 0{,}1 \cdot Xt \quad 2 + 0{,}3 \cdot Xt \quad 1 + 0{,}6 \cdot Xt + 0{,}3 \cdot Xt{+}1 \quad 0{,}1 \cdot Xt{+}2 .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Оценка циклических колебаний позволяет:

- Определите максимумы или минимумы, которые временной ряд может достигать.
- Выполнение краткосрочного или среднесрочного прогноза инж.
- Определите циклические компоненты.

Ограничения и преимущества использования взвешенное скользящее среднее при оценке циклические колебания следующие:

- Взвешенные скользящие средние могут сглаживать кривая с циклическими колебаниями, все еще сохраняя большую часть исходной флуктуации, потому что они точно сохраняют амплитуды циклов.
- Использование нечетного количества месяцев для установить скользящую среднюю облегчает лучшее центрирование полученных значений .
- Трудно изучать циклические колебания временного ряда , потому что циклы обычно различаются по длине и амплитуде. Этот обусловлено наличием множества факторов, где влияние этих факторов может переход от одного цикла к другому. Никто моделей, используемых для объяснения и предсказания такие колебания были обнаружены полностью удовлетворительно.

ПРИМЕРЫ

Давайте установим скользящее среднее значение за пять месяцев скорректированных данных . по светскому тренду и сезонному вариации.

Используем также веса 0,1, 0,3, 0,6, 0,3, 0,1; так как их сумма равна 1, то нет необходимости в нормализации.

Пусть Xi — скорректированные значения Ci · Ii:

$$Ci = 0{,}1 \cdot Xi \quad 2 + 0{,}3 \cdot Xi \quad 1 + 0{,}6 \cdot Xi + 0{,}3 \cdot Xi{+}1 \quad 0{,}1 \cdot Xi{+}2 .$$

В приведенной ниже таблице показана электроэнергия, потребляемая уличными фонарями каждый месяц в г. миллионов киловатт-часов за годы 1952 и 1953 гг. Данные уточнены по светскому тренду и сезонному вариации.

Год	Месяц	Данные Си	Скользящее среднее на 5 месяцев Си
1952	Дж.	99,9	
	Ф	100,4	
	М	100,2	100,1
	А	99,0	99,1
	М	98,1	98,4
	Дж	99,0	98,7
	Дж	98,5	98,3
	А	97,8	98,3
	С	100,3	99,9
	О	101,1	101,3
	Н	101,2	101,1
	Д	100,4	100,7

Machine Translated by Google

Год Месяц Данные	Си	Скользящее среднее на 5 месяцев	Си
1953 Дж.	100,6	100,3	
Ф	100,1	100,4	
М	100,5	100,1	
А	99,2	99,5	
М	98,9	98,7	
Дж	98,2	98,2	
Дж	98,4	98,5	
А	99,7	99,5	
С	100,4	100,5	
О	101,0	101,0	
Н	101.1		
Д	101.1		

Никаких существенных циклических эффектов в этих данные и незначительные эффекты не означают никаких эффектов. Начало экономической часто ищут цикл, но это только кажется каждые 20 лет.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Прогнозирование
- Неправильная вариация
- Скользящее среднее
- Сезонная вариация
- Светская тенденция
- Временная последовательность

ССЫЛКА

Коробка, GEP, Дженкинс, GM: временные ряды Анализ: прогнозирование и управление (серия inTimeSeriesAnalysis).HoldenDay,San Франциско (1970)

Уолд, НО (ред.): Библиография о времени Ряды и случайные процессы. Оливер и Бойд, Эдинбург (1965)

C

Machine Translated by Google

Дэниелс, Генри Э.

Дэниэлс, Генри Эллис родился в Лондоне в Октябрь 1912 г. После окончания Эдинбургского университета в 1933 г. он продолжил свою учится в Кембриджском университете. Получив докторскую степень в Эдинбургском университете, он вернулся в Кембридж читать лекции по математике в 1947 г. В 1957 г. Дэниелс, Генри стал первым профессором математики в Бирмингемский университет, должность, которую он занимал до 1978 г., до выхода на пенсию. Область исследования что интересовало Дэниэлса, Генри, так это статистика вывода, аппроксимация седловой точки в статистике, эпидемиологическом моделировании и статистическая теория применительно к текстильной технологии. Когда он был в Бирмингеме, он основал ежегодное собрание статистиков в Gregynog Hall в Поуисе. Эти ежегодные встречи постепенно стали одним из наиболее хорошо принят среди английских статистиков.

С 1974 по 1975 год он был президентом Королевское статистическое общество, присудившее ему серебряная медаль Гая в 1957 году и в золота в 1984 г. В 1980 г. он был избран членом Королевского общества, а в 1985 г. – почетным членом Международного статистического общества. Общество. Он умер в 2000 году.

Некоторые основные работы и статьи Дэниэлс, Генри Э.:

1954 г. Седловые приближения в статистике. Анна. Мат. Стат. 25, 631–650.

1955 г. Обсуждение статьи Бокса, GEP. и Андерсон, С.Л. Дж. Рой. Стат. соц. сер. Б 17, 27–28.

1958 г. Обсуждение статьи Кокса, Д.Р. Дж. Рой. Стат. соц. сер. Б 20, 236–238.

1987 Аппроксимации вероятности хвоста. Междунар. Стат. Откр. 55, 137–48.

Данные

Датум (множественное число данных) является результатом наблюдение за популяцией или за образец . Слово «датум» латинское и означает «что-то дано»; он используется в математике для обозначения элемента информации (не обязательно числового), из которого делается вывод могут быть сделаны.

Обратите внимание, что число (или любая другая форма описание), которое не обязательно содержит любую информацию нельзя путать с данные, которые действительно содержат информацию.

Machine Translated by Google

Данные, полученные в результате наблюдений, относящийся к изучаемой переменной . Эти данные являются количественными, качественными, дискретными или непрерывными, если соответствующая переменная является количественной, качественной, дискретной или непрерывной соответственно.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Двоичные данные

Категориальные данные

Неполные данные

Наблюдение

Население

Образец

Пространственные данные

Стандартизированные данные

Ценить

Переменная

ССЫЛКИ Federer, WT: Сбор данных. В: Kotz, S., Johnson, NL (eds.) Encyclopedia of Statistical Sciences, vol. 2. Уайли, Нью-Йорк (1982)

Анализ данных

Часто одним из первых шагов, выполняемых в научных исследованиях, является сбор данных.

Эти данные обычно организованы в двумерные таблицы. Такая организация обычно упрощает извлечение информации о данных, другими словами, их анализ.

В самом широком смысле анализ данных можно считать сущностью статистики , с которым связаны все остальные аспекты предмета.

ИСТОРИЯ

Поскольку анализ данных охватывает множество различных методов статистического анализа, трудно кратко осветить историю анализа данных. Тем не менее, мы можем быстро просмотреть хронологию наиболее фундаментальных аспектов предмета: • Первая публикация по исследовательскому анализу данных датируется 1970–1971 гг. и была написана Tukey, JW. Это была первая версия его работы «Исследовательский анализ данных», опубликованной в 1977 г. • Теоретические принципы анализа корреспонденции принадлежат Хартли, Х.О. (1935) (опубликовано под его первоначальным немецким именем Хиршфельд) и Фишеру , Р.А. (1940). Однако эта теория была в значительной степени развита в 1970-х годах Бензекри, Дж. П.

• Первые исследования по классификации были в биологии и в зоологии. Самая старая форма типологии была разработана Галеном (129–199 гг. н. э.). Методы численной классификации основаны на идеи Адансона (в восемнадцатом веке) и были развиты, среди прочих, Зубином (1938) и Торндайком (1953).

ОБЛАСТИ И ОГРАНИЧЕНИЯ Область анализа данных включает множество различных статистических методов.

Типы анализа данных можно классифицировать следующим образом:

1. Исследовательский анализ данных, который включает (как следует из названия) изучение данных посредством: • графического представления данных, • преобразования данных (при необходимости), • обнаружения выбросов. , • Разработка исследовательских гипотез , которые не были предусмотрены в начале эксперимента , • Надежная оценка.

Machine Translated by Google

2. Анализ исходных данных, который включает:
   - Выбор статистических методов для применяется к данным.
3. Многомерный анализ данных, который включает:
   - Дифференциальный анализ, • Преобразование данных , уменьшающее количество измерений и функций.
   интерпретация высказываний, • Поиск структуры.
4. Специальные формы анализа данных , подходящие для различных аналитических задач; к этим формам относятся:
   - Анализ соответствий, • Анализ множественных соответствий, • Классификация.
5. Подтверждающий анализ данных, включающий оценку и проверку результатов анализа; это включает: • оценку параметров, • проверку гипотез,
   - обобщение и заключение.
   сион.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Классификация

Кластерный анализ

Анализ корреспонденции

Данные

Исследовательский анализ данных

Проверка гипотезы

Статистика

Трансформация

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бензекри, JP: L'Analyse des données. Том. 1: Таксиномия. Том. 2: L'Factoryelle des корреспонденции. Дюно, Париж (1976)

Бензекри, JP: Histoire et préhistoire de l'analyse des données, les cahiers de l'analyse des données 1, вып. 1–4. Дюно, Париж (1976)

Эверитт, Б.С.: Кластерный анализ. Холстед, Лондон (1974)

Фишер, Р.А.: Точность дискриминантных функций. Энн. Ойген. (Лондон) 10 422–429 (1940)

Гауэр, Дж. К.: Классификация, геометрия и анализ данных. В: Бок, НН (ред.) Классификация и родственные методы Анализ данных. Эльзевир, Амстердам (1988)

Hirschfeld, HO: Связь между корреляцией и случайностью. проц. Камб. Филос. соц. 31, 520–524 (1935)

Торндайк, Р.Л.: Кто принадлежит семье? Психометрика 18, 267–276 (1953).

Tukey, JW: Explanatory Data Analysis, ограниченное предварительное издание. Эддисон Уэсли, Рединг, Массачусетс (1970–1971)

Тьюки, Дж. В.: Исследовательский анализ данных. Аддисон-Уэсли, Рединг, Массачусетс (1977)

Зубин, Дж.: Метод измерения единомыслия. Дж. Ненормальный. Социальная психология. 33, 508–516 (1938)

Сбор данных

При сборе данных нам необходимо рассмотреть несколько вопросов. Во-первых, необходимо определить, зачем нам нужно собирать данные и для чего эти данные будут использоваться. Во-вторых, нам необходимо рассмотреть тип данных, которые следует собирать: важно, чтобы собранные данные связаны с целью исследования. Наконец, мы должны рассмотреть, как данные должны быть собраны изд. Существует три подхода к сбору данных:

Machine Translated by Google

1. Зарегистрируйтесь

2. Выборка и перепись

3. Экспериментальные исследования

1. Зарегистрируйтесь

Одной из форм доступных данных является зарегистрированная один. Например, есть зарегистрированные данные на рождаемость, смертность, браки, дневную температуру, месячное количество осадков и продажи автомобилей.

Не существует общих статистических методов, которые убедиться, что ценные выводы сделаны из этих данных. Каждый набор данных должен быть рассматривать по существу, и следует будьте осторожны, делая выводы о Население.

2. Выборка и перепись

Методы выборки можно разделить на два категории: случайные методы и неслучайные методы.

Неслучайные методы предполагают построение, эмпирическим путем, выборка , напоминающая генеральную совокупность , из которой она взята, насколько это возможно. Наиболее часто используется неслучайным методом является квотная выборка.

Случайные методы используют вероятностную процедуру для получения выборки из совокупности.

Учитывая тот факт, что вероятность того, что данная единица будет выбрана в выборке, известна, затем можно рассчитать ошибку из-за выборки.

Основными случайными методами являются простая случайная выборка, стратифицированная выборка, систематическая выборка и групповая выборка.

При переписи наблюдают за всеми объектами населения , что дает данные обо всем населения. Ясно, что этот тип исследования очень дорого, когда население очень большой.Поэтому переписи не проводятся очень часто.

3. Экспериментальные исследования

В эксперименте сбор данных осуществляется на основе конкретного эксперимента. ал дизайн. Эти экспериментальные конструкции применяется для сбора данных во всех исследованиях поля.

ИСТОРИЯ

Вероятно, самая старая форма сбора данных восходит к переписям населения, проводившимся в древности.

Антилле и Уджвари (1991) упоминают о существовании должности официального статистика. в Китае во времена династии Чжоу (111–211 гг. ДО Н.Э).

Римский автор Тацит говорит, что император или Август приказал всем солдатам, кораблям и богатство в Империи, чтобы быть сосчитанным.

Доказательства переписей также можно найти в Библия; Святой Лука сообщает, что «Цезарь Август приказал издать декрет, предписывающий перепись всего мира (...), и все отправились в записаться каждый в своем городе».

Форма статистики также может быть найдена во время на этот раз. Название выдает административную происхождения, так как происходит от латинского слова «Статус:» Государство.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Перепись
- План экспериментов
- Выборка
- Опрос

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Антилль, Г., Уджвари, А.: Pratique de la Statis tique Inferentielle. ПАН, Невшатель, Швейцария (1991)

Бурсен, Ж.-Л.: Lesstructures du hasard. Сёй, Париж (1986)

дециль

Децили – это меры положения, вычисляемые на наборе данных.

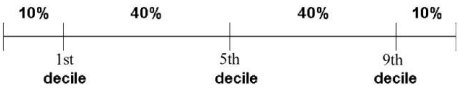
Machine Translated by Google

Децили — это значения , которые делят распределение на десять равных частей, где каждая часть содержит одинаковое количество наблюдений. Дециль является членом более широкого семейства квантилей.

X -й дециль указывает значение, при котором 10x% наблюдений происходят ниже этого значение и (100 10x)% наблюдений выше этого значения. Например, восьмой дециль — это значение, при котором 80% наблюдений находятся ниже этого значения, а над ним.

Пятый дециль представляет собой медиану.

Следовательно, для данного распределения будет девять децилей:



МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Процесс, используемый для расчета децилей, подобен тому, который используется для расчета медианы или квартили.

Когда все необработанные наблюдения доступны

способный, процесс, используемый для расчета децилей, выглядит следующим образом:

1. Организуйте n наблюдений в свободную качественное распределение
2. Децили соответствуют наблюдениям, для которых кумулятивные относительные частоты превышают 10%, 20%, 30%,..., 80%, 90%.

   Некоторые авторы предлагают использовать следующую формулу для точного определения значений различных децилей:

   Вычисление j-го дециля: n+1

   Возьмем i за целую часть j · и 10 k за дробную часть j · $\frac{n+1}{10}$.

   Возьмем xi и xi+1 в качестве значений наблюдений в i-й и (i+1)-й позициях .

ций (когда наблюдения расположены в порядке возрастания).

Тогда j -й дециль равен:

$$D_j = x_i + k \cdot (x_{i+1} \quad x_i) .$$

При группировке наблюдений в классы децили определяются следующим образом: 1. Определяют класс, содержащий желаемый дециль расположен:

- Первый дециль: первый класс, для которого совокупная относительная частота превышает 10%.
- Второй дециль: первый класс, для которого совокупная относительная частота превышает 20%.

  ...
- Девятый дециль: первый класс, для которого кумулятивная относительная частота превышает 90%.

2. Рассчитайте значение дециля на основе гипотеза о том , что наблюдения равномерно распределены в каждом классе:

$$\text{дециль} = L1 + \frac{(N \cdot q) \quad \text{finf fdecile}}{} \cdot \text{в.}$$

где

L1 = нижний предел класса дециля

N = общее количество наблюдений

д = 1/10 для первого дециля

д = 2/10 для второго дециля

...

q = 9/10 для девятого дециля

finf = сумма частот ниже класса дециля

fdecile = частота класса дециля

c = размер интервала класса дециля

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Расчет децилей имеет смысл только для количественной переменной , которая может принимать значения на заданном интервале. Количество наблюдений должно быть относительно высоким, поскольку расчет децилей предполагает деление набора наблюдений на десять частей.

Децили относительно часто используются в упражняться; например, при интерпретации распределение доходов в городе или состояние.

ПРИМЕРЫ

Рассмотрим пример, где децили рассчитывается для частотного распределения непрерывная переменная , где наблюдаемая ции сгруппированы в классы.

Следующая таблица частот представляет прибыль (в тысячах евро) 500 пекарен:

Прибыль (в тысячи евро)	Фре quen cies	Куму поздно частота	Родственник накопленный частота
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
Общий	500		

Первый дециль относится к классу 150–200 (это где кумулятивная относительная частота превышает 10%).

Предполагая, что наблюдения распределены равномерно в каждом классе, получаем следующее значение для первого дециля:

$$\text{1-й дециль} = 150 + \frac{(500 \cdot \frac{1}{10}) \quad 40}{50} \cdot 50$$

$$= 160 .$$

Второй дециль приходится на класс 200–250. Значение второго дециля равно

$$\text{2-й дециль} = 200 + \frac{(500 \cdot \frac{2}{10}) \quad 90}{60} \cdot 50$$

$$= 208{,}33 .$$

Мы можем вычислить другие децили таким же образом. путь, который дает:

дециль	Сорт	Ценить
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

Мы можем заключить, например, что 10% 500 пекарен получают прибыль в размере от 100 000 и 160 000 евро, 50% составляет прибыль ниже 315 000 евро, и 10% получают прибыль от 441670 до 500000 евро.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Мера местоположения
- медиана
- Процентиль
- квантиль
- Квартиль

Machine Translated by Google

Степень свободы

Число степеней свободы равно параметр из распределения хи-квадрат. Этот параметр также используется в других распределения вероятностей , связанных с распределение хи-квадрат, такое как Стьюдент распределение и распределение Фишера.

В другом контексте количество градусов свободы относится к числу линейных независимые члены, участвующие в вычислении суммы квадратов на основе n независимых наблюдений.

ИСТОРИЯ

Термин «степень свободы» был введен Фишером Р.А. в 1925 г.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть Y1, Y2,...,Yn — случайная выборка размер n, взятый из популяции с неизвестное среднее $\bar{Y}$. Сумма отклонений n наблюдений по отношению к их среднему арифметическому всегда равно нулю :

$$\sum_{i=1}^{n} (Y_i - \bar{Y}) = 0 .$$

Это требование является ограничением на каждое отклонение $Y_i - \bar{Y}$, используемое при расчете вариации анс:

$$S^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1} .$$

Это ограничение означает, что n 1 отклонение полностью определяет n-е отклонение. n отклонений (а также сумма их квадраты и дисперсия в S2 выборка), следовательно, имеют n 1 степень свободы.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение хи-квадрат
- Распределение Фишера
- Параметр
- Студенческое распределение
- Дисперсия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Приложения «Студенческих» распределение. Метрон 5, 90–104 (1925)

Деминг, В. Эдвардс

Деминг, У. Эдвардс родился в 1900 г. в г. Су-Сити, штат Айова. Он изучал науки в Вайомингский университет, который окончил в 1921 г. и в Университете Колорадо, где он получил степень магистра математики. и физику в 1924 году. Затем он поступил в Йель Университет, где он получил докторскую степень по физике в 1928 году. После защиты докторской диссертации он проработал десять лет в лаборатории Министерство сельского хозяйства. В 1939 году Деминг перешел в Бюро переписи населения в Вашингтоне. Там он использовал свои теоретические знания, чтобы инициировать первые переписи, которые должны были проводиться выборочным методом. Эти переписи использовали методы, которые позже послужили примером для аналогичные переписи проводились вокруг остальных мира. Он покинул Бюро переписи населения в 1946 г. - стать консультантом в статистическом исследовании и профессор статистики в Нью-Йорке Университет. Он умер в Вашингтоне, округ Колумбия, в 1993.

Долгое время (до 1980 г.) Деминга, мы игнорировались американскими компаниями. В 1947 году Деминг отправился в Токио в качестве консультанта по применению его методов выборки. С 1950 года японцы промышленность приняла теории управления Деминга. В течение десяти лет многие японцы

Д

Machine Translated by Google

продукты экспортировались в Америку, потому что они были лучше и дешевле, чем аналогичные продукты, произведенные в Соединенных Штатах.

Некоторые основные работы и статьи Деминга, В. Эдвардса:

1938 г. Статистическая корректировка данных. Уайли, Нью-Йорк.

1950 Некоторая теория выборки. Уайли, Нью-Йорк.

1960 SampleDesigninBusinessResearch. Уайли, Нью-Йорк.

Демография

Демография – это изучение человеческих популяций. Он включает в себя анализ таких явлений, как рождаемость, смертность, миграция, браки, разводы, коэффициент рождаемости, коэффициент смертности и возрастные пирамиды.
Эти явления рассматриваются как с биологической, так и с социально-экономической точек зрения.
Методы, используемые в демографии, включают в себя высшую математику и статистику.

как многие области социальных наук.

ИСТОРИЯ

Первые демографические исследования относятся к середине XVII века. В этот период Граунт, Джон, английский ученый, проанализировал единственные доступные данные о населении: список смертей в районе Лондона, классифицированный в соответствии с их причиной. В 1662 г. он опубликовал исследование, в котором пытался оценить средний размер семьи, значение миграционного движения и другие элементы, связанные со структурой населения.

В сотрудничестве с Петти, сэром Уильямом, Граунт предложил провести более серьезные исследования должна быть сформирована популяция и создан центр сбора статистических данных .

В течение восемнадцатого века те же типы анализов проводились и постепенно улучшались, но только в девятнадцатом веке некоторые европейские страны, а также Соединенные Штаты предприняли национальные исследования.

переписи и установили регулярные записи рождений, браков и смертей.
Эти отчеты показывают, что разные регионы предлагали разные шансы на выживание для своих

жителей. Эти выводы в конечном итоге привели к улучшению условий труда и гигиены.

Используя собранные демографические данные, стало возможным делать прогнозы, и примерно в конце века появились первые демографические журналы и обзоры. К ним относятся «Демография» в Соединенных Штатах, «Население» во Франции и «Исследования населения» в Великобритании.

В середине XX века демографические исследования стали сосредотачиваться на мировом населении, поскольку важным вопросом стали демографические проблемы стран третьего мира.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Перепись
- Население

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кокс, PR: Демография, 5-е изд. Издательство Кембриджского университета, Кембридж (1976)

Хаузер, П. М., Дункан, О. Д.: Изучение населения: инвентаризация и оценка. University of Chicago Press, Чикаго, Иллинойс (1959)

Machine Translated by Google

Дендрограмма

Дендрограмма – это графическое представление различных агрегаций, полученных в ходе кластерного анализа. Он состоит из узлов, соответствующих группам, и ветвей, представляющих ассоциации, возникающие на каждом этапе. Структура дендрограммы определяется порядок, в котором производятся агрегаты. Если к дендрограмме добавить шкалу , то можно представить расстояния, на которых скопления происходили.

ИСТОРИЯ

В более общем смысле дендрограмма (от греческое «дендрон» — дерево) — дерево диаграмма, иллюстрирующая отношения, которые существуют между членами множества.

Первыми примерами дендрограмм были филогенетические деревья, используемые систематиками. Термин «дендрограмма» кажется впервые использованы в работе

Майра и др. (1953).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

При кластерном анализе набора объектов агрегации достигаются с помощью таблица расстояний и выбранная связь метод (метод одиночной связи, полный способ соединения и др.).
Каждая крайняя точка дендрограммы представляет отдельный класс, созданный автоматическим классификация.

Затем строится дендрограмма путем представления каждой группы узлом, расположенным в определенное положение по отношению к горизонтальной шкале, где положение зависит на расстоянии , на котором совокупность сформирован.

Объекты размещаются на нулевом уровне шкала. Чтобы остановить ответвления, соединяющие узлы не запутываются, выполняется процедура построения дендрограммы систематическим образом.

- Первые два сгруппированных элемента размещены, один над другим, на нулевом уровне шкалы . Тогда горизонтальная линия рисуется рядом с каждым элементом, начиная с нуля уровень до aggregationdistance для каждого элемент. Результирующий класс затем представляется вертикальной линией, соединяющей концы линий для элементов (образующих «узел»). Середина этой вертикальной линии обеспечивает начальную точку для новой горизонтальной линии для этого класса, когда он объединяется с другим классом. Дендрограмма поэтому состоит из ветвей конструкции ed из горизонтальных и вертикальных линий.
- Во время следующий шаг:
  1. Два класса, каждый из которых состоит из одного элемент агрегируется; эти классы помещаются на нулевой уровень, а образовавшийся узел располагается на расстоянии агрегации, как описано выше;
  2. Элемент агрегируется в класс сформировались ранее. Элемент агрегированы с нулевого уровня, и вставляется между существующими классами если необходимо). Узел, представляющий это скопление расположено на (горизонтальном) расстоянии , соответствующем расстояние агрегации между классом и элемент.

Эта процедура продолжается до тех пор, пока не будет достигнуто желаемое получается конфигурация. Если есть еще классы, состоящие из одного элемента в В конце процесса агрегирования они просто добавляются с нулевого уровня дендрограммы. Когда классификация завершена , каждый класс будет соответствовать определенной части дендрограммы.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Дендрограмма описывает упорядоченный путь совокупности операций, выполняемых в течение кластерный анализ. Он иллюстрирует этот тип классификации очень точным образом.
Этот строго определенный подход к построению дендрограммы иногда модифицируется из-за к обстоятельствам. Например, как мы будем см. в одном из приведенных ниже примеров расстояния агрегации двух или более последовательных шаги могут быть одинаковыми, и поэтому процедура затем необходимо изменить, чтобы убедиться, что ветви дендрограммы не перепутываются.

В случае, который мы описали выше, дендрограмма строится горизонтально. Очевидно, что можно построить и дендрограмму. сверху вниз или снизу вверх.
Ветви могут состоять даже из диагональных линии.

ПРИМЕРЫ

Проведем кластерный анализ с помощью метод одной связи и построить соответствующую дендрограмму, иллюстрирующую различные скопления, которые могут возникнуть.
Рассмотрим оценки, полученные пятью учениками во время экзаменов по четырем разным предметам: английскому, французскому, математике и физике.
Мы хотели бы разделить этих пятерых студентов на две группы, используя единую ссылку метод, чтобы проверить новое учение метод.

Эти оценки, которые могут принимать значения от 1 до 6, сведены в следующую таблицу:

	английский	французский	математика	физика
Ален	5.0	3,5	4.0	4,5
Джин	5.5	4.0	5,0	4,5
4.5 марта		4,5	4.0	3,5
Пол	4.0	5,5	3,5	4.0
Пьер	4.0	4,5	3.0	3,5

Таблица расстояний получается путем вычисления евклидовых расстояний между учениками и добавления их в следующую таблицу:

	Ален	Жан	Марк	Поль	Пьер
Ален	0	1,22	1,5	2,35	2
Жан	1,22	0	1,8	2,65	2,74
1,5 марта		1,8	0	1,32	1,12
Пол	2,35	2,65	1,32	0	1,22
Пьер	2	2,74	1,12	1,22	0

Используя метод одной ссылки, основанный на минимальное расстояние между объектами в двух классах следующие разделы получается на каждом этапе:
Первый шаг: Марк и Пьер сгруппированы в расстояние агрегации 1,12: {Ален}, {Жан}, {Марк,Пьер}, {Поль} Второй шаг: Пересматриваем таблицу расстояний. Новый дис таблица приведена ниже:

	Жан	Марк и Пьер	Поль	
Ален	1,22	1,5		2,35
Жан		1,8		2,65
Марк и Пьер				1,22

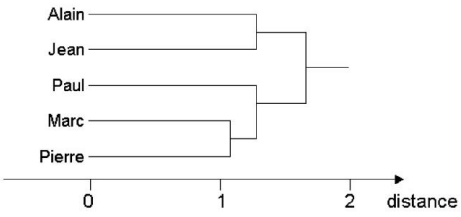
Затем Ален и Жан группируются на расстоянии. из 1.22, и мы получаем следующее разбиение: {Ален, Жан}, {Марк, Пьер}, {Поль}
Третий шаг: дана новая таблица расстояний ниже:

	Марк и Пьер	Поль	
Ален и Жан	1,5		2,35
Марк и Пьер			1,22

Затем Пол объединяется с Марком и Пьером в расстояние 1,22, что дает следующее разделение: {Ален, Жан}, {Марк, Пьер, Поль}
Четвертый шаг: две группы объединяются на расстоянии 1,5.

Это дает следующую дендрограмму:

Machine Translated by Google



Обратите внимание, что Пол должен стоять перед группу Марка и Пьеррена, чтобы избежать перепутывание ветвей диаграммы.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Классификация
- Кластерный анализ
- Полный метод связи
- Расстояние
- Таблица расстояний

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Майр Э., Линсли Э. Г., Юзингер Р. Л.: Методы и принципы систематической зоологии. Макгроу-Хилл, Нью-Йорк (1953)

Sneath, PHA, Sokal, RR: Числовой налог ономия: принципы и практика Численная классификация (серия книги по биологии). WH Фриман, Сан Франциско, Калифорния (1973)

Сокал, Р. Р., Снит, П. Х.: Принципы Числовая таксономия. Фримен, Сан Франциско (1963)

Функция плотности

Функция плотности непрерывной случайной величины позволяет определить вероятность того, что случайная величина X примет значения в заданном интервале.

ИСТОРИЯ

См. вероятность.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим P(a X b), вероятность что непрерывная случайная величина X принимает avalueintheinterval[a, b].Эта вероятность определяется:

$$P(a \quad X \quad b) = \int_a^b f(x)dx \ .$$

где f(x) — функция плотности случайной величины X.

Функция плотности представлена графически на системе координат. Различные значения случайной величины X помещаются на абсцисс, а взятые функцией f размещаются на ординате.
График функции f не позволяет нам определить вероятность для одной конкретной точки, а вместо этого визуализировать вероятность для интервала на поверхности.
Общая поверхность под кривой соответствует значению 1:

$$e(x)\,dx = 1 \quad .$$

Переменная X является непрерывной случайной величиной возможно, если существует неотрицательная функция f , которая определяется для действительных чисел и для которых следующее свойство выполняется для каждого интервала [a, b]:

$$P(a \quad X \quad b) = \int_a^b f(x)\,dx \quad ,$$

Здесь P (a X b) — функция вероятности.
Поэтому вероятность того, что непрерывная случайная величина X примет значение в интервал [a, b] можно получить интегрированием функции вероятности по [a, b].
Существует также следующее условие на f :

$$f(x)\,dx = P(\quad X \quad) = 1 \quad .$$

Мы можем изобразить функцию плотности графически в системе координат. Разные

Machine Translated by Google

по оси абсцисс отложены значения случайной величины X , по оси ординат — функции f .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Если a = b в P(a X b) = $\int_a^b f(x)dx$, то

$$P(X = a) = \int_a^a f(x)dx = 0 .$$

Это означает, что вероятность того, что непрерывная случайная величина принимает изолированный фиксированное значение всегда равно нулю.

Поэтому не имеет значения, будем ли мы включать или исключать границы интервал при расчете вероятности , связанной с интервалом.

ПРИМЕРЫ

Рассмотрим X, непрерывную случайную величину, для которой функция плотности

$$e(x) = \begin{cases} \frac{3}{4}(x2 + 2x), & \text{если } 0 < x < 2 \\ 0, & \text{если не} \end{cases} .$$

Мы можем рассчитать вероятность того, что X больше 1. Мы получаем:

$$P(X > 1) = \int_1 f(x)dx ,$$

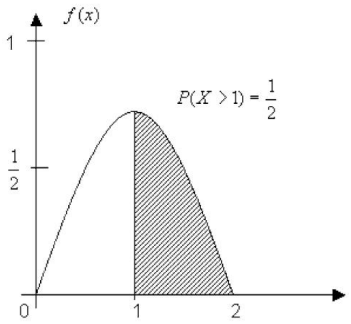
которые можно разделить на:

$$P(X > 1) = \int_1^2 \frac{3}{4}(x2 + 2x)dx + \int_2 0\,dx$$

$$"=" \frac{3}{4}\left[\frac{x3}{3} + x2\right]_1^2$$

$$"=" \frac{1}{2} .$$

Функция плотности и область, в которой P(X > 1) можно представить графически:



Заштрихованная поверхность представляет собой область, где P(X > 1).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция непрерывного распределения
- Непрерывное распределение вероятностей
- Совместная функция плотности
- Вероятность

Зависимость

Понятие зависимости может иметь два значения. Первый касается событий случайного эксперимента. Говорят, что два события являются зависимыми, если наступление одного зависит от наступления другого.

Второе значение слова касается отношения, как правило, функционального отношения, которое может существовать между случайными величинами.

Эту зависимость можно измерить, и в большинстве случаев эти измерения требуют случайной ковариации.

Наиболее часто используемой мерой зависимости является коэффициент корреляции.

нулевая гипотеза, которая утверждает, что две случайные величины X и Y являются независимыми

независимая, может быть проверена против альтернативной гипотезы зависимости. Тесты, которые используются для этой цели, включают критерий независимости хи-квадрат, а также тесты, основанные на ранговой корреляции Спирмена.

Machine Translated by Google

ляционный коэффициент и ранг Кендалла коэффициент корреляции.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Критерий независимости хи-квадрат
- Коэффициент корреляции
- Ковариация
- Коэффициент ранговой корреляции Кендалла
- Коэффициент ранговой корреляции Спирмена
- Тест на независимость

Зависимая переменная

Зависимая переменная (или переменная отклика) в регрессионной модели — это переменная , которая, как считается, изменяется в зависимости от других переменных, называемых независимыми переменными, включенных в анализ.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Независимая переменная
- Множественная линейная регрессия
- Регрессивный анализ
- Простая линейная регрессия

Описательный Эпидемиология

Описательная эпидемиология включает описание частоты и характера заболеваний среди населения.

Он основан на сборе информации о здоровье.

информацию (в таблицах смертности, регистрах заболеваемости, заболеваниях, подлежащих декларированию...) и информацию, которая может оказать влияние на здоровье населения (данные о загрязнении атмосферы, рискованном поведении...), которые используются для получения статистическая картина общего состояния здоровья населения.

ИСТОРИЯ

См. эпидемиологию.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ См.

причину и следствие в эпидемиологии, отношение шансов и шансов, относительный риск, относимый риск, предотвратимый риск, уровень заболеваемости, уровень распространенности.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Несмотря на то, что описательная эпидемиология дает только «элементарную» информацию, она все же жизненно важна: • Она позволяет изучать масштабы и закономерности явлений здоровья (путем изучения их распространенности и частотности), облегчает эпидемиологический надзор,

• Это помогает принимать решения, связанные с планированием и администрированием медицинских учреждений и программ, • Это может привести к гипотезам о факторах риска заболевания (пример: повышенный уровень заболеваемости раком кожи на юге Франции, что, как полагают, связано с повышенной интенсивностью солнечных лучей на юге Франции по сравнению с севером.).

Описательную эпидемиологию нельзя использовать для установления связи между причиной и следствием, это не инструмент прогнозирования.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА См. эпидемиологию.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

См. эпидемиологию.

Описательная статистика

Информация, собранная в ходе исследования, часто может принимать различные формы, например данные о частоте

Machine Translated by Google

(например, количество голосов, отданных за кандидата на выборах) и данные шкалы.

Эти данные часто изначально расположены или организованы таким образом, что их трудно читать и интерпретировать.

Описательная статистика предлагает нам некоторые процедуры, которые позволяют нам представлять данные в удобочитаемой и полезной форме.

Некоторые из этих процедур позволяют нам получить графическое представление данных, например, в следующих формах: • гистограмма, • столбчатая диаграмма, • круговая диаграмма, • стержень и лист, • блочная диаграмма и т. д., в то время как другие позволяют нам получить набор параметров, обобщающих важные свойства исходных данных:

• Среднее значение, • Стандартное отклонение, • Коэффициент корреляции, • Номер индекса и т. д.

Описательная статистика также включает в себя методы анализа данных, такие как анализ соответствия, который заключается в графическом представлении связей между строками и столбцами таблицы сопряженности.

ИСТОРИЯ

Первой известной формой описательной статистики была перепись населения. Переписи устраивались правителями древних цивилизаций, которые хотели пересчитать своих подданных и следить за их профессиями и товарами.

См. статистику и графическое представление.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коробчатый сюжет

Перепись

Анализ корреспонденции

Анализ данных

Дендрограмма

Графическое представление

Гистограмма

Индекс

Линейный график

Круговая диаграмма

Схема стебля и листа

План экспериментов

Планирование эксперимента в некотором роде похоже на его программирование . Каждый фактор , участвующий в эксперименте, может принимать определенное количество различных значений.

(называемые уровнями), а используемый план эксперимента определяет уровни одного или многих факторов (или комбинаций факторов), используемых в эксперименте.

ИСТОРИЯ

Экспериментальные конструкции были впервые использованы в 1920-х годах, в основном в сельскохозяйственной сфере. Сэр Фишер, Рональд Эйлмер был первым, кто использовал математическую статистику при планировании экспериментов. В 1926 году он написал статью, излагающую принципы планирования экспериментов в нематематических терминах.

Федерер В.Т. и Балаам Л.Н. (1973) предоставили очень подробную библиографию литературы по планированию экспериментов до 1969 г., включающую 8000 ссылок.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Цель экспериментального дизайна состоит в том, чтобы найти наиболее эффективный и экономичный метод.

результаты, которые позволяют нам сделать твердые и адекватные выводы по результатам эксперимента.
мент.

Machine Translated by Google

Наиболее часто применяемый экспериментальный дизайны полностью рандомизированы дизайн, рандомизированный дизайн блоков и дизайн латинского квадрата.
Каждая схема подразумевает другой математический анализ, чем те, которые используются для других. дизайны, так как дизайны действительно соответствуют к разным математическим моделям. Примеры из этих типов анализа включают дисперсионный анализ, ковариационный анализ и регрессионный анализ.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Эксперимент
- Модель
- Оптимальный дизайн
- Регрессивный анализ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Федерер, В.Т., Валаам, Л.Н.: библиография. об эксперименте и дизайне лечения до 1968 года. Хафнер, Нью-Йорк (1973).

Фишер, Р.А.: Организация полевых экспериментов. Дж. Министерство сельского хозяйства. 33, 503–513 (1926)

определитель

Любая квадратная матрица A порядка n имеет связанное с ней специальное число, известное как определитель матрицы A.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим A = (aij), квадратную матрицу порядка н. Определитель A, обозначаемый det(A) или |A|, задается следующим сумма:

$$\det(A) = (\pm)a1i \cdot a2j \cdot \cdots \cdot anp ,$$

где суммирование производится по всем перестановкам второго индекса · Знак положителен, если количество инверсий в (i, j,..., r) четно (четная перестановка) и отрицательно, если число нечетное (нечетная перестановка).

Определитель также может быть определен как «развивающиеся» по линии или столбцу. Для например, мы можем разработать матрицу вдоль первая линия:

$$\det(A) = \sum_{j=1}^{n} (-1)^{1+j} \cdot a1j \cdot A1j ,$$

где A1j — определитель квадрата “подматрица” порядка n 1, полученная из A удалив первую строку и j-й столбец.

A1j называется кофактором элемента a1j.
Поскольку мы можем произвольно выбрать линию или столбец, мы можем написать:

$$\det(A) = \sum_{j=1}^{n} (-1)^{я+} \cdot айж \cdot айдж$$

при фиксированном i (развивающемся по i-й линии), или

$$\det(A) = \sum_{я=1}^{n} (-1)^{я+} \cdot айж \cdot айдж$$

при фиксированном j (развивающемся по j-му столбцу).

Этот второй способ определения определителя является рекурсивным, поскольку определитель квадратная матрица порядка n вычисляется с использованием определители матриц порядка n - 1.

Свойства определителя

- Определитель квадратной матрицы A равен равно определителю его транспонированного матрица A :

$$\det(A) = \det(A) .$$

Machine Translated by Google

- Если в матрице A поменять местами две строки (или два столбца) , знак определителя изменится.
- Если все элементы строки (или столбца) равны нулю, определитель также равен нулю. •

Если к другой строке из матрицы A добавить кратное число строк из матрицы A, определитель останется прежним. Это означает, что если

две строки из A идентичны, легко получить строку, в которой все элементы обнулить, просто вычитая одну строку из другой; из-за предыдущего свойства определитель A равен нулю. Такая же ситуация возникает, когда одна строка представляет собой несколько пле другого. •

Определитель матрицы, имеющей нули только под (или над) диагональю, равен произведению элементов на диагонали (такая матрица обозначается как «треугольная»).

- Рассмотрим A и B, две квадратные матрицы порядка n. Определитель произведения двух матриц равен произведению определителей двух матриц.

$$\det(A \cdot B) = \det(A) \cdot \det(B) .$$

ПРИМЕРЫ

1) Рассмотрим A, квадратную матрицу порядка 2:

$$A = \begin{matrix} a & b \\ c & d \end{matrix}$$

$$\det(A) = a \cdot d \quad b \cdot c .$$

2) Рассмотрим B, квадратную матрицу порядка 3:

$$Б = \begin{matrix} 3 & 2 & 1 \\ 4 & 1 & 0 \\ 2 & 2 & 0 \end{matrix} .$$

Развивая по последнему столбцу:

$$\det(B) = 1 \cdot \begin{matrix} 4 & 1 \\ 2 & 2 \end{matrix} + 0 \cdot \begin{matrix} 3 & 2 \\ 2 & 2 \end{matrix} + 0 \cdot \begin{matrix} 3 & 2 \\ 4 & 1 \end{matrix}$$

$$= 1 \cdot \{4 \cdot (\quad 2) \quad (\quad 1) \cdot 2\}$$

$$= \quad 8 + 2$$

$$= \quad 6 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Инверсия
- Матрица
- Перестановка

Отклонение

Понятие отклонения описывает разницу между наблюдаемым значением и фиксированным значением.

значение из множества возможных значений количественной переменной.

Это фиксированное значение часто является арифметическим среднее значение набора значений или медиана.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим множество значений x1, x2,...,xn. Среднее арифметическое этих значений обозначается по $\bar{x}$. Отклонение заданного значения xi относительно среднего арифметического равно:

$$\text{отклонение} = (x_i \quad \bar{x}) .$$

Аналогичным образом, если медиану этих же значений обозначить через m, отклонение значения xi по отношению к медиане равно:

$$\text{отклонение} = (x_i \quad m) .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Среднее абсолютное отклонение
- Среднеквадратичное отклонение
- Дисперсия

Machine Translated by Google

Дихотомическая переменная

Переменная называется дихотомической, если она может принимать только два значения.

Простейшим примером является качественная категориальная переменная «пол», которая может принимать два значения: «мужской» и «женский».

Обратите внимание, что количественные переменные всегда можно сократить и дихотомизировать. Переменная «доход» можно, например, свести к двум категориям: «низкий доход» и «высокий доход».
энуэ».

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Двоичные данные
- Категория
- Качественная категориальная переменная
- Переменная

Дихотомия

Дихотомия — это разделение особей популяции или выборки на две группы в зависимости от заранее определенных критериев.
Переменная называется дихотомической, если она может принимать только два значения. Такие данные называются двоичными данными.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Двоичные данные
- Дихотомическая переменная

Дискретная функция распределения

Функция распределения дискретного ранга переменная dom определена для всех действительных чисел как вероятность того, что случайная величина примет значение , меньшее или равное этому действительному числу.

ИСТОРИЯ

См. вероятность.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

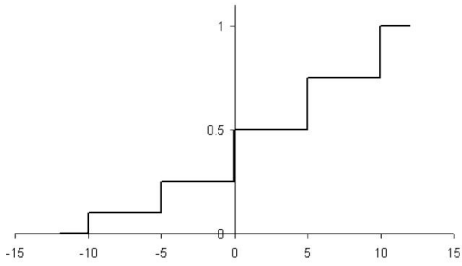
Функция F, определяемая формулой

$$F(b) = P(X \quad b) .$$

называется функцией распределения дис крит случайная величина X.

Таким образом, при заданном действительном числе b функция распределения соответствует вероятности того, что X меньше или равно b.

Функция распределения может быть представлена графически в системе координат. Различные значения дискретной случайной возможные значения X отображаются по оси абсцисс, а совокупные вероятности , соответствующие различным значениям X, F(x), показаны по оси ординат.



В случае, когда возможные значения дискретной случайной величины равны b1, b2,. . . при b1 < b2 < дискретная функция распределения является ступенчатой. F(bi) постоянна на интервале [bi, bi+1[.

Свойства дискретного Функция распределения

1. F — неубывающая функция; другими словами, если a < b, то F(a) F(b)
2. F принимает значения из интервала [0, 1].

Machine Translated by Google

3. $\lim_{b} F(б) = 0.$

4. $\lim_{б} F(б) = 1.$

ПРИМЕРЫ

Рассмотрим случайный эксперимент , состоящий в одновременном бросании двух игральных костей. Рассмотрим дискретную случайную величину X, соответствует сумме баллов от двух

игральная кость.

Найдем вероятность события _ {X 7}, что по определению является значением функция распределения для x = 7.

Дискретная случайная величина X принимает свое значения из множества E = {2, 3,..., 12}.

Вероятности , связанные с каждым значением X , приведены в следующей таблице:

Икс	2	3	4	5	6
П (Икс)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$

Икс	7	8	9	10	11	12
П (Икс)	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

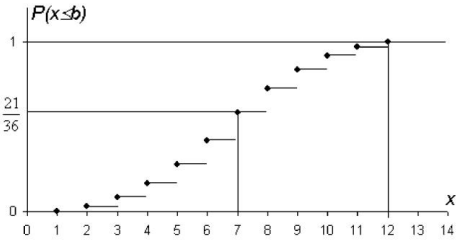
Чтобы установить функцию распределения, должны вычислить, для каждого значения b из X, сумма вероятностей всех значений меньше или равно б. Поэтому мы создаем новая таблица, содержащая кумулятивные вероятности:

б	2	3	4	5	6	7
П (Икс б)	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$

б	8	9	10	11	12	
П (Икс б)	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$	= 1

Вероятность события {X 7 } равна поэтому равно $\frac{21}{36}$.

Мы можем представить функцию распределения дискретной случайной величины X следующим образом минимумы:



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Вероятность
- Функция вероятности
- Случайный эксперимент
- Ценить

Дискретное распределение вероятностей

Если каждое возможное значение дискретного случайного переменная связана с некоторой вероятностью, можно получить дискретную вероятность распределения этой случайной величины.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Распределение вероятности дискретного случайная величина X задается функцией вероятности P(x) или функцией распределения F(x).

В целом его можно охарактеризовать ожидаемое значение:

$$E[X] = \sum_{D} x \cdot P(X = x),$$

и его дисперсия:

$$Вар(X) = \sum_{D} (x - E[X])^2 \cdot P(X = x),$$

где D представляет набор, из которого X может принять его значения.

Machine Translated by Google

ПРИМЕРЫ

Дискретные распределения вероятностей, которые Наиболее часто используются распределение Бернулли, биномиальное распределение, отрицательное биномиальное распределение , геометрическое распределение, полиномиальное распределение, гипергеометрическое распределение и распределение Пуассона.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Распределение Бернулли

Биномиальное распределение

Непрерывное распределение вероятностей
Дискретная функция распределения

Ожидаемое значение
Геометрическое распределение

Гипергеометрическое распределение

Совместная функция распределения вероятностей
Полиномиальное распределение

Отрицательное биномиальное распределение
распределение Пуассона

Вероятность

Распределение вероятностей

Функция вероятности

Случайная переменная

Дисперсия случайной величины

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

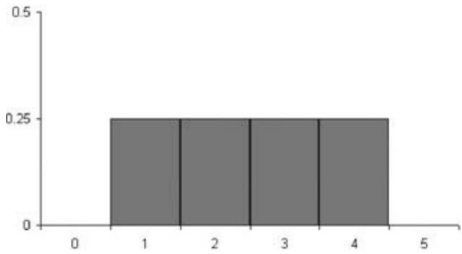
Джонсон, Н.Л., Коц, С.: Распространение в Статистика: дискретные распределения. Уайли, Нью-Йорк (1969)

Дискретная униформа Распределение

Дискретное равномерное распределение критическое распределение вероятностей. Соответствующее непрерывное распределение вероятностей является (непрерывным) равномерным распределением.

Рассмотрим n событий, каждое из которых имеет такая же вероятность P(X = x) = $\frac{1}{n}$; побежал _ переменная dom X следует за дискретной униформой распределение и его функция вероятности :

$$P(X = x) = \begin{cases} \frac{1}{n}, & \text{если } x = x1, x2, \ldots, xn \\ 0, & \text{если не} \end{cases}.$$



Дискретное равномерное распределение, n = 4

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение дискретной униформы распределение по множеству firstnnaturalnum bers по определению определяется как:

$$\begin{aligned} E[X] &= \sum_{x=1}^{n} x \cdot P(X = x) \\ &= \frac{1}{n} \cdot \sum_{x=1}^{n} x \\ &= \frac{1}{n} \cdot (1 + 2 + \cdots + n) \\ &= \frac{1}{n} \cdot \frac{n^2 + n}{2} \\ &= \frac{n + 1}{2}. \end{aligned}$$

Дисперсия этого дискретного равномерного распределения равна:

$$\mathrm{Var}(X) = E\left[X^2\right] - (E[X])^2.$$

С

$$\begin{aligned} E[X^2] &= \sum_{x=1}^{n} x^2 \cdot P(X = x) \\ &= \frac{1}{n} \cdot \sum_{x=1}^{n} x^2 \end{aligned}$$

Machine Translated by Google

$$\text{"="} \frac{1}{n} \cdot 12 + 22 + \cdots + n2$$

$$\text{"="} \frac{1}{n} \cdot \frac{n\,(n+1)\,(2n+1)}{6}$$

$$\text{"="} \frac{2n2 + 3n + 1}{6},$$

и

$$(\text{БЫВШИЙ})^2 \text{"="} \frac{(n+1)2}{22},$$

у нас есть:

$$\text{Вар}(X) = \frac{2n2 + 3n + 1}{} \quad \frac{(n+1)2}{22}$$

$$\text{"="} \frac{6\,n2}{12} \quad \frac{1}{}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Дискретное равномерное распределение часто используется для генерации случайных чисел из любого дискретное или непрерывное распределение вероятностей.

ПРИМЕРЫ

Рассмотрим случайную величину X, оценку получается броском кубика. Если кубик не загружен, вероятность получения любого конкретная оценка равна $\frac{1}{6}$. Поэтому мы иметь:

$$P(X = 1) = \frac{1}{6}$$

$$P(X = 2) = \frac{1}{6}$$

...

$$P(X = 6) = \frac{1}{6}.$$

Количество возможных событий равно n = 6. Мы поэтому есть:

$$P(X = x) = \begin{cases} \frac{1}{6}, & \text{для } x = 1, 2, \ldots, 6 \\ 0, & \text{если не} \end{cases}.$$

Другими словами, случайная величина X подчиняется дискретному равномерному распределению.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дискретное распределение вероятностей
- Равномерное распределение

Рассеивание

См. меру дисперсии.

Расстояние

Расстояние – это числовое описание расстояние между двумя объектами. Дистанция поэтому соответствует действительному числу:

- Ноль, если оба объекта одинаковы;
- Строго положительный, если нет.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим три объекта X, Y и Z, и расстояние между X и Y, d(X, Y).
Расстояние имеет следующие свойства:

- положительно,

$$д(X, Y) > 0,$$

или ноль тогда и только тогда, когда объекты являются такой же

$$d(X, Y) = 0 \quad X = Y.$$

- Он симметричен, что означает, что

$$d(X, Y) = d(Y, X).$$

- Он проверяет следующее неравенство, называемое треугольное неравенство:

$$d(X, Z) \quad d(X, Y) + d(Y, Z).$$

Это говорит о том, что расстояние от одного объекта до другого меньше или равно расстоянию, полученному при прохождении третьего объект.

Рассмотрим X и Y, которые являются двумя векторами с n компонентами,

$$X = (xi) \text{ и } Y = (yi),$$

для i = 1, 2,...,n.

Machine Translated by Google

Семейство расстояний, которое часто используется в такой случай приведен ниже:

$$d(X, Y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \quad c \; p \geq 1 .$$

Эти расстояния называются дисками Минковского. танцы

Расстояние при p = 1 называется абсолютным расстояние или норма L1 .

Расстояние для p = 2 называется Евклидово расстояние, и оно определяется:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} .$$

Евклидово расстояние является наиболее часто используемым расстоянием.

Когда тип используемого расстояния не указано, обычно имеется в виду евклидово расстояние.

Существуют также взвешенные расстояния Минковского, определяемые формулой:

$$d(X, Y) = \sqrt[p]{\sum_{i=1}^{n} w_i \cdot |x_i - y_i|^p} \quad c \; p \geq 1 ,$$

где wi,i = 1,...,n представляют различные веса, сумма которых равна 1.

Когда p = 1 или 2, мы получаем взвешенную абсолютное и евклидово расстояния соответственно.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Для набора из r различных объектов X1, X2,...,Xr расстояние между объектом i и объектом j для каждая пара (Xi, Xj) может быть вычислена. Эти расстояния используются для создания таблицы расстояний или, в более общем смысле, таблица различий, содержащая термины, которые можно обобщить до

$$d_{ij} = d(X_i, X_j) , \quad \text{для } i, j = 1, 2, \ldots, r,$$

используется в методах кластерного анализа. Обратите внимание, что термин мера несходства ty используется, если треугольное неравенство не удовлетворен.

ПРИМЕРЫ

Рассмотрим пример на основе оценок получено пятью студентами в их Экзамены по английскому, французскому, математике и физике.

Эти оценки, которые могут принимать любое значение от 1-6, сведены в следующую таблицу:

	английский	французский	математика	физика
Ален	5.0	3,5	4.0	4,5
Жан	3,5	4.0	5,0	5,5
4.5 марта		4,5	4.0	3,5
Павел	6,0	5,5	5,5	5,0
Пьер	4.0	4,5	2,5	3,5

Евклидово расстояние между Аленом и Жан, обозначаемый d2, определяется как:

$$\begin{aligned} d2(\text{Ален, Жан}) &= \sqrt{(5{,}0 - 3{,}5)^2 + \cdots + (4{,}5 - 5{,}5)^2} \\ &= \sqrt{2{,}25 + 0{,}25 + 1 + 1} \\ &= \sqrt{4{,}5} = 2{,}12 . \end{aligned}$$

Абсолютное расстояние между Аленом и Жан, обозначаемый d1, определяется как:

$$\begin{aligned} d1(\text{Ален, Жан}) &= |5{,}0 - 3{,}5| + \cdots \\ &\quad + |4{,}5 - 5{,}5| \\ &= 1{,}5 + 0{,}5 + 1 + 1 \\ &= 4{,}0 . \end{aligned}$$

Продолжая расчеты, получаем:

$$d2(\text{Ален, Марк}) = \sqrt{2{,}25} = 1{,}5$$

$$d2(\text{Ален, Поль}) = \sqrt{7{,}5} = 2{,}74$$

Machine Translated by Google

d2(Ален, Пьер) = $\sqrt{5,25} = 2,29$

d2(Жан, Марк) = $\sqrt{6,25} = 2,5$

d2(Жан, Поль) = $\sqrt{9} = 3$

d2(Жан, Пьер) = $\sqrt{10,75} = 3,28$

d2(Марк, Поль) = $\sqrt{7,75} = 2,78$

d2(Марк, Пьер) = $\sqrt{2,5} = 1,58$

d2(Поль, Пьер) = $\sqrt{16,25} = 4,03$

d1(Ален, Марк) = 2,5

d1(Ален, Поль) = 5.0

d1(Ален, Пьер) = 4,5

d1(Жан, Марк) = 4,5

d1(Жан, Поль) = 5,0

d1(Жан, Пьер) = 5,5

d1(Марк, Поль) = 6,5

d1(Марк, Пьер) = 2,0

d1(Поль, Пьер) = 7,5 .

Таблица евклидовых расстояний получается помещая эти результаты в следующую таблицу:

	Ален	Жан	Марк	Поль	Пьер
Ален	0	2,12	1,5	2,74	2,29
Жан	2,12	0	2,5	3	3,28
1,5 марта		2,5	0	2,78	1,58
Пол	2,74	3		2,78	0
Пьер	2,29	3,28	1,58	4,03	0

Аналогично можно получить следующую таблицу абсолютных расстояний:

	Ален	Жан	Марк	Поль	Пьер
Ален	0	4	2,5	5	4,5
Жан	4	0	4,5	5	5,5
2.5 марта		4,5	0	6,5	2
Пол	5	5	6,5	0	7,5
Пьер	4,5	5,5	2	7,5	0

Обратите внимание, что порядок близости особей может меняться в зависимости от выбранной расстояние. Например, если используется евклидово расстояние, ближе всех к нему находится Ален. Марк (d2(Ален, Марк) = 1,5), тогда как Алена выходит на второе место, если абсолютный используется расстояние (d1(Ален, Марк) = 2,5), и именно Пьер ближе всех к Марку, с расстояние 2 (d1(Pierre.Marc) = 2).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Расстояние хи-квадрат
- Кластерный анализ
- Таблица расстояний
- Расстояние Махаланобиса
- Мера непохожести

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Абди, Х .: Расстояние. В: Салкинд, Нью-Джерси (ред.) Энциклопедия измерений и статистики. Мудрец, Тысяча дубов (2007)

Таблица расстояний

Для набора из r особей таблица расстояний – квадратная матрица порядка r, а общий член (i, j) равен расстоянию между i -й и j-й индивидуум. Благодаря свойства расстояния, таблица расстояний симметричен, а его диагональ равна нулю.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Предположим, у нас есть r индивидуумов, обозначенных X1, X2,...,Xr; таблица расстояний, составленная из этих r особей, представляет собой квадрат матрица D порядка r и общий член

$$dij = dXi , Xj ,$$

где d Xi, Xj — расстояние между i-й и j-й индивидуум.
Свойства расстояния позволяют утверждать что:

Machine Translated by Google

- Диагональные элементы dii множества D равны нулю поскольку d (Xi, Xi) = 0 (при i = 1, 2, ...,p),
- Несимметричный, т. е. dij = dji для каждого i и для значений j от 1 до r.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Использование таблицы расстояний обеспечивает тип кластерного анализа, целью которого является выявление самые близкие или самые похожие люди в чтобы можно было их сгруппировать.
Расстояния, введенные в таблицу расстояний обычно предполагаются евклидовыми расстояниями, если не указан другой тип расстояния. был указан.

ПРИМЕРЫ

Возьмем пример с оценками получили пять студентов на четырех экзаменах: английский, французский, математика и физика.

Эти оценки, которые могут принимать любое число, кратное 0,5 с 1 по 6, сведены в следующую таблицу:

	английский	французский	математика	физика
Ален 5.0	3,5	4.0	4,5	
Жан 3,5	4.0	5,0	5,5	
4.5 марта	4,5	4.0	3,5	
Пол 6.0	5,5	5,5	5,0	
Пьер 4.0	4,5	2,5	3,5	

Мы пытаемся измерить расстояния , которые отдельные пары учащихся. Для этого мы сравните две строки таблицы (соответствующие двум ученикам) одновременно: если быть точным, каждая оценка во второй строке вычитается из соответствующей оценки в первой строке. линия.

Например, если мы вычисляем расстояние от Алена к Жану, используя евклидову дис тэнс, мы получаем:

$$d(\text{Ален.Жан}) = ((5{,}0 - 3{,}5)^2 + (3{,}5 - 4{,}0)^2 + (4{,}0 - 5{,}0)^2 + (4{,}5 - 5{,}5)^2)^{\frac{1}{2}}$$
$$= \sqrt{2{,}25 + 0{,}25 + 1 + 1}$$
$$= \sqrt{4{,}5} = 2{,}12 .$$

Точно так же расстояние между Аленом и Марк равен:

$$d(\text{Ален, Марк}) = ((5{,}0 - 4{,}5)^2 + (3{,}5 - 4{,}5)^2 + (4{,}0 - 4{,}0)^2 + (4{,}5 - 3{,}5)^2)^{\frac{1}{2}}$$
$$= \sqrt{0{,}25 + 1 + 0 + 1}$$
$$= \sqrt{2{,}25} \text{ знак равно } 1{,}5 .$$

И так далее,

$$d(\text{Ален, Поль}) = \sqrt{7{,}5} = 2{,}74$$
$$d(\text{Ален, Пьер}) = \sqrt{5{,}25} = 2{,}29$$
$$d(\text{Жан, Марк}) = \sqrt{6{,}25} = 2{,}5$$
$$d(\text{Жан, Поль}) = \sqrt{9} = 3$$
$$d(\text{Жан, Пьер}) = \sqrt{10{,}75} = 3{,}28$$
$$d(\text{Марк, Поль}) = \sqrt{7{,}75} = 2{,}78$$
$$d(\text{Марк, Пьер}) = \sqrt{2{,}5} = 1{,}58$$
$$d(\text{Поль, Пьер}) = \sqrt{16{,}25} = 4{,}03 .$$

Организовав эти результаты в виде таблицы, мы получить следующую таблицу расстояний:

	Ален	Жан	Марк	Поль	Пьер
Ален	0	2,12	1,5	2,74	2,29
Жан	2,12	0	2,5	3	3,28
1,5 марта	2,5	0	2,78	1,58	
Пол	2,74	3	2,78	0	4.03
Пьер	2,29	3,28	1,58	4,03	0

Machine Translated by Google

Мы видим, что значения по диагонали равны нулю и что матрица симметрична. Это означает, что расстояние от каждого ученика до одного и того же ученика равно нулю, а расстояние, отделяющее первого ученика от второго, идентично расстоянию что отделяет второе от первого. Итак, d (Xi, Xi) = 0 и d Xi, Xj = d Xj, Xi .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Кластерный анализ
- Дендрограмма
- Расстояние
- Матрица

Функция распределения

Функция распределения случайной величины способ определяется для каждого действительного числа как вероятность того, что случайная величина примет значение , меньшее или равное этому вещественному числу.

В зависимости от того, является ли случайная величина дискретной или непрерывной, мы получаем дискретную функцию распределения или конечную величину. плавная функция распределения.

ИСТОРИЯ

См. вероятность.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Определим функцию распределения ран dom переменная X при значении b (c 0 < b <) по:

$$F(b) = P(X \quad b) .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция непрерывного распределения
- Дискретная функция распределения
- Вероятность
- Случайная переменная
- Ценить

Точечный график

Точечный график — это тип частотного графика. Когда набор данных относительно небольшой (примерно до 25 точек данных), данные и их частоты могут быть представлены в виде точечной диаграммы. Этот тип графического представления особенно полезен для выявления выбросов.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Точечный график строится следующим образом. Устанавливается шкала, содержащая все значения в наборе данных . Каждое данное тогда индивидуально отмечены на графике точкой над соответствующим значением; если есть несколько данных с одинаковым значением, точки выравниваются одна над другой.

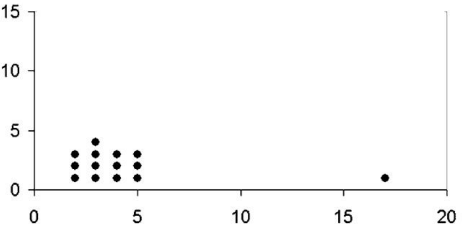
ПРИМЕРЫ

Рассмотрим следующие данные:

5, 2, 4, 5 , 3 , 2 , 4, 3, 5, 2, 3, 4, 3, 17 .

Для построения точечной диаграммы мы устанавливаем шкалу от 0 до 20.

Затем мы представляем данные точками над каждым значением. Мы получаем:



Обратите внимание, что в этом ряду есть выброс, равный 17.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Линейный график

Распределение частоты

Гистограмма

Графическое представление

Тест Дурбина-Ватсона

Тест Дарбина-Ватсона вводит статус
tic d , который используется для проверки автокорреляции
остатки, полученные из модели линейной регрессии. Это проблема, которая часто возникает при применении линейной модели к временному ряду, когда мы хотим проверить независимость полученных таким образом остатков.

ИСТОРИЯ

Дурбин Дж. и Уотсон Г.С. изобрели этот тест в 1950 году.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим случай модели множественной линейной регрессии , содержащей p 1 независимые переменные. В модели написано:

$$Yt = \beta 0 + \sum_{j=1}^{p-1} \beta j Xjt + \varepsilon t , \qquad \tau = 1,\ldots,T,$$

где

Yt — зависимая переменная,

Xjt, j = 1,...,p 1 — независимые переменные

βj, где j = 1,...,p 1 — оцениваемые параметры,

εt при t = 1,...,T является ненаблюдаемой случайный ошибочный термин.

В матричной форме модель записывается так:

$$Y = X\beta + \varepsilon ,$$

где

Y — вектор (n × 1) наблюдений, связанных с зависимой переменной (n наблюдений),

β — вектор (p × 1) оцениваемых параметров,

ε — вектор (n × 1) ошибок,

$$\text{и} X = \begin{matrix} 1 & X11 & \ldots & X1(p\ \ 1) \\ \vdots & \vdots & & \vdots \\ 1 & Xn1 & \ldots & Xn(p\ \ 1) \end{matrix}$$

матрица (n×p) плана, связанная с независимыми переменными.
Остатки, полученные методом наименьших квадратов, имеют вид:

$$e = Y - Y = I - XXX \quad {}^{1} XY \quad .$$

Статистика d определяется следующим образом :

$$\text{д} = \frac{\sum_{\tau=2}^{T} (\text{эт} - \text{эт-1})^{2}}{\sum_{\tau=1}^{T} e_{\tau}^{2}} .$$

Статистика d проверяет гипотезу о том, что ошибки не зависят от альтернативы, согласно которой они следуют авторегрессии первого порядка. процесс:

$$\varepsilon t = \rho \varepsilon t \quad 1 + ut ,$$

где |p| < 1 и theut независимы и нормально распределены со средним значением , равным нулю, и дисперсией σ2 . Мы называем термин ρ автокорреляцией . В терминах ρ проверка гипотезы записывается следующим образом:

H0 : ρ знак равно

0 H1 : ρ знак равно 0 .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Построение этой статистики означает, что
он может принимать значения от 0 до 4. У нас есть

Machine Translated by Google

d = 2 при ρ = 0. Обозначим выборочную оценку наблюдения ρ через ρ. Чтобы проверить гипотезу H0, Дурбин и Ватсон свели в таблицу критические значения d при

уровень значимости 5%; эти критические значения зависят от числа наблюдений T и числа объясняющих переменных (p 1). Таблица для проверки положительной автокорреляции на уровне α, статистический тест d сравнивается с нижним и верхним критическими значениями (dL, dU), и •

отбрасываем H0 : p = 0, если d <

dL • не отклоняем H0 , если d >

dU • при dL < d < dU тест не дает результатов.

Аналогично проверке отрицательной автокорреляции на уровне α статистика 4-d сравнивается с нижними и верхними критическими значениями (dL, dU), и • отклоняется H0 :

p = 0, если H4 d < dL • не отклоняется

H0 , если 4 d > dU • при dL < 4 d < dU тест неокончательный.

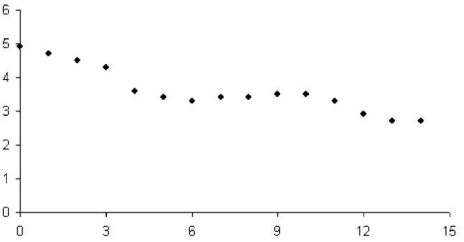
Обратите внимание, что модель должна содержать константу срок, поскольку установленные таблицы Дурбина–Ватсона обычно применяются к моделям с постоянным сроком. Объясняемая переменная не должна быть среди объясняющих переменных.

ПРИМЕРЫ

Рассмотрим пример количества дорожно-транспортных происшествий в США на миллион миль, пройденных автомобилем в период с 1970 по 1984 год. Данные за эти годы приведены ниже:

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.

На следующем рисунке показано, как данные распределяются во времени. По горизонтали по оси ноль соответствует 1970 году, а 15 — 1984 году.



Количество аварий = a + b · t ,

при t = 0, 1,..., 14.

Уравнение для простой линии линейной регрессии, полученное для этих данных, приведено ниже:

Количество аварий = 4,59 0,141 · t.

Тогда мы получим следующие остатки:

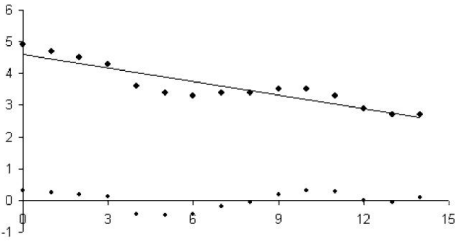
0,31, 0,25, 0,19, 0,13 , -0,43 , -0,49, -0,45, -0,21, -0,07 , 0,17 , 0,32, 0,26, 0,00, -0,06, 0,08.

Значение статистики d равно 0,53. в Таблица Дарбина – Ватсона для одной переменной регрессии, находим значения:

дл,0,05 = 1,08

дЕ,0,05 = 1,36 .

Таким образом, мы имеем случай, когда 0 < d < dL, поэтому мы отвергаем гипотезу H0 в пользу положительной автокорреляции между остатками модели. На следующем рисунке показана положительная корреляция между остатками во времени:



Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Автокорреляция

Множественная линейная регрессия

Остаток

Простая линейная регрессия

Статистика

Временная последовательность

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бурбонне, Р.: Econométrie, manuel et exercices corrigés, 2-е изд. Дюно, Париж (1998)

Дурбин, Дж.: Метод, альтернативный d-критерию. Биометрика 56, 1–15 (1969).

Дурбин, Дж., Уотсон, Г.С.: Тестирование последовательной корреляции в регрессии по методу наименьших квадратов, I. Биометрика 37, 409–428 (1950).

Дурбин, Дж., Уотсон, Г.С.: Тестирование последовательной корреляции в регрессии по методу наименьших квадратов, II. Биометрика 38, 159–177 (1951).

Харви, АС: Эконометрический анализ временных рядов. Филип Аллан, Оксфорд (Уайли, Нью-Йорк) (1981)

Machine Translated by Google

Эконометрика

Эконометрика касается применения статистических методов к экономическим данным.

Экономисты используют статистику для проверки своих теорий.

или делать прогнозы.

Поскольку экономические данные не являются экспериментальными по своей природе и часто содержат некоторую случайность, эконометрика использует скорее стохастические модели .

чем детерминированные модели.

ИСТОРИЯ

Преимущества применения математики и статистики в области экономики

были быстро реализованы. Во второй половине XVII века Петти, сэр Уильям (1676 г.) опубликовал важную статью, в которой представил методологические основы эконометрики.

По словам Яффе, В. (1968), Вальрас, Леон, профессор Лозаннского университета, считается создателем экономической теории общего равновесия, которая обеспечивает теоретическую основу для современной экономики.

метрики.

29 декабря 1930 года в Кливленде

Огайо, группа экономистов, статистиков и математиков основала Экономет.

rics Society для содействия исследованиям в области математических и статистических теорий, связанных

со сферами экономики. Они создали выходящий раз в два месяца обзор «Econometrica».

который впервые появился в январе 1933 г.

Разработка моделей одновременных уравнений , которые можно было бы использовать во временных рядах.

для экономических прогнозов наступил момент

эконометрика возникла как отдельная область, и она остается важной частью эконометрики сегодня.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Множественная линейная регрессия
- Простая линейная регрессия

Временная последовательность

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Жаффе, В .: Вальрас, Леон. В: Силлс, Д.Л. (ред.) Международная энциклопедия социальных наук, том. 16. Macmillan and Free Press, Нью-Йорк, стр. 447–453 (1968).

Петти, В.: Политическая арифметика. В: Уильям Петти, Экономические сочинения. . . , том 1. Келли, Нью-Йорк, стр. 233–313 (1676 г.).

Воннакотт, Р. Дж., Воннакотт, Т. Х.: Экономические показатели. Уайли, Нью-Йорк (1970)

Эджворт, Фрэнсис Ю.

Эджуорт, Фрэнсис Исидро родился в 1845 году в Эджуортстауне (Ирландия) и умер в

Machine Translated by Google

1926. Он участвовал в различных темах, такие как мораль, экономические науки, вероятность и статистика. Его раннее и важное работы над индексами и теорией полезности были позже последовала статистическая работа, которая сейчас иногда называют «байесовским». Вместе с ним возникли термины «модуль» и «флуктуация». Он стал первым редактором Экономического Журнал 1881 года.

Основная работа Эджворта, Фрэнсис Исидро:

1881 г. Математическая физика: очерк Применение математики к Нравственные науки. Кеган Пол, Лондон.

Эфрон, Брэдли

Эфрон, Брэдли родился в Сент-Поле, штат Миннесота, в мае 1938 года в семье Эфрона, Эстер и Эфрон, Майлз, еврейско-русские иммигранты. В 1960 году он получил диплом по математике. В том же году он поступил в Стэнфорд, где он получил докторскую степень. под руководством Миллера, Руперта и Соломона, Херба в отдела статистики. Он занял несколько должности в университете: кафедра статистики, заместитель декана по науке, председатель Консультативного совета университета и председатель Сената факультета. В настоящее время он профессор статистики и биостатистики в Стэнфорде. Университет.

Он был награжден многими премиями, в том числе премией Форда, премией Макартура и медалью Уилкса за исследования в области использование компьютерных приложений в статистике, особенно в отношении начальной загрузки и техники складного ножа. Он является членом Национальной академии наук и Американской академии наук и искусств. стипендия в IMS и ASA. Он также стал президентом Американского статистического Ассоциация в 2004 году.

Термин «компьютерные статистические Методы» возникли у него и Диакониса, П. Его исследовательские интересы охватывают широкий ряд статистических тем.

Некоторые основные работы и статьи Эфрона, Брэдли:

1993 (совместно с Тибширани, Р. Дж.) Введение в Bootstrap. Чепмен & Зал.

1983 Оценка частоты ошибок правила прогнозирования: улучшение перекрестной проверки. Варенье. Стат. доц. 78, 316-331.

1982 Складной нож, бутстрап и прочее Планы передискретизации. СИАМ, Филадельфия, Пенсильвания.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Начальная загрузка
- Повторная выборка

собственное значение

Пусть A — квадратная матрица порядка n. Если есть вектор x = 0, который дает кратное (k · x) сам при умножении на A, то этот вектор называется собственным вектором , а мультипликативный множитель k называется собственным значением матрица A.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть A — квадратная матрица порядка n , а x — ненулевой собственный вектор. k является собственным значением A если

$$A \cdot x = k \cdot x \quad .$$

Machine Translated by Google

т. е. (A k · In) · x = 0, где In – единичная матрица порядка n.

Поскольку x отличен от нуля, мы можем определить возможные значения k , найдя решения уравнения |A k · In| = 0.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Для квадратной матрицы A порядка n определитель |A k · In| — полином степени n по k , имеющий не более n не обязательно уникальных решений. Следовательно, будет не более n собственных значений; некоторые из них могут быть нулями.

ПРИМЕРЫ

Рассмотрим следующую квадратную матрицу третьего порядка:

$$A = \begin{matrix} 4 & 3 & 2 \\ 0 & 1 & 0 \\ 2 & 2 & 0 \end{matrix}$$

$$A - \kappa \cdot I3 = \begin{matrix} 4 & 3 & 2 \\ 0 & 1 & 0 \\ 2 & 2 & 0 \end{matrix} \kappa \cdot \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$

$$A - \kappa \cdot I3 = \begin{matrix} 4 - \kappa & 3 & 2 \\ 0 & 1 - \kappa & 0 \\ 2 & 2 & -\kappa \end{matrix} .$$

Определитель A k · I3 определяется выражением :

$$|A \quad \kappa \cdot I3| = (1 - \kappa) \cdot [(4 - \kappa) \cdot (k) 2 \cdot (2)]$$
$$= (1 k) \cdot (k2 4 \cdot k + 4)$$
$$= (1 - \kappa) \cdot (\kappa - 2) \quad ^2 .$$

Собственные значения получаются путем нахождения решений уравнений (1 k)·(k 2)2 = 0. Находим

$$k1 = 1 \text{ и}$$
$$\kappa 2 = \kappa 3 = 2 \quad .$$

Следовательно, два из трех собственных значений смешаны.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- определитель
- собственный вектор
- Матрица
- Вектор

E

собственный вектор

Пусть A — квадратная матрица порядка n. «Собственный вектор» A — это вектор x , содержащий n компонент (не все нули), где матричное произведение A на x кратно x, например, k · x.

Мы говорим, что собственный вектор x связан с собственным значением k.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Пусть A — квадратная матрица порядка n; тогда вектор x (не нулевой) является собственным вектором A , если существует число k такое, что

$$A \cdot x = \kappa \cdot x \quad .$$

Поскольку x отличен от нуля, мы определяем возможные значения k , находя решения уравнения |A k · In| = 0, где In — единичная матрица порядка n.

Мы находим соответствующие собственные векторы, решая систему уравнений, определяемую формулой:

$$(A \quad k \cdot In) \cdot x = 0 .$$

Каждое ненулевое кратное собственного вектора можно считать собственным вектором. Если с

Machine Translated by Google

ненулевая константа, то

$$A \cdot x = к \cdot x$$
$$c \cdot (A \cdot x) = c \cdot (к \cdot x)$$
$$A \cdot (c \cdot x) = k \cdot (c \cdot x) .$$

Следовательно, c · x также является собственным вектором, если c не ноль.

Собственным вектором можно назвать вектор нормы1 или единичный вектор на оси c·x, а собственный вектор будет лежать на факториальной оси. Этот вектор уникален (с точностью до знака).

ПРИМЕРЫ

Рассмотрим следующую квадратную матрицу третьего порядка:

$$A = \begin{matrix} 4 & 3 & 2 \\ 0 & 1 & 0 \\ 2 & 2 & 0 \end{matrix} .$$

Собственные значения этой матрицы A получаются, если определитель матрицы (A k · I3) равен нулю.

$$A - к \cdot I3 = \begin{matrix} 4 & 3 & 2 \\ 0 & 1 & 0 \\ 2 & 2 & 0 \end{matrix} - к \cdot \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$

$$A - к \cdot I3 = \begin{matrix} 4 - к & 3 & 2 \\ 0 & 1 - к & 0 \\ 2 & 2 & -к \end{matrix} .$$

Определитель A k · I3 определяется выражением:

$$|A \quad к \cdot I3| = (1 \quad k) \cdot [(4 \quad k) \cdot (\quad k) \quad 2 \cdot (\quad 2)]$$
$$= (1 \quad k) \cdot k2 \quad 4 \cdot k + 4$$
$$= (1 - к) \cdot (к - 2) \quad ^2 .$$

Следовательно, собственные значения равны:

$$k1 = 1 \text{ и}$$
$$к2 = к3 = 2 .$$

Теперь вычислим собственные векторы, связанные с собственным значением k1:

$$к1 = 1,$$

$$A \quad k1 \cdot I3 = \begin{matrix} 4 & 1 & 32 \\ 0 & 1 & 1 & 0 \\ 2 & & 2 & 1 \end{matrix}$$

$$"=" \quad \begin{matrix} 33 & 2 \\ 00 & 0 \\ 2 & 2 \quad 1 \end{matrix} .$$

Обозначая компоненты собственного вектора тор x как

$$x = \begin{matrix} x1 \\ x2 \\ x3 \end{matrix} .$$

то нам нужно решить следующую систему уравнений:

$$(A \quad к1 \cdot I3) \cdot x = 0$$

$$\begin{matrix} 33 & 2 \\ 00 & 0 \\ 2 & 2 \quad 1 \end{matrix} \cdot \begin{matrix} x1 \\ x2 \\ x3 \end{matrix} "=" \begin{matrix} 0 \\ 0 \\ 0 \end{matrix} .$$

то есть

$$3x1 + 3x2 + 2x3 = 0$$
$$2x1 + 2x2 \quad x3 = 0 .$$

где второе уравнение исключено из-за его тривиальности.

Решение этой системы уравнений, зависящее от x2, определяется выражением:

$$x = \begin{matrix} 7 \cdot x2 \\ x2 \\ 12 \cdot x2 \end{matrix} .$$

Вектор x является собственным вектором матрицы A для каждого ненулевого значения x2. Чтобы найтиединицу

Machine Translated by Google

собственный вектор, положим:

$$1 = x \quad 2 = (7 \cdot x2)^2 + (x2)^2 + (12 \cdot x2)^2 = (49 + 1 + 144) \cdot x2^2 = 194 \cdot x2^2.$$

откуда $x2 = \frac{1}{\sqrt{194}}$,

и поэтому $x2 = \frac{1}{194} = 0{,}0718$.

Тогда единичный собственный вектор матрицы A , связанный с собственным значением k1 = 1, равен:

$$x = \begin{matrix} 0{,}50 \\ 0{,}07 \\ 0{,}86 \end{matrix}.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- определитель
- собственное значение
- собственный вектор
- Факторная ось
- Матрица

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Додж, Ю.: Математика основы для заливки экономисты. Springer, Берлин Гейдельберг Нью-Йорк (2002)

Мейер, компакт-диск: матричный анализ и прикладные Линейная алгебра, SIAM (2000)

Эпидемиология

Эпидемиология (слово происходит от Греческие слова «эпидемос» и «логос», означающие «учение об эпидемиях») — это изучение частота и распространение болезни, параметры и факторы риска, определяющие состояние здоровья населения по отношению к время, пространство и группы людей.

ИСТОРИЯ

Эпидемиология возникла в XVII в. века, когда Петти, Уильям (1623–1687), английский врач, экономист и ученый собрал информацию о населении, которую описал в работе «Политическая арифметика». Граунт, Джон (1620–1694) предложил первый строгий анализ причин смерти в 1662 г. Кетле, Адольф (1796–1874), бельгийский астроном и математик, считается основоположником современной статистики населения, мать дисциплина эпидемиологии, статистики, эконометрики и другие количественные дисциплины описание социальных и биологических характеристик человеческой жизни. Благодаря работе Фарра, Уильям (1807–1883), эпидемиология была признан отдельной областью от статистики, так как он изучал причины смерти и как она менялась в зависимости от возраста, пола, времени года, место жительства и профессия. Шотландский врач Линд Джеймс (1716–1794) доказал что употребление цитрусовых остановило цингу, которая стал важным этапом в истории эпидемиология. Сноу, Джон (1813–1858), врач в Лондоне, показывал эпидемии размножаются инфекциями, что было подтверждено исследованием in situ, доказавшим, что пить воду Темзы, загрязненную мусором из канализации способствует распространению холеры.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

См. причину и следствие в эпидемиологии, отношение шансов и шансов, относительный риск, относимый риск, предотвратимый риск, заболеваемость частота, распространенность.

ПРИМЕРЫ

Возьмем пример со смертностью в двух фиктивных регионах. Предполагать

Machine Translated by Google

что мы хотим проанализировать ежегодную смертность в двух регионах A и B, каждый из которых населен с 10000 жителей. Данные о смертности указаны в следующей таблице:

	Регион А	Регион Б
Смертей/Всего Население	500/ 10000	1000/ 10000
Общая смертность на год	5%	10%
Годовой коэффициент смертности среди тех, кто < 60 лет	5%	5%
Годовой коэффициент смертности среди тех, кто > 60 лет	15%	15%
% населения < 60 лет	100%	50%

Исходя из этих данных, можем ли мы утверждать, что жители района Б сталкиваются с повышенным вероятность смертности, чем у жителей регион А? Другими словами, будем ли мы рекомендовать людям покидать регион А и жить в Б?

Если мы разделим обе популяции по возрасту, мы обратите внимание, несмотря на то, что разные коэффициенты общей смертности в регионах A и B не кажутся объясняется различной возрастной структурой из А и Б.

Предыдущий пример подчеркивает множество факторов, которые следует учитывать во время анализ данных. Основные факторы, которые следует учитывать при эпидемиологическом исследовании, в том числе: возраст (как абсолютный возраст и как поколение, которому принадлежит человек к), пол и этническая принадлежность, среди прочего.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Атрибутивный риск
- Предотвратимый риск
- Биостатистика
- Причина и следствие в эпидемиологии
- Уровень заболеваемости
- Шансы и отношение шансов
- Уровень распространенности
- Относительный риск

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Лилиенфельд, А.М., Лилиенфельд, Д.Э.: Основы эпидемиологии, 2-е изд. Кларен Дон, Оксфорд (1980)

MacMahon, B., Pugh, TF: Эпидемиология: Принципы и методы. Маленький Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Ротманн, Дж. К.: Эпидемиология. Введение. Издательство Оксфордского университета (2002)

Ошибка

Ошибка – это разница между esti сопряженное значение и истинное значение (или ссылка значение) соответствующего количества.

ИСТОРИЯ

Распределения ошибок — это распределения вероятностей , описывающие ошибку. возникающее при повторных измерениях одно и то же количество в тех же условиях. Они были введены во второй половине восемнадцатого века, чтобы показать, как из среднего арифметического можно получить хорошее приближение к эталонному значению изучаемой величины.

В письме президенту Королевского общества Симпсону Т. (1756 г.) вероятность получения определенной ошибки при наблюдении может быть описана дискретным распределением вероятностей. Он предложил первые два распределения ошибок. Были предложены и другие дискретные распределения. изучал Лагранж, Дж. Л. (1776 г.).

Machine Translated by Google

Симпсон, Т. (1757) также предложил непрерывное распределение ошибок, как и Лаплас, PS (1774 и 1781). Однако самое важное распределение ошибок, нормальное распределение, было предложено Gauss, CF (1809).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В области метрологии можно выделить следующие типы погрешностей: • Абсолютная погрешность, т. е. абсолютная погрешность. лютовое значение разницы между наблюдаемое значение и эталонное значение определенное количество; • Относительная ошибка, представляющая собой отношение между абсолютной ошибкой и значением самого количества. Он характеризует точность физической меры; • Случайная ошибка, связанная не с измерительным инструментом, а с самим экспериментатором; • Экспериментальная ошибка, которая является ошибкой из-за неконтролируемых переменных; • Случайная ошибка, которая представляет собой случайную ошибку, возникающую в результате комбинации ошибок, связанных с прибором и/или пользователь. Изучаются статистические свойства случайных ошибок и их оценки . использование вероятностей; • Остаточная ошибка, которая соответствует разнице между оценкой значение и наблюдаемое значение. Термин для этой ошибки иногда используется остаток ; • Систематическая ошибка, ошибка, возникающая из-за последовательных причин (например, неправильная калибровка измерительного прибора) и всегда происходящая в одном и том же направлении. Предвзятость , используемая в статистика является частным примером этого; • Округленная ошибка, которая возникает, когда числовое значение заменяется усеченным значением, близким к нему. При выполнении расчетов с числа, имеющие n знаков после запятой, находится значение округленной ошибки между

$$\frac{1}{2}\cdot 10^{\text{н и 2}} \quad \text{и} \quad \frac{1}{12}\cdot 10^{\text{н.}}$$

При длительных расчетных процедурах могут накапливаться ошибки округления, что приводит к очень неточным результатам. Этот тип ошибки часто встречается при использовании компьютера.

Каждая ошибка, возникающая в статистической задаче, представляет собой комбинацию различных типов ошибок, перечисленных выше.

В статистике, а точнее во время проверки гипотез, также существуют: • ошибка типа I, соответствующая ошибке, совершаемой при отклонении нулевой гипотезы , когда она верна; • Ошибка типа II, ошибка, совершаемая, когда нулевая гипотеза не отвергается, если она ложна.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Анализ остатков

Предвзятость

Оценка

Проверка гипотезы

Остаток

Ошибка I типа

Ошибка второго рода

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Гаусс, CF: Theoria Motus Corporum Coelestium. Верке, 7 (1809 г.)

Лагранж, Дж. Л.: Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs наблюдения; данс lequel на изучить ле avantages де cette метод пар ле исчисления вероятностей; et où l'on résoud différents problèmes relat-

E

Machine Translated by Google

ifs à cette matière. Разное Тауриненсия 5, 167–232 (1776)

Лаплас, PS de: Mémoire sur la probabi lite des причины par les événements. Мем. акад. Рой. науч. (представлено разными учеными) 6, 621–656 (1774) (или Лаплас, PS de (1891) Œuvres complètes, том 8. Gauthier-Villars, Париж, стр. 27–65).

Лаплас, PS de: Mémoire sur les probabil ités.Mem. акад. Рой. науч. Париж, 227–332 гг. (1781) (или Лаплас, PS de (1891) Œuvres полные, том. 9. Готье-Виллар, Париж, стр. 385–485.)

Симпсон, Т.: Письмо достопочтенному Джордж Эрл Маклсфилд, президент Королевское общество, опираясь на ряд наблюдений в практической астрономии. Филос.Транс.Рой. соц. Лонд. 49, 82–93 (1755)

Симпсон, Т .: Разные трактаты о некоторых Любопытные и очень интересные темы в механике, физико-астрономии и Спекулятивная математика. Норс, Лондон (1757 г.)

Оценка

Оценка – это процедура, которая используется для определяют значение того или иного параметра, связанного с популяцией. Для оценки параметра берется выборка из население и значение оценщика для неизвестного параметра вычисляется. Оценка делится на две большие категории: точечная оценка и интервальная оценка.

ИСТОРИЯ

Понятие оценки восходит к первые работы по математической статистике, особенно Бернулли, Жака (1713 г.), Лаплас, PS (1774) и Бернулли, Даниэль (1778 г.).

Наибольший прогресс в теории оценки после введения наименьшего

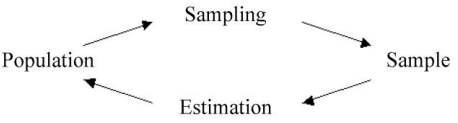
Метод квадратов , вероятно, был формулировкой метода моментов Пирсоном, К. (1894, 1898). Тем не менее, основы теория оценки принадлежит Фишеру Р.А. В своей первой работе 1912 г. он ввел метод максимального правдоподобия .

В 1922 году он написал фундаментальную статью, в которой четко описано, что такое оценка на самом деле в первый раз.

Fisher, RA (1925) также представил набор определений, которые были приняты для описания оценщики. Были введены такие термины, как «предвзятые», «эффективные» и «достаточные» оценки. выведенный им в его теории оценивания.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

На следующем графике показана зависимость между выборкой и оценкой. «Отбор проб» – это процесс получения пробы от населения, в то время как «оценка» — это обратный процесс: от выборки к популяции.



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Доверительный интервал
- Оценщик
- оценка L1
- Балльная оценка
- Надежная оценка
- Образец
- Выборка

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Арматт, М.: Построение понятий. d'estimation et de vraisemblance chez Рональд А. Фишер. В: Майресс, Дж. (ред.) Оценка и зондаж. Экономика, Париж (1988)

Бернулли, Д.: Наиболее вероятный выбор между несколькими противоречивыми наблюдениями, перевод и комментарии (1778 г.). В: Пирсон, Э.С. и Кендалл, М.Г. (ред.) Исследования по истории статистики и Вероятность. Гриффин, Лондон (1970)

Бернулли, Дж.: ArsConjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Базель (1713 г.)

Фишер, Р.А.: Об абсолютном критерии подгонка частотных кривых. Беспорядок. Мат. 41, 155–160 (1912)

Фишер, Р.А.: О математических основах теоретической статистики. Филос. Транс. Рой. соц. Лонд. сер. А 222, 309–368 (1922)

Фишер, Р.А.: Теория статистической оценки. проц. Камб. Филос. соц. 22, 700–725 (1925)

Гаусс, CF: Theoria Motus Corporum Целестий. Верке, 7 (1809 г.)

Лаплас, PS de: Mémoire sur la probabi lite des причины par les événements. Мем. акад. Рой. науч. (представлено разными учеными) 6, 621–656 (1774) (или Лаплас, PS de (1891) Œuvres complètes, том 8. Gauthier-Villars, Париж, стр. 27–65).

Лаплас, PS de: Аналитическая теория вероятности. Курсье, Париж (1812 г.)

Лежандр, AM: Новые методы для определение орбит комет. Курсье, Париж (1805 г.)

Пирсон, К.: Вклад в математическую теорию эволюции. И. В: Карл Ранние статистические статьи Пирсона. Издательство Кэмбриджского университета, Кембридж, стр. 1–40 (1948). Впервые опубликовано в 1894 г. как: О рассечении асимметричных частотных кривых. Филос. Транс. Рой. соц. Лонд. сер. А 185, 71–110

Пирсон, К., Филон, СПГ: математический вклад в теорию эволюции. IV: О возможных ошибках частоты константы и на влияние случайных отбор по вариации и корреляции.Phi лос. Транс. Рой. соц. Лонд. сер. А 191, 229–311 (1898 г.)

E

Оценщик

Любая статистическая функция выборки, используемая для оценка неизвестного параметра населения называется оценкой .

Любое значение , полученное из этой оценки, является оценка параметра населения . _

ИСТОРИЯ

См. оценку.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим θ, неизвестный параметр, определяемый для населения и (X1, X2,...,Xn), выборка , взятая из этой популяции. Статистическая функция (или просто статистика) этого выборка g(X1,...,Xn) используется для оценки θ. Чтобы выделить статистику g(X1,...,Xn), которая является случайной величиной, от значения то, что принимает в конкретном случае, является общим статистику g(X1,...,Xn) назвать оценкой, а значение , которое она принимает в конкретном случае, - оценкой .

Machine Translated by Google

Оценщика не следует путать с
метод оценки.

Параметр популяции обычно _ _
обозначается греческой буквой, а его оценка
этой же буквой, но с циркумфлексом
акцент (ˆ) или шляпу, чтобы отличить ее от
параметр населения. Иногда
вместо него используется соответствующая латинская буква.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

При расчете оценщика, такого как
среднего арифметического, ожидается, что
значение этой оценки приблизится к значению
истинный параметр (в данном случае среднее значение
генеральной совокупности) по мере увеличения размера выборки.
э.

Отсюда ясно, что хороший оценщик должен
быть близким к истинному параметру. Чем ближе
чем оценщики группируются вокруг истинных
параметров, тем лучше оценщики.
В этом смысле хорошие оценщики сильно
связанные со статистическими мерами дисперсии
например дисперсия.

В целом оценщики должны обладать определенным
качества, подобные описанным ниже.

1. Оценщик без смещения:

   Оценка θˆ неизвестного параметра
   эфир θ называется несмещенным, если его
   ожидаемое значение равно θ,

$$E[\hat{\theta}] = \theta .$$

   Среднее $\bar{x}$ и дисперсия S2 выборки
   являются оценками без смещения
   соответственно среднего μ и дисперсии
   σ2 населения , с

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} xi \text{ и}$$

$$S2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (xi - \bar{x})^2 .$$

2. Эффективный оценщик:

   Говорят, что оценка θˆ параметра θ
   быть эффективным, если дисперсия θˆ меньше
   чем дисперсия любой другой оценки
   θ.

   Следовательно, для двух оценок без смещения
   θ, один будет более эффективным, чем
   другой, если его дисперсия меньше.

3. Непротиворечивая оценка:

   Говорят, что оценка θˆ параметра θ
   быть непротиворечивым, если вероятность того, что он
   отличается от θ, уменьшается по мере увеличения размера выборки
   увеличивается.

   Следовательно, оценка непротиворечива, если

$$\lim_{n} P(|\hat{\theta} - \theta| < \varepsilon)\hat{} = 1 ,$$

каким бы малым ни было число ε >ˆ 0.

ПРИМЕРЫ

Рассмотрим (x1, x2 , ..., xn) выборку
из популяции со средним значением μ и
дисперсия σ2 . _
Оценка $\bar{x}$, рассчитанная по выборке и
используемая для оценки параметра μ
населения , является оценкой без смещения.
Затем:

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} xi\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} xi\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[xi]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$= \frac{1}{n} n \cdot \mu$$

$$= \mu .$$

Machine Translated by Google

Для оценки дисперсии σ2 генеральной совокупности заманчиво использовать оценку S2 , рассчитанную по выборке (x1, x2,...,xn), которая определяется здесь как:

$$S2 = \frac{1}{n} \sum_{i=1}^{n} 2 (xi \quad \bar{x}) \cdot$$

Эта оценка является смещенной оценкой σ2, поскольку:

$$\sum_{i=1}^{n} (xi \quad \bar{x})^2 = \sum_{i=1}^{n} (xi \quad \mu)2 \quad n(\bar{x} \quad \mu)2$$

следующим образом:

$$\sum_{i=1}^{n} (xi \quad \bar{x})^2$$

$$= \sum_{i=1}^{n} [(xi \quad \mu) \quad (\bar{x} \quad \mu)]^2$$

$$= \sum_{i=1}^{n} (xi \quad \mu)2 \quad 2(xi \quad \mu)(\bar{x} \quad \mu) + (\bar{x} \quad \mu)2$$

$$= \sum_{i=1}^{n} (xi \quad \mu)2 \quad 2(\bar{x} \quad \mu) \cdot \sum_{i=1}^{n} (\text{кси} - \text{м}) + n(\bar{x} \quad \mu)2$$

$$= \sum_{i=1}^{n} (xi \quad \mu)2 \quad 2(\bar{x} \quad \mu)n \sum_{i=1}^{n} xi \quad n \cdot \mu + n(\bar{x} \quad \mu)2$$

$$= \sum_{i=1}^{n} (xi \quad \mu)2 \quad 2(\bar{x} \quad \mu)(n \cdot \bar{x} \quad n \cdot \mu) + n(\bar{x} \quad \mu)2$$

$$= \sum_{i=1}^{n} (xi \quad \mu)2 \quad 2n(\bar{x} \quad \mu)(\bar{x} \quad \mu) + n(\bar{x} \quad \mu)2$$

$$= \sum_{i=1}^{n} (xi \quad \mu)2 \quad 2n(\bar{x} \quad \mu)2 + n(\bar{x} \quad \mu)2$$

$$= \sum_{i=1}^{n} (xi \quad \mu)2 \quad n(\bar{x} \quad \mu)2 .$$

Следовательно:

$$E\,S2 = E\left[\frac{1}{n} \sum_{i=1}^{n} (xi \quad \bar{x})^2\right]$$

$$= E\left[\frac{1}{n} \sum_{i=1}^{n} (xi \quad \mu)2 \quad (\bar{x} \quad \mu)2\right]$$

$$= \frac{1}{n} E \sum_{i=1}^{n} (xi \quad \mu)2 \quad E(\bar{x} \quad \mu)2$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(xi \quad \mu)2 \quad E(\bar{x} \quad \mu)2$$

$$= \text{Var}(xi) \quad \text{Var}(\bar{x}) \quad \sigma2 = \frac{\sigma2}{n}$$

$$= \frac{n - 1}{n} \sigma2 = \sigma2 .$$

Поэтому, чтобы получить несмещенную оценку дисперсии σ2 , мы избегаем использования S2 и вместо этого используем то же значение, $\frac{n \quad 1}{n}$ деленное на:

$$C^2 = \frac{n}{n - 1} S2$$

$$= \frac{n}{n - 1} \cdot \frac{1}{n} \sum_{i=1}^{n} (xi \quad \bar{x})^2$$

$$= \frac{1}{n - 1} \sum_{i=1}^{n} (xi \quad \bar{x})^2 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Предвзятость

- Оценка
- Образец
- Выборка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Хогг, Р.В., Крейг, А.Т.: Введение в математическую статистику, 2-е изд.
Макмиллан, Нью-Йорк (1965)

E

Machine Translated by Google

Леманн, Э.Л.: Теория точечной оценки, 2-е изд.
Уайли, Нью-Йорк (1983)

Событие

В случайном эксперименте событие является подмножеством множества выборочного пространства. Другими словами, событие — это набор возможных исходов случайного эксперимента.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Учитывать , выборочное пространство случайного эксперимента. Любое подмножество называется событие.

Различные типы событий •

Невозможные события

Невозможное событие соответствует пустому множеству. Вероятность невозможного события равна 0 :

если B = φ , P(B) = 0 .

• Конечное событие

Надежное событие — это событие, соответствующее выборочному пространству. Вероятность достоверного события равна 1 :

P () = 1 .

• Простое событие

Простое событие — это событие, имеющее только один исход.

Операции над событиями

• Отрицание события

Рассмотрим событиеE. Отрицанием этого события является другое событие, называемое противоположным событием E, которое реализуется, когда E не реализуется, и наоборот. Событие, противоположное E, обозначается множеством E^{-}, дополнением к E:

$$\overline{E} = \backslash E = \{\omega \quad ; \omega \; / \quad E\} .$$

• Пересечение событий

Рассмотрим два события E и F. пересечение этих двух событий другое событие, которое реализуется, когда оба события реализуются одновременно. Оно называется событием «E и F» и обозначается E F.

E F = {ω ; ω E и ω F} .

В более общем смысле, если E1, E2,...,Er являются r событиями, то

$$\bigcap_{i=1}^{r} E_i = \{\omega \quad ; \omega \quad E_i, \quad i = 1,...,n\}$$

определяется как пересечение r событий, которое реализуется, когда все r событий реализуются одновременно. • Объединение

событий. Рассмотрим два одинаковых события E и F. Объединение этих двух событий — еще одно событие, которое реализуется, когда реализуется хотя бы одно из двух событий E или F. Иногда его называют событием «E или F» и обозначают E F:

E F = {ω ; ω E или ω F} .

Свойства операций над событиями.

Рассмотрим три события E, F и G. Пересечение и объединение этих событий характеризуют следующие свойства: • Коммутативность:

E F = F E ,

E F = F E .

• Ассоциативность:

E (F G) = (E F) GE (F ,

G) = (E F) G .

Machine Translated by Google

• Идемпотентность:

$$E \quad E = E \quad ,$$
$$E \quad E = E \quad .$$

Дистрибутивность пересечения относительно объединения:

$$E \quad (F \quad G) = (E \quad F) \quad (E \quad G) .$$

Дистрибутивность объединения относительно пересечения:

$$E \quad (F \quad G) = (E \quad F) \quad (E \quad G) .$$

К основным отношениям отрицания относятся:

$$\overline{E \quad F} = \overline{E} \quad \overline{F} \quad ,$$
$$\overline{E \quad F} = \overline{E} \quad \overline{F} \quad .$$

ПРИМЕРЫ

Рассмотрим случайный эксперимент , в котором одновременно бросают желтую и синюю кости.

Выборочное пространство этого эксперимента формируется из набора пар возможных результатов двух игральных костей:

$$= \{(1, 1), (1, 2), (1, 3),..., (2, 1) ,$$
$$(2, 2),..., (6, 5), (6, 6)\}.$$

где, например: (1, 2) = «1 на желтом кубике и «2» на синем кубике. Число

Количество пар (или простых событий), которые могут быть образованы, определяется количеством аранжировок с повторением двух объектов из шести, что равно 36. Для этого выборочного пространства можно описать, например,

следующие события: 1. Сумма баллов равна шести:

$$A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}.$$

2. Сумма баллов четная:

$$B = \{(1, 1), (1, 3), (1, 5),$$
$$(2, 2), (2, 4),..., (6, 4), (6, 6)\}.$$

3. Сумма баллов меньше 5:

$$C = \{(1, 1), (1, 2), (1, 3),$$
$$(2, 1), (2, 2), (3, 1)\}.$$

4. Сумма баллов нечетная:

$$D = \{(1, 2), (1, 4), (1, 6), (2, 1),$$
$$(2, 3),..., (6, 5)\} = B \quad .$$

5. Сумма баллов четная и меньше 5:

$$E = \{(1, 1), (1, 3), (2, 2), (3, 1)\}$$

знак равно B C .

6. Синий кубик останавливается на «2»:

$$F = \{(1, 2), (2, 2), (3, 2), (4, 2),$$
$$(5, 2), (6, 2)\}.$$

7. Синий кубик приземляется на «2» или на общий счет. меньше 5:

$$G = \{(1, 2), (2, 2), (3, 2), (4, 2), (5, 2),$$
$$(6, 2), (1, 1), (1, 3), (2, 1), (3, 1)\}$$

знак равно F C .

8. A и Г несовместимы.

9. A подразумевает Б.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Совместимость
- Условная возможность
- Независимость
- Вероятность
- Случайный эксперимент
- Образец пространства

E

Machine Translated by Google

Ожидаемое значение

Ожидаемое значение случайной величины представляет собой средневзвешенное значение значений , может принимать случайная величина, где вес вероятность того , что конкретное значение принимается случайной величиной.

ИСТОРИЯ

Математический принцип ожидаемой стоимости впервые появился в работе под названием De Ratiociniis в Aleae Ludo, опубликованном в 1657 г. голландским ученым Гюйгенсом Кристианом (1629–1695). Его мысли на эту тему, по-видимому, сильно повлияли на более поздние работы Паскаля и Ферма по теории вероятностей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В зависимости от того, является ли случайная величина дискретной или непрерывной, мы называем либо ожидаемое значение дискретной случайной величины, либо ожидаемое значение непрерывной случайной величины.

Рассмотрим дискретную случайную величину X у которого есть функция вероятности p(X). ожидаемое значение X, обозначаемое E[X] или μ, определяется

$$\mu = E[X] = \sum_{i=1}^{n} x_i p(x_i)$$

если X может принимать n значений.

Если случайная величина X непрерывна, то ожидаемое значение становится

$$\mu = E[X] = \int_D x f(x)\,dx ,$$

если X принимает значения на интервале D, где f(x) — функция плотности X.

Свойства ожидаемого значения

Рассмотрим две константы a и b, a случайная величина X. Тогда мы имеем:

1. E[aX + b] = aE[X] + b;
2. E[X + Y] = E[X] + E[Y];
3. E[X - Y] = E [X] - E [Y];
4. Если X и Y независимы, то:

$$E[X \cdot Y] = E[X] \cdot E[Y] .$$

ПРИМЕРЫ

Мы рассмотрим два примера, один из которых относится к дискретной случайной величине, а другой — к непрерывной случайной величине.

Рассмотрим азартную игру, в которой кубик брошено и счет отмечен. Предположим, что вы выиграть евро, если выпадет четное число, два евро, если он приземлится на «1» или «3», и проиграет три евро, если выпадет 5.

Случайная величина X , рассматриваемая здесь, равна количество выигранных или проигранных евро.

В следующей таблице представлены различные значения X и их соответствующие вероятности:

Икс	3	1	2
П (Икс)	1/6	3/6	2/6

Таким образом, ожидаемое значение равно:

$$E[X] = \sum_{i=1}^{3} x_i p(x_i)$$

$$= 3 \cdot \frac{1}{6} + 1 \cdot \frac{3}{6} + 2 \cdot \frac{2}{6}$$

$$= \frac{2}{3} .$$

Другими словами, игрок в среднем выигрывает $\frac{2}{3}$ евро за бросок.

Рассмотрим непрерывную случайную величину X со следующей функцией плотности:

$$e(x) = \begin{cases} 1 \text{ для } 0 < x < 1 \\ 0 \text{ в другом месте.} \end{cases}$$

Machine Translated by Google

Ожидаемое значение этой случайной величины X равно:

$$E[X] = \int_0^1 x \cdot 1 \, dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция плотности
- Функция вероятности
- Средневзвешенное арифметическое

Эксперимент

Эксперимент – это операция, проводимая в контролируемых условиях с целью охватить ранее неизвестный эффект, проверить или установить гипотезу или продемонстрировать известный закон.

Цель эксперимента — выяснить связь между контролируемыми условиями. и результат эксперимента. Экспериментальный анализ проводится по наблюдениям, на которые влияют не только контролируемые условия, но и неконтролируемые условия и погрешности измерений.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Когда возможные результаты эксперимента могут быть описаны и каждому возможному результату можно приписать вероятность реализации, эксперимент называется случайным экспериментом. Множество всех возможных исходов эксперимента называется выборочным пространством. Факторный эксперимент — это когда экспериментатор организует эксперимент с двумя или более факторами. Факторы – это контроль лабильные условия эксперимента. Опыт умственные ошибкипроисходят из неконтролируемого состояния и от ошибок измерения , которые присутствует в любом типе эксперимента. Экспериментальный план устанавливается в зависимости от цели экспериментатора или возможных доступных ресурсов.

ПРИМЕРЫ

См. случайный эксперимент и факторный эксперимент.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- План экспериментов
- Факторный эксперимент
- Проверка гипотезы
- Случайный эксперимент

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Box, GEP, Hunter, WG, Hunter, JS: Статистика для экспериментаторов. Введение в проектирование, анализ данных и построение моделей. Уайли, Нью-Йорк (1978)

Рюэгг, А.: Вероятности и статистика. Press es Polytechniques Romandes, Лозанна, Швейцария (1985)

Экспериментальная установка

Экспериментальная единица – это наименьшая часть экспериментального материала, к которой мы применяем уход.

План эксперимента определяет количество экспериментальных единиц, подвергаемых различным обработкам.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- План экспериментов
- Эксперимент
- Фактор
- Уход

Machine Translated by Google

Исследовательский анализ данных

Исследовательский анализ данных – это подход к анализу данных, при котором функции и характеристики данных рассматриваются с «открытый ум»; другими словами, без попытка применить ту или иную модель к данные. Часто используется при первом контакте. с данными, прежде чем какие-либо модели были выбран для структурного или стохастического анализа компонентов, а также используется для поиска отклонений от общепринятых моделей.

ИСТОРИЯ

Исследовательский анализ данных представляет собой набор методов, разработанных в основном Тьюки, Джон Уайлдерс, с 1970 г. Философия, лежащая в основе этого подхода, заключается в изучении данные перед применением конкретной вероятностной модели. По словам Тьюки, JW, Исследовательский анализ данных подобен детективной работе. При исследовательском анализе данных эти подсказки могут быть числовыми и (очень часто) графический. Действительно, Тьюки ввел несколько новых полуграфических представлений данных. инструменты, помогающие в исследовательском анализе данных, включая «график с коробкой и усами» (также известный каккоробочный сюжет) в 1972 году, а стебель и диаграмма листьев в 1977 году. Эта диаграмма похож на гистограмму, которая датируется восемнадцатого века.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Хороший способ обобщить важную информацию в данных с исследовательским анализом данных обеспечивается сводкой из пяти чисел, которая представлена в виде таблицы следует:

медиана	
Первый квартиль	Третий квартиль
Минимум	Максимум

Пусть n будет количеством наблюдений; мы расположите данные в порядке возрастания. Мы определять:

$$\text{средний ранг} = \frac{n+1}{2}$$

$$\text{квартильный ранг} = \frac{\text{средний ранг} + 1}{2},$$

где x - значение x , усеченное вниз до следующего наименьшего целого числа. В своей книге Тьюки Д.В. называет первое и четвертые квартили «шарниры».

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В исследовательском анализе данных, определяемом Тьюки, Дж. В., есть четыре основные темы. Это графическое представление, повторное выражение (которое является просто преобразованием переменных), остатки и сопротивление (что является синонимом понятия прочности). Сопротивление является мерой чувствительности анализа или резюме к «плохие данные». Необходимость изучения сопротивления отражает тот факт, что даже «хорошие данные редко содержит менее 5% ошибок и поэтому важно иметь возможность защитить анализ от неблагоприятные последствия ошибки». Линия сопротивления Тьюки дает надежное соответствие набору точек, это означает, что эта линия не слишком чувствительна к какому-либо конкретному наблюдению. Медиана очень устойчива, а средняя нет.

Графические представления используются для анализа поведения данных, подбора данных, диагностических мер и остатков. Таким образом мы может обнаружить любые неожиданные характеристики и узнаваемая закономерность в данных.

Разработка разведочных данных Анализ тесно связан с упором на визуальную репрезентацию и использованием множество относительно новых графических методов, таких как диаграмма стебля и листа и коробочный сюжет.

Machine Translated by Google

ПРИМЕРЫ

Индексы доходов швейцарских кантонов в 1993 г.

Кантон	Индекс	Кантон	Индекс
Цюрих	125.7	Шаффхаус	99,2
Берн	86.2	Аппенцелль Прав.-Внешн.	84,3
Люцерн	87,9	Аппенцелль Rh.-Int.	72,6
Ури	88.2	Сен-Галль	89,3
Швиц	94,5	Граубюнден	92,4
Обвальд	80.3	Аргови	98,0
Нидвальд	108.9	Тургови	87,4
Гларис	101.4	Тессин	87,4
Зуг	170,2	Во	97,4
Фрибург	90,9	Вале	80,5
Солер	88.3	Невшатель	87,3
Базель-Штадт	124.2	Женева	116,0
Базель-Ланд	105.1	Юра	75,1

В таблице представлены показатели доходов Швейцарские кантоны на душу населения (Швейцария = 100) в 1993 году. Мы можем рассчитать сводную статистику пяти чисел:

$$\text{средний ранг} = \frac{n+1}{2} \text{ "=" } \frac{26+1}{2} = 13{,}5 .$$

Таким образом, медиана Md является средним значением тринадцатое и четырнадцатое наблюдения:

$$\text{Мд} = = \frac{89{,}3 + 90{,}9}{2} \text{ 90,1 .}$$

Затем вычисляем:

$$\text{квартильный ранг} = \frac{13+1}{2} = 7 \quad .$$

Первый квартиль будет седьмым наблюдением от начала данных (при расположении в порядке возрастания) и третий квартиль будет быть седьмым наблюдением с конца данные:

$$\text{первый квартиль} = 87{,}3$$

$$\text{третий квартиль} = 101{,}4 .$$

Минимум и максимум составляют соответственно: 72,6 и 170,2.

Это дает нам следующую пятизначную сумму Мэри:

90,1	
87,3	101,4
72,6	170,2

E

Сопротивление (или надежность) медианы по отношению к среднему

Считаем следующие числа: 3, 3, 7, 7, 11, 11.

Сначала мы вычисляем среднее значение этих чисел:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{"="} \frac{3+3+7+7+11+11}{6} = 7 \quad .$$

Затем вычисляем медиану Md:

$$\text{средний ранг} = \frac{n+1}{2} \text{ "=" } \frac{6+1}{2} = 3{,}5$$

$$\text{Мд} = = \frac{7+7}{2} \quad .$$

Отметим, что в этом случае среднее и медиана обоих равна 7.

Теперь предположим, что мы добавляем еще одно число: 1000. Поэтому мы пересчитываем среднее и медиана следующих чисел: 1000, 3, 3, 3, 7, 11, 11. Среднее значение:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{"="} \frac{1000+3+3+7+7+11+11}{7}$$

$$= -136{,}86 .$$

Понятно, что наличие даже одного «плохого datum» может в значительной степени повлиять на среднее значение. Медиана:

$$\text{средний ранг} = \frac{n+1}{2} \text{ "=" } \frac{7+1}{2} = 4$$

$$\text{Мд} = 7 \quad .$$

Machine Translated by Google

В отличие от среднего, медиана не
изменение, которое говорит нам, что медиана
более устойчивы к экстремальным значениям («плохие данные»)
чем среднее.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коробчатый сюжет

Графическое представление
Остаток

Схема стебля и листа
Трансформация

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

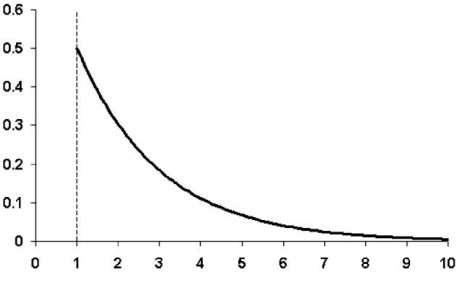
Tukey, JW: Некоторые графические и полуграфические
дисплеи. В: Бэнкрофт, Т.А. (ред.) Статистические
статьи в честь Джорджа У.
Снедекор. Издательство государственного университета Айовы,
Эймс, ИА, стр. 293–316 (1972).

Тьюки, Дж. В.: Исследовательский анализ данных.
Аддисон-Уэсли, Рединг, Массачусетс (1977)

экспоненциальный
Распределение

Случайная величина X подчиняется экспоненциальному
распределению с параметром θ, если ее функция
плотности определяется выражением:

$$e(x) = \begin{cases} \theta \cdot e^{\theta x}, \text{ если } x \geq 0; \theta > 0 \\ 0 \quad \text{если не} \end{cases}.$$



Экспоненциальное распределение, θ = 1, σ = 2

Экспоненциальное распределение также называют
отрицательная экспонента.
Экспоненциальное распределение является непрерывным
распределение вероятностей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Функция распределения случайного _
переменная X , которая следует экспоненциальному
распределению с параметром θ, выглядит следующим образом:

$$\Phi(x) = \int_0^{x} \theta \cdot e^{-\theta t}\, dt = 1 - e^{-\theta x}.$$

Ожидаемое значение определяется по формуле:

$$E[X] = \theta \frac{1}{_}.$$

Поскольку у нас есть

$$E\,X2 = \theta\, 2 \frac{2}{_}.$$

дисперсия равна :

$$\mathrm{Var}(X) = E\,X2 - (E[X])^2 = \frac{2}{2_} - \frac{1}{2_} = \frac{1}{2_}.$$

Экспоненциальное распределение – это
непрерывное распределение вероятностей, аналог
геометрическое распределение.
На самом деле это частный случай гаммы
распределения, где параметр α
гамма-распределение равно 1.
Оно становится распределением хи-квадрат с
две степени свободы , когда параметр θ
экспоненциального распределения равен
до 2^{-1}.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Экспоненциальное распределение используется для
описывать случайные события во времени. Для экзамена
Например, продолжительность жизни является
характеристикой, которая часто представляется
экспоненциальной случайной величиной.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение хи-квадрат
- Непрерывное распределение вероятностей
- Гамма-распределение
- Геометрическое распределение

Экстраполяция

Со статистической точки зрения экстраполяция — это оценка зависимой переменной для значений независимой переменной, находящихся за пределами набора наблюдений.

Экстраполяция часто используется во временных рядах , когда модель , определяемая из значения u1, u2,...,un , наблюдаемые в моменты времени t1, t2,...,tn (при t1 < t2 < < tn) , используется для предсказания значения переменной в момент времени tn+1.

Как правило, простая функция линейной регрессии , основанная на наблюдениях (Xi, Yi), i = 1, 2,...,n, может использоваться для оценки значений Y для значений X , которые расположены за пределами множество X1, X2,...,Xn. Также возможно получить экстраполяции из модели множественной линейной регрессии.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Требуется осторожность при использовании экстраполяций, так как они получаются на основе гипотезы о том, что модель не меняется при значениях независимой переменной, используемой для экстраполяции.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Прогнозирование
- Множественная линейная регрессия
- Простая линейная регрессия

Временная последовательность

E

Machine Translated by Google

Ф

Фактор

Фактор — это термин, используемый для описания любого контролируемого условия в эксперименте. Каждый фактор может принимать заранее определенное число значений, называемых уровнями фактора. Эксперименты, в которых используются факторы, называются факторными экспериментами.

ПРИМЕРЫ

Социолог хочет узнать мнение жителей города о проблеме наркотиков. Его интересует влияние пола и дохода на отношение к этому вопросу.

Уровни фактора «пол»: мужской (M) и женский (Ж); уровни фактора «доход» соответствуют низкому (H), среднему (A) и высокому (H), где деление между этими уровнями задается опытным путем.
наставник

Факторное пространство G этого эксперимента состоит из следующих шести комбинаций уровней двух факторов :

$$G = \{(M,L);(M,A);(M,H);(F,L);$$
$$(F,A);(F,H)\} .$$

В данном примере фактор «пол» является качественным и принимает номинальные значения; тогда как фактор «доход» является количественным и принимает порядковые значения.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- План экспериментов
- Эксперимент
- Факторный эксперимент
- Уход

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Ракто, Б.Л., Хедаят, А., Федерер, В.Т.: Факторные планы. Уайли, Нью-Йорк (1981)

Уровень фактора

Уровни фактора – это разные значения что это может занять. План эксперимента определяет, какие уровни из всех возможных комбинаций таких уровней используются в эксперименте .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- План экспериментов
- Эксперимент
- Фактор
- Уход

Факторная ось

Рассмотрим S, квадратную матрицу порядка n. Если Xi — собственный вектор S , связанный с собственным значением ki, мы говорим, что Xi — это i-й фактор.

Machine Translated by Google

торическая ось; точнее, это ось , несущая вектор Xi.

Обычно мы располагаем собственные значения в порядке убывания, что объясняет необходимость нумерации осей факториала.

ИСТОРИЯ

См. анализ соответствия.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Понятие факториальной оси используется в анализе соответствий.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Анализ корреспонденции
- собственное значение
- собственный вектор
- Инерция

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

См. анализ соответствия.

Факторный эксперимент

Факторный эксперимент — это эксперимент , в котором изучаются все возможные методы лечения , которые могут быть получены из двух или более факторов, где каждый фактор имеет два или более уровней, таким образом, чтобы можно было исследовать основные эффекты и взаимодействия . Термин «факторный эксперимент» описывает эксперимент, в котором все различные факторы комбинируются всеми возможными способами, но не определяет план эксперимента, используемый для проведения такого эксперимента.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В факторном эксперименте можно использовать модельное уравнение , чтобы связать зависимую переменную с независимой переменной.

Следовательно, если в факторном эксперименте участвуют два фактора, ассоциированная модель выглядит следующим образом:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

$i = 1, 2, \ldots, a$ (уровни фактора A),

$j = 1, 2, \ldots, b$ (уровни фактора B),

$k = 1, 2, \ldots, n_{ij}$ количество бывших экспериментальные единицы , получающие лечение ij ,

где:

- μ – общее среднее, общее для всех обработок,
- α_i – эффект i -го уровня фактора A,
- β_j — эффект j-го уровня фактора B,
- $(\alpha\beta)_{ij}$ — эффект взаимодействия i -го уровня A j-го уровня B,
- ε_{ijk} — экспериментальная ошибка наблюдения Y_{ijk}.

В общем случае μ — константа, а ε_{ijk} — случайная величина, распределенная согласно нормальному распределению со средним значением , равным нулю, и дисперсией σ^2 .

Если член взаимодействия $(\alpha\beta)_{ij}$ опущен, получается аддитивная модель , означающая, что влияние каждого фактора является константой, которая складывается с общим средним μ, и это не зависит от уровня другого фактора. Это известно как модель без взаимодействия. ции.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В факторном эксперименте, где все факторы имеют одинаковое число уровней, количество применений, используемых в эксперименте, обычно определяется числом уровней, возведенных в степень, равную числу

Machine Translated by Google

факторов. Например, для эксперимента
в котором используются три фактора, каждый из которых
имеет два уровня, этот эксперимент известен как 23-
факторный эксперимент, в котором используется восемь
воздействий.

Наиболее часто используемое обозначение для
комбинации методов лечения , используемых на самом деле
Экспериментальный эксперимент можно описать следующим образом.
Используйте заглавную букву для обозначения фактора
и числовой индекс для обозначения его уровня.
Рассмотрим, например, факторный эксперимент.
с тремя факторами: фактор A с двумя уровнями,
фактор B с двумя уровнями и фактор C с
три уровня. 12 комбинаций будут:

A1B1C1 , A1B1C2 , A1B1C3 ,
A1B2C1 , A1B2C2 , A1B2C3 ,
A2B1C1 , A2B1C2 , A2B1C3 ,
A2B2C1 , A2B2C2 , A2B2C3 .

Иногда только индексы (написанные в
в том же порядке, что и факторы) формулируются:

111 , 112 , 113 ,
121 , 122 , 123 ,
211 , 212 , 213 ,
221 , 222 , 223 .

Двухфакторные эксперименты или, в более общем
смысле, многофакторные эксперименты
важно по следующим причинам:

- Двухфакторный эксперимент использует
  ресурсы более эффективно, чем два
  эксперимента, каждый из которых использует один фактор.
  Первый занимает меньше времени и требует
  меньше экспериментальных единиц для достижения
  заданный уровень точности.
- Двухфакторный эксперимент позволяет
  влияние одного фактора, изучаемого на
  каждый уровень другого фактора, потому что
  уровни обоих факторов различны.
  к выводам, которые справедливы на более широком
  диапазона экспериментальных условий, чем если бы
  используется серия однофакторных планов.
- Наконец, одновременное исследование двух
  факторы необходимы при взаимодействии
  между факторами присутствуют, что означает
  что действие фактора зависит от
  уровень другого фактора.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

План экспериментов

Эксперимент

Ф

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

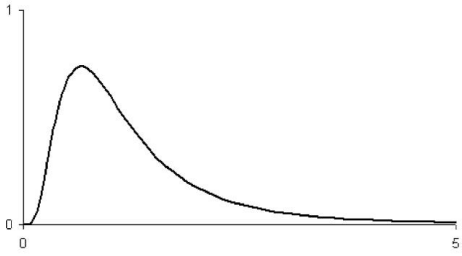
Додж, Ю.: Основные планы опыта.
В: Эшлиманн, Дж., Бонжур, К., Стокер, Э. (ред.)
Методологии и методы.
de plans d'expériences: 28ème Cours de
perfectionnement de l'Association vau doise des
chercheurs en physique, Саас-Фе, 2–8 марта 1986
г. (1986 г.)

Распределение Фишера

Случайная величина X следует закону Фишера
распределения, если его функция плотности принимает
форма:

$$\phi(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{x^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}x\right)^{\frac{m+n}{2}}},$$

где – гамма-функция (см. гамма-распределение) ,
m, n – степени
свободы (m, n = 1, 2,...).



Распределение Фишера, m = 12, n = 8

Machine Translated by Google

Распределение Фишера является непрерывным распределение вероятностей.

ИСТОРИЯ

Распределение Фишера было открыто Фишер, Р.А. , 1925 г. Используемый символ F для обозначения распределения Фишера, был введен Снедекором в 1934 г. в честь Фишера Р.А.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если U и V — два независимых случайных переменные, каждая из которых следует за хи-квадрат распределение с соответственно m и n степеней свободы, то случайная величина:

$$Ф = \frac{Е/м}{В/н}$$

следует распределению Фишера с m и n степени свободы.

Ожидаемое значение случайной переменной F следующее за распределением Фишера:

$$E[F] = п - \frac{н}{2} \quad \text{для } n > 2 \quad ,$$

а дисперсия равна:

$$Var(F) = 2 м (n \frac{2n2 (м + п - 2)}{2)} \quad п > 4 \quad .$$

$$(п - 4)$$

Распределение Фишера с 1 и v степенями свободы совпадает с квадратом Распределение Стьюдента с v степенями свободы.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Важность распределения Фишера в статистическая теория связана с ее применением к распределению отношения независимых оценщики дисперсии . В настоящее время это дистрибутив чаще всего используется в стандартные тесты, связанные с анализом дисперсионный и регрессионный анализ.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Распределение хи-квадрат
- Непрерывное распределение вероятностей
- Таблица Фишера
- Тест Фишера
- Студенческое распределение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Статистические методы для Исследовательские работники. Оливер и Бойд, Эдин Бург (1925)

Снедекор, Г.В.: Расчет и интерпретация дисперсионного и ковариационного анализа. Коллегиальный, Эймс, Айова (1934)

Индекс Фишера

Индекс Фишера представляет собой составной индекс , который позволяет нам изучать увеличение стоимость жизни (инфляция). Это среднее геометрическое двух индексов:

- Индекс Ласпейреса и
- Индекс Пааше.

$$\text{Индекс Фишера} = \sqrt{\text{(индекс Ласпейреса)} \times \text{(индекс Пааше)}} \quad .$$

ИСТОРИЯ

В 1922 году экономист и математик Фишер, Ирвинг создали модель, основанную на индекс Фишера, чтобы обойти некоторые вопросы, связанные с использованием индекса num Беры Ласпейреса и Пааше:

Поскольку индекс Ласпейреса всегда использует количество товаров, проданных изначально в качестве эталона, он может переоценить любое увеличение стоимости жизни (поскольку потребление более дорогие товары со временем могут упасть).

И наоборот, поскольку индекс Пааше всегда использует количество товаров, проданных в текущем

Machine Translated by Google

период как эталон, он может недооценивать увеличение стоимости жизни.

Поскольку индекс Фишера является геометрическим среднего значения этих двух индексов, его можно считать идеальным компромиссом между их.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Используя определения индекса Ласпейреса и индекс Пааше, индекс Фишера Fn/0 рассчитывается следующим образом:

$$Фн/0 = 100 \cdot \sqrt{\frac{PnQ0 \cdot PnQn}{P0Q0 \cdot P0Qn}},$$

где Pn и Qn представляют собой цену товара и количество их продано в текущем период n, а P0 и Q0 — цена и количество, проданное одного и того же товара в течение отчетного периода 0. Получаются разные значения для разных товаров.

Обратите внимание, что индекс Фишера обратим. означающий, что:

$$Фн/0 = \frac{1}{F0/n}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Несмотря на то, что индекс Фишера является своего рода «идеального индекса компромисса», это редко используется на практике.

ПРИМЕРЫ

Рассмотрим следующую воображаемую таблицу, в которой указаны цены и соответствующие количества. продано из трех товаров народного потребления по ссылке год 0 и в текущем году n.

	Количество (тысячи)		Цена (евро)	
Товары	1970	1988	1970	1988
	(Q0)	(Qn)	(P0)	(Пн)
Молоко	50,5	85,5	0,20	1,20
Хлеб	42,8	50,5	0,15	1.10
Масло	15,5	40,5	0,50	2.00

Из следующей таблицы имеем:

$$PnQn = 239{,}15,$$

$$P0Qn = 44{,}925,$$

$$PnQ0 = 138{,}68 \text{ и}$$

$$P0Q0 = 24{,}27.$$

Товары	PnQn	P0Qn	PnQ0	P0Q0
Молоко	102,60	17,100	60,60	10,10
Хлеб	55,55	7,575	47,08	6,42
Масло	81,00	20,250	31,00	7,75
Общий	239,15	44,925	138,68	24,27

Затем мы можем найти индекс Пааше:

$$В/0 = \frac{Пн \cdot Qн}{P0 \cdot Qn} \cdot 100$$

$$= \frac{239{,}15}{44{,}925} \cdot 100 = 532{,}3,$$

и индекс Ласпейреса:

$$В/0 = \frac{Pn \cdot Q0}{P0 \cdot Q0} \cdot 100$$

$$= \frac{138{,}68}{24.27} \cdot 100 = 571{,}4.$$

Индекс Фишера – это квадратный корень из произведение порядковых чисел Пааше и Ласпейреса (или среднее геометрическое два индекса):

$$\text{Индекс Фишера} = \sqrt{532{,}3 \times 571{,}4} = 551{,}5.$$

По индексу Фишера цена рассматриваемые товары подорожали на 451,5% (551,5 100) между отчетным годом и текущий год.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Составной порядковый номер
- Индекс
- Индекс Ласпейреса
- Индекс Пааше
- Простой порядковый номер

Ф

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, И.: Создание индексов.
Хоутон Миффлин, Бостон (1922 г.)

Фишер, Ирвинг

Фишер, Ирвинг родился в Нью-Йорке 27 февраля 1867 года. Он получил докторскую степень в Йельском университете в 1892 году. Его математический подход к теории ценностей и цен, впервые описанный в его диссертации, привел к тому, что он стал известен как один из первых американских математические экономисты. Он основал Эконометрическое общество в 1930 г. вместе с Фришем, Рагнаром и Роосом, Чарльзом Ф., и стал его первым президентом. Фишер И. оставался в Йельском университете на протяжении всей своей карьеры. Он начал с преподавания математики, затем экономики. После пребывания в Европе он был назначен профессором социальных наук в 1898 году и умер в Нью-Йорке в 1947 году.

Фишер был плодовитым автором: его список публикаций насчитывает более 2000 наименований. Он интересовался как экономической теорией, так и научными исследованиями. В 1920 году он предложил эконометрические и статистические методы для расчет индексов. Его «идеальный индекс», который мы теперь называем индексом Фишера, представляет собой среднее геометрическое индекса Ласпейреса и индекса Пааше.

Некоторые основные работы и статьи Фишера, Ирвинга:

1892 г. Математические исследования в теории стоимости и цен. Connecticut Academy of Arts and Science, Нью-Хейвен, Коннектикут.

1906 г. Природа капитала и дохода. Макмиллан, Нью-Йорк.

1907 Процентная ставка: ее природа, Определение и отношение к экономические явления. Макмиллан, Нью-Йорк.

1910 Введение в экономическую науку. Макмиллан, Нью-Йорк.

1911 Покупательная способность денег. Макмиллан, Нью-Йорк.

1912 Элементарные принципы экономики. Макмиллан, Нью-Йорк.

1921 Лучшая форма порядкового номера. Являясь. Стат. доц. кв., 17, 533–537.

1922 Создание индексов. Хоутон-Миффлин, Бостон, Массачусетс.

1926 Статистическая связь между unem применение и изменение цен. Междунар. Трудовая редакция, 13, 785–792.

1927 г. Статистический метод измерения «предельной полезности» и проверки справедливости прогрессивного подоходного налога. В: Hollander, JH (ed) Economic Essays Contributed in Honor of John Bates Clark. Макмиллан, Нью-Йорк.

1930 Теория процента, определяемая нетерпением тратить доход и возможностью инвестировать его. Макмиллан, Нью-Йорк.

1932 Бумы и депрессии. Аделькфи, Нью-Йорк.

1933 Теория долговой дефляции Великих депрессий. Эконометрика, 1, 337–357.

1935 100% Деньги. Аделькфи, Нью-Йорк.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Индекс Фишера

Индекс

Machine Translated by Google

Фишер, Рональд Эйлмер

Родившийся в 1890 году в Ист-Финчли, недалеко от Лондона, Фишер, Рональд Эйлмер изучал математику в Харроу и в колледже Гонвилля и Кайуса в Кембридже. Он начал свою статистическую карьеру в 1919 году, когда он начал работать в Институте tute сельскохозяйственных исследований в Ротамстеде («Экспериментальная станция Ротамстеда»). Время, которое он провел там, продолжавшееся до 1933 года, было очень продуктивным. В 1933 году он стал профессором евгеники в Университетском колледже Лондона, а затем с 1943 по 1957 год он занимал кафедру генетики Бальфура в Университете Лондона. Кембридж. Даже после выхода на пенсию Фишер провел ряд исследований для Отдела математической статистики Общего собрания. богатство Научные и промышленные исследования организации в Аделаиде, Австралия, где он умер в 1962 году.

Фишер признан одним из основоположников современной статистики. Хорошо известно в области генетики он также внес важный вклад в общую теорию оценки максимального правдоподобия. Его работа над

Следует также упомянуть план эксперимента, необходимый для изучения сельскохозяйственных экспериментальных данных, а также его разработку методики, используемой для анализа различных анс.

Фишер разработал принципы рандомизации, рандомизированных блоков, планов латинских квадратов и факторных схем.

Некоторые из основных работ и статей Фишера Р.А.:

1912 г. Об абсолютном критерии подгонки частотных кривых. Mess.Math., 41, 155–160.

1918 Соотношение между родственниками по предположение Менделя

это. Транс. Рой. соц. Эдинбург, 52, 399–433.

1920 г. Математическое исследование методы определения точности наблюдения по средней ошибке и по среднеквадратической ошибке. Пн. Нет. Р. Астрон. Соц., 80, 758–770.

1921 г. О «вероятной ошибке» коэффициента корреляции, полученного из небольшой образец. Метрон, 1(4), 1–32.

1922 г. К математическому обоснованию теоретическая статистика. Фил. Транс. А, 222, 309–368.

1925 Теория статистического оценивания. проц. Камб. Фил ос. Соц., 22, 700–725.

1925 г. Статистические методы исследования Рабочие. Оливер и Бойд, Лондон.

1934 г. Вероятность, правдоподобие и количество информации в логике неопределенного вывода. Proc.Roy.Soc.A, 146, 1–8.

1935 План экспериментов. Оливер и Бойд, Эдинбург.

1971–1974 Беннетт, Дж Х. (редактор) Сборник статей Р. А. Фишера. ун-т Аделейд, Аделаида, Австралия

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение Фишера
- Таблица Фишера
- Тест Фишера

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер-Бокс, Дж: Р. А. Фишер, жизнь ученого. Уайли, НьюЙорк (1978)

Machine Translated by Google

Фишер Стол

В таблице Фишера приведены значения функции распределения случайной величины способным, который следует распределению Фишера.

ИСТОРИЯ

См. Распределение Фишера.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть случайная величина X подчиняется закону Фишера. распределения с m и n степенями свободы. Функция плотности случайной величины X определяется как:

$$\phi(т) = \frac{\frac{м+п}{2}}{\frac{м}{2}\quad\frac{н}{2}} \quad \frac{м}{н}^{\frac{м}{2}} \cdot \frac{т^{\frac{м}{2}-2}}{1+\frac{м}{н}т^{\frac{м+п}{2}}}, \quad т \quad 0 ,$$

где представляет собой гамма-функцию (см. гамма-распределение).

При заданных степенях свободы m и n таблица Фишера позволяет определить значение x, при котором вероятность P(X x) равна определенному значению.

Наиболее часто используемые таблицы Фишера дают значения x для

$$P(X \quad x) = 95\% \text{ и } (X \quad x) = 97{,}5\% .$$

Обычно мы используем $(x =) Fm,n,\alpha$ для обозначения значения случайной величины X, для которой

$$PX \quad Fm,n,\alpha = 1 \quad \alpha .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Таблица Фишера используется при проверке гипотез, когда в ней используется распределенная статистика. по распределению Фишера и особенно при дисперсионном анализе, а также при простом и множественном линейном регрессионном анализе.

Меррингтон и Томпсон (1943) предоставили таблицы Фишера для значений длиной до пяти знаков после запятой, а Фишер и Йейтс (1938) — таблицы длиной до двух знаков после запятой.

ПРИМЕРЫ

Если X следует распределению Фишера с m = 10 и n = 15 степеней свободы значение x, соответствующее вероятности 0,95, равно 2,54:

$$P(X \quad 2{,}54) = 0{,}95 .$$

Точно так же значение x, соответствующее вероятности 0,975, равно 3,06:

$$P(X \quad 3{,}06) = 0{,}975 .$$

Пример использования таблицы Фишера см. в разделе Тест Фишера.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Распределение Фишера
- Множественная линейная регрессия
- Простая линейная регрессия
- Статистическая таблица

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер Р.А., Йейтс Ф.: Статистические таблицы для биологических, сельскохозяйственных и медицинских исследований. Оливер и Бойд, Эдинбург и Лондон (1963)

Меррингтон, М., Томпсон, К.М.: Таблицы процентных пунктов обратного бета- (F) -распределения. Биометрика 33, 73–88 (1943).

Machine Translated by Google

Тест Фишера

Тест Фишера — это проверка гипотезы о наблюдаемое значение статистики вида :

$$F = \frac{U}{M} \cdot \frac{N}{B},$$

где U и V — независимые случайные величины, каждая из которых подчиняется распределению хи-квадрат с m и n степенями соответственно. свободы.

ИСТОРИЯ

См. Распределение Фишера.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим линейную модель с одним фактором применяется на t уровнях:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$
$$i = 1, 2, \ldots, t ; j = 1, 2, \ldots, n_i ,$$

где

Y_{ij} представляет j-е наблюдение, получающее лечение я ,

μ – общее среднее, общее для всех обработок,

τ_i — реальный эффект лечения i на наблюдение и

ε_{ij} – экспериментальная ошибка наблюдения Y_{ij}.

Случайные величины ε_{ij} являются независимыми и распределены нормально со средним значением 0 и дисперсией $\sigma^2 : N(0, \sigma^2)$.

В этом случае нулевая гипотеза , утверждающая что есть существенная разница между Лечение заключается следующем:

$$H0 : \tau 1 = \tau 2 = \quad \cdots \quad = \tau t .$$

Альтернативная гипотеза состоит в следующем:

H1 : значения $\tau i (i = 1, 2, \ldots, t)$ не все идентичны.

Чтобы сравнить изменчивость внутри обработки, также называемую ошибкой или остатками, с изменчивости между обработками, мы строим отношение , где числитель является оценкой разница между лечением _ а в знаменателе оценка отклонение в лечении или ошибка. Рыба ertеststatistic, также называемый коэффициентом F, тогда предоставлено:

$$F = \frac{\sum_{i=1}^{t} n_i \cdot (\bar{Y}_{i.} - \bar{Y}_{..})^2 / (t-1)}{\sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 / (N-t)},$$

где

n_i – количество наблюдений в тракте мент $(i = 1, \ldots, t)$

$N = \sum_{i=1}^{t} n_i$ – общее количество наблюдений ции,

$\bar{Y}_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$ – среднее значение i-й обработки, для я $i = 1, \ldots, t$ и

$\bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$ является глобальным средним значением.

После выбора уровня значимости α для теста сравниваем расчетное значение F с соответствующим критическим значением в таблице Фишера для t 1 и N t степеней свободы, обозначаемых $F_{t-1, N-t, \alpha}$.

Если $F \geq F_{t-1, N-t, \alpha}$, мы отвергаем нулевую гипотезу , что означает, что хотя бы одна из трактовок отличается от других.

Если $F < F_{t-1, N-t, \alpha}$, мы не отвергаем нулевую гипотеза; иными словами, лечение не представляют существенной разницы, что также означает, что образцы взяты из одного и того же Население .

Ф

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Критерий Фишера чаще всего используется при дисперсионном анализе, вковариационном анализа, так и регрессионном анализе.

ПРИМЕРЫ

См. однофакторный дисперсионный анализ и двусторонний дисперсионный анализ.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Распределение Фишера
- Таблица Фишера
- Проверка гипотезы
- Регрессивный анализ

Прогнозирование

Прогнозирование – это использование статистических данные, экономическая теория и неэкономические условия для формирования разумного мнение о будущих событиях.
Для прогнозирования используется множество методов. будущего. Все они основаны на ограниченном числе базовых гипотез.

Одна из таких гипотез утверждает, что значительные тенденции, наблюдаемые в его будет сохраняться и в будущем.
Другой утверждает, что измеримое колебание ции в переменной будут воспроизводиться в регулярными интервалами, так что эти изменения будут стать предсказуемым.

Анализ временных рядов, выполняемый в процессе прогнозирования, основан на двух гипотезах:

1. Долгосрочный прогноз:
   Когда мы делаем долгосрочные прогнозы(подробнее чем через пять лет), анализ и экстраполяция вековых тенденций становятся важными.
   Долгосрочные прогнозы на основе светских тенденция обычно не принимает циклический флюк во внимание. Более того, сезонные колебания не влияют на годовые данные.
   и не рассматриваются в долгосрочной перспективе бросать.
2. Среднесрочный прогноз:
   Для того, чтобы принять вероятное влияние учет циклических колебаний в механизм, используемый для получения среднесрочной прогноз (от одного до пяти лет вперед), умножаем значение проекции вековой тенденции на циклическую юфлуктуацию, что дает прогнозируемое значение. Это, из конечно, предполагает, что те же самые структуры данных, которые производят циклические изменения будет держаться в будущем, и так наблюдается вариации будут продолжать повторяться регулярно. В действительности последующие циклы имеют тенденциюсильно различаться по промежуткам времени. периодичность, их амплитуда и модели, которым они следуют.
3. Краткосрочный прогноз:
   Необходимо учитывать сезонный индекс. учитывать в краткосрочном прогнозе (несколько месяцев вперед), наряду с прогнозируемым значения для вековой тенденции и циклического колебания.
   Как правило, мы не пытаемся прогнозировать нерегулярные колебания.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть Y будет временным рядом. Мы можем описать это используя следующие компоненты:

- Светская тенденция T;
- Сезонные колебания S;
- циклические колебания C;
- Неправильные вариации I.

Мы различаем:

- Мультипликативная модель $Y = T \cdot S \cdot C \cdot I$;
- Аддитивная модель $Y = T + S + C + I$.

Для значения t, превышающего время принятых для выполнения наблюдения, мы указываем

Machine Translated by Google

что:

Yt — прогноз значения Y в момент времени t; Tt — проекция вековой тенденции в момент времени t; Ct — циклическое колебание в момент времени t; St – прогноз сезонных колебаний в момент времени t.

В этом случае:

1. Для долгосрочного прогноза значение временного ряда можно оценить с помощью

$$Yt = Tt.$$

2. Для среднесрочного прогноза значение временного ряда можно оценить с помощью

$$Yt = Tt \cdot Ct \text{ (мультипликативная модель)}$$

или $Yt = Tt + Ct$ (аддитивная модель).

3. Для краткосрочного прогноза значение временного ряда можно оценить с помощью

$$Yt = Tt \cdot Ct \cdot St$$

(для мультипликативной модели)

или $Yt = Tt + Ct + St$

(для аддитивной модели).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Планирование и принятие решений — это два вида деятельности, которые в некоторой степени предполагают предсказание будущего. Другими словами, администраторы, выполняющие эти задачи, должны делать прогнозы.

Стоит отметить, что получение прогнозов с помощью анализа временных рядов имеет смысл, даже несмотря на то, что прогноз обычно оказывается неточным. В действительности, даже когда он неточен, прогноз, полученный таким образом, вероятно, будет более точным, чем другой прогноз, основанный только на интуиции. Таким образом, несмотря на свои ограничения и проблемы, анализ временных рядов является полезным инструментом в процессе прогнозирования.

ПРИМЕРЫ

Понятие «прогноз» зависит от предмета исследования. Например: • В метеорологии:

- Краткосрочный прогноз выглядит часами вперед,
- Долгосрочный прогноз смотрит на дни вперед. • В экономике:
- Краткосрочный прогноз выглядит месяцами предстоящий;
- Среднесрочный прогноз выглядит от одного до пяти годы вперед; –

Долгосрочный прогноз рассчитан более чем на пять лет вперед.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Циклическое колебание
Неправильная вариация
Сезонный индекс
Сезонная вариация
Светская тенденция
Временная последовательность

Дробный факторный план

Дробный факторный план эксперимента — это факторный эксперимент, в котором только часть комбинаций уровней факторов возможное реализуется.

Этот тип плана используется, когда эксперимент содержит ряд факторов, которые считаются более важными, чем другие, и/или существует большое количество уровней факторов. Конструкция позволяет использовать для уменьшения количества необходимых экспериментальных единиц.

ПРИМЕРЫ

Для факторного эксперимента, содержащего два фактора, каждый из которых имеет три уровня, имеется девять возможных комбинаций. Если четыре из этих наблюдений подавить, часть

Machine Translated by Google

остаются наблюдения, которые могут быть схемой следующим образом:



где Yij представляет наблюдение , сделанное для i-го уровня фактора A и j-го уровня фактора B.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- План экспериментов
- Эксперимент

ССЫЛКИ Yates, F.: Дизайн и анализ факторных экспериментов, Techn. комм. 35, Имперское бюро почвоведения, Харпенден (1937 г.)

Ч астота

Ч астота или абсолютная ч астота соответствует количеству появлений конкретного наблюдения или результата в эксперименте.

Абсолютную ч астоту можно отличить от относительной ч астоты. Относительная ч астота наблюдения определяется как отношение ч исла появлений к общему ч ислу наблюдений.

Для некоторых анализов желательно знать количество наблюдений, для которых значение , скажем, «меньше или равно» или «больше, чем» заданное значение. Количество значений, меньших или равных этому заданному значению, называется кумулятивной ч астотой.

Также может потребоваться процент наблюдений, в которых значение «меньше или равно» или « больше » заданного значения . Доля значений, меньших или равных этому данное значение называется кумулятивной относительной ч астотой.

ПРИМЕРЫ

Рассмотрим, например, следующие 16 наблюдений, которые представляют рост (в см) группы людей:

174 169 172 174
171 179 174 176
177 161 174 172
177 168 171 172

Здесь ч астота наблюдения 171 см равна 2; тогда как ч астота наблюдения 174 см равна 4.

В этом примере относительная ч астота 174 = $0{,}25 = \frac{4}{16}$ (или 25% всех наблюдений). Абсолютная ч астота равна 4, а общее количество наблюдений равно 16.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Кривая ч астоты
- Распределение ч астоты
- Таблица ч астот

Ч астотная кривая

Если полигон ч астот сгладить, получается кривая, называемая кривой ч астот. Это сглаживание может быть выполнено, если количество наблюдений в ч астотном распределении становится бесконечно большим, а ширина классов становится бесконечно малой.

Ч астотная кривая соответствует предельной форме ч астотного полигона.

ИСТОРИЯ

К концу девятнадцатого века ry, общей тенденцией было рассматривать все

Machine Translated by Google

распределения должны быть нормальными распределениями.

Гистограммы, представляющие несколько мод, были скорректированы для формирования набора кривых нормальной частоты, а те, которые представляли асимметрию, были проанализированы путем преобразования данных таким образом, чтобы гистограмму, полученную в результате этого преобразования, можно было сравнить с нормальной кривой.

В 1895 г. Пирсон К. предложил создать набор различных теоретических частотных кривых, чтобы иметь возможность получить лучшие приближения к гистограммам.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Непрерывное распределение вероятностей
- Распределение частоты
- Полигон частот
- Нормальное распределение

ССЫЛКИ Пирсон, К.: Вклад в математическую теорию эволюции. II: Косая вариация в однородном материале. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, Кембридж, стр. 41–112 (1948). Впервые опубликовано в 1895 году в Philos. Транс. Рой. соц. Лонд. сер. А 186, 343–414

Распределение частоты

Распределение населения может быть выражено в относительном выражении (например, в процентах), в долях или в абсолютном выражении. Во всех этих случаях результат называется частотным распределением.

Мы также получаем либо дискретное частотное распределение, либо непрерывное частотное распределение, в зависимости от того, является ли изучаемая переменная дискретной или непрерывной. Количество появлений определенного значения или количество наблюдений за видом, определенного класса, называется частотой этого значения или класса.

Полученное таким образом распределение частот может быть представлено либо в виде таблицы частот, либо графически, например, в виде гистограммы, полигона частот или линейной диаграммы.

Любому частотному распределению соответствует относительное частотное распределение, которое является распределением каждого значения или каждого класса по отношению ко общему количеству наблюдений.

ИСТОРИЯ

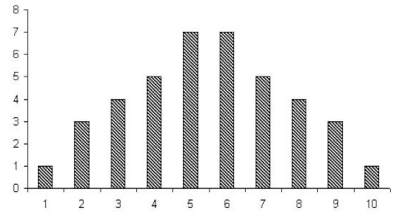
См. графическое представление.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

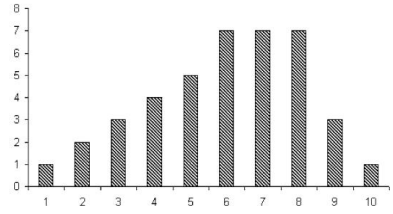
На практике частотные распределения могут значительно различаться. Например:

- Некоторые симметричны, другие области

Рик:



Симметричное распределение

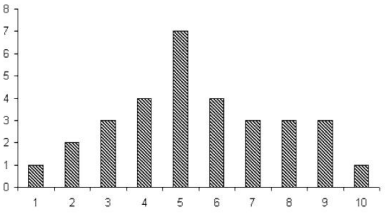


Асимметричное распределение

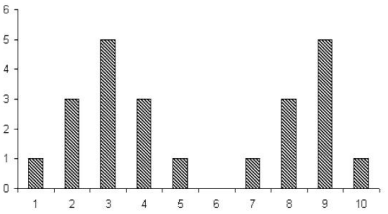
- Некоторые имеют только одну моду (унимодальные распределения); у других есть несколько (множественных модальных распределений):

Ф

Machine Translated by Google



Унимодальное распространение



Бимодальное распределение

Это означает, что частотные распределения могут можно разделить на четыре основных типа:

1. Унимодальные симметричные распределения;
2. Унимодальные асимметричные распределения;
3. J-образные распределения;
4. U-образные распределения.

При статистических исследованиях часто возникает необходимость сравнить два распределения. Если оба распределения однотипны, их можно сравнить, изучив их основные характеристики, такие как мера центральной тенденции, мера дисперсии или мера формы.

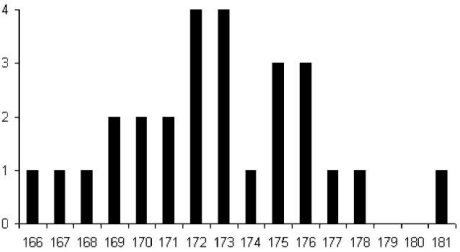
ПРИМЕРЫ

Следующие данные представляют высоты, измерено в сантиметрах, наблюдаемое для совокупности 27 учащихся младших классов средней школы:

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

Упорядочив и перегруппировав эти наблюдения по значению, мы получим следующее распределение частот:

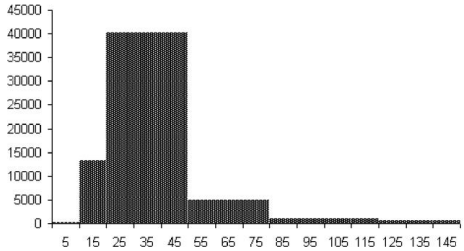
Высота	Частота	Высота	Частота
166	1	174	1
167	1	175	3
168	1	176	3
169	2	177	1
170	2	178	1
171	2	179	0
172	4	180	0
173	4	181	1



Вот еще один пример, который касается частотное распределение непрерывного переменная. После получения и классификации личных доходов населения г. Невшатель (штат в Швейцарии), для период 1975–1976 гг., получаем следующее таблица частот:

Чистая выручка (в тыс. швейцарских франков)	Фре качества	Родственник частота
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



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Ч астота
- Кривая ч астоты
- Пол иг он ч астот
- Табл иц а ч астот
- Г истог рамма
- Инте рвал
- Л инейный г рафик

Ч астотный пол иг он

Пол иг он ч астот пре дставл яе т собой г рафич е ское пре дставл ение ч астотного распре де л е ния, ч то соответствуе т г истог рамме ч астоты распре де л е ние . Э то тип ч астотного г рафика. Он принимае т форму се г ме нтированной л инии, сое диняющей се ре дина ве ршины каждого прямоу г ол ьника в г истог рамма.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Мы строим пол иг он ч астот, отме ч ая точ ку дл я каждого кл асса ч астоты распре де л е ние . Каждая точ ка соответствуе т се ре дина кл асса по оси абсцисс и ч астота кл асса по ординате .
З ате м мы сое диняе м каждую точ ку с сосе дними точ ками л иниями. Таким образом, л инии э ффе ктивно сое диняют сре днюю точ ку ве ршины каждого прямоу г ол ьник на г истог рамме ч астоты распре де л е ние .

Обы чно мы з амыкае м пол иг он ч астот в точ ке два конц а распре де л е ния, выбрав соответствующие точ ки на г ориз онтал ьной оси и сое диняя их с точ ками дл я пе рвого и после дние з анятия.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Ч астотный пол иг он обы чно строится из данны х, г ру ппированны х в инте рвал ы. В э том сл у ч ае пре дпоч тите л ьне е испол ьзовать проме жу тки одинаковой ширины. Е сл и мы подходим г л адку ю кривую к пол иг ону ч астот, мы пол у ч ить ч астотну ю кривую. Э тот проце сс сг л аживания вкл ю ч ае т пе ре г ру ппировку данны х таким образом, ч то ширина кл асса как можно ме ньше , и з ате м пе ре строить распре де л е ние . Ясно, ч то мы може м вы пол нять тол ько э тот тип сг л аживание , е сл и кол ич е ство е диниц в совоку пности достаточ но ве л ико, ч тобы дать з нач ите л ьну ю кол ич е ство наб л юде ний в каждом кл ассе после пе ре г ру ппировка.

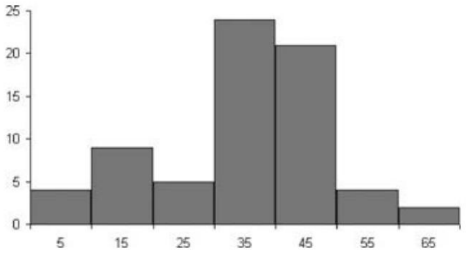
ПРИМЕРЫ

В приве де нной ниже табл ице ч астот у каз аны е г одны е сре дне е кол ич е ство осадков дл я 69 г ородов США:

Сре дне г одовое кол ич е ство осадков (в дюймах)	Ч астота
0–10	4
10–20	9
20–30	5
30–40	24
40–50	21
50–60	4
60–70	2
Общий	69

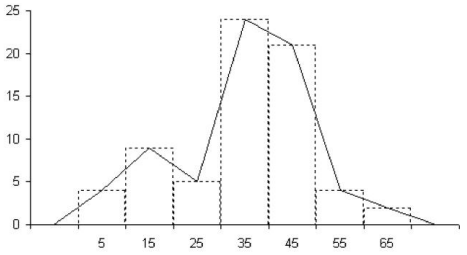
Пол иг он ч астот строится по г истог рамма ч астотного распре де л е ния:

Machine Translated by Google



Для построения полигона частот построим график балл за каждый класс. Положение этого точка вдоль абсцисс задается средней точкой класса, а ее положение вдоль ордината задается частотой сорт. Другими словами, каждая точка находится в середина вершины прямоугольника для класс на гистограмме. Затем каждая точка соединяется со своими ближайшими соседями линиями. В чтобы замкнуть многоугольник, ставится точка на абсциссе на ширину одного класса перед середина первого класса и еще одна точка помещается по оси абсцисс шириной в один класс после середины последнего занятия. Эти затем точки соединяются линиями с точками для первого и последнего классов соответственно.

полигон частот, который мы получаем в результате этой процедуры, выглядит следующим образом:



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение частоты
- Таблица частот
- Графическое представление
- Гистограмма
- ожив

Таблица частот

Таблица частот представляет собой инструмент, используемый для представления частотного распределения. Это обеспечивает способность представлять статистические наблюдения.

Частота значения переменной _ _ _ _ сколько раз это значение появляется в население. Частотное распределение может тогда определяется как список частот получ енные для различных значений, принимаемых переменная.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим набор из n единиц, описываемых переменной X , которая может принимать k значений x1, x2,...,xk. Letni bekoличество единиц, имеющих значение xi; ni – это частота значения xi. Относительная частота xi равна fi = $\frac{ни}{н}$. Поскольку значения различны и исчерпывающи, сумма всех частот ni равна общее количество n единиц в наборе или сумма относительных частот fi равно единице; в другие слова:

$$\sum_{я=1}^{к} п = п_и \qquad \sum_{я=1}^{к} фи = 1 \cdot$$

Соответствующая таблица частот для этого сценарий следующий:

Ценности переменная Икс	Относительные частоты	частоты
x1	n1	f1
x2	n2	f2
...	...	...
xk	нк	фк
Общий	н	1

Иногда важно понятие кумулятивной частоты (см. второй пример

Machine Translated by Google

например). это сумма всех частых

я

до частоты fi : fj включительно .

j=1

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Построение таблицы частот является самым простым и наиболее часто используемым подходом. к представлению данных. Однако есть некоторые правила, которые следует соблюдать при создание такой таблицы:

- Таблица должна иметь исчерпывающую краткое название, в котором упоминаются единицы трудоустроенный;
- Названия строк и столбцов должны быть точными и кратким;
- Должны быть указаны итоги столбца;
- Источник данных, приведенных в таблице должны быть указаны.

В общем, таблица должна быть достаточно понятной, чтобы ее можно было понять без необходимости прочитайте сопроводительный текст.
Количество классов, выбранных для представления данные зависят от размера изучаемого множества. Использование большого количества классов для небольшого приведет к нерегулярным частотам из-за малое количество единиц в классе. На с другой стороны, используя небольшое количество классов для большой набор приводит к потере информации о структуре данных.

Однако не существует общего правила для определения числа классов, которые должны быть используется для построения частотной таблицы. Для относительно небольших наборов (n 200) от 7 до рекомендуется 15 классов, хотя это правило не является абсолютным.

Для упрощения часто используют классы одинаковой ширины, а интервалы классов должны не перекрываются. В некоторых случаях, например, если у нас есть переменная, которая может принимать большой диапазон значений, но некоторый интервал значений ожидается, что будет гораздо чаще, чем остальные, может быть полезно использовать открытый класс, например «1000 и более».

ПРИМЕРЫ

Частотное распределение жителей Австралии (в тысячах) в соответствии с их матримониальным статусом на 30 июня 1981 г.

приведены в следующей таблице.

Супружеский положение дел	Частота (в тысячах)	Родственник частота
Одинокий	6587	0,452
Женатый	6837	0,469
В разводе	403	0,028
Овдовела	749	0,051
Общий	14576	1.000

Источник: ABS (1984) Карманный ежегодник Австралии. Австралийское бюро статистики, Канберра, п. 11.

Данные представлены в виде таблицы частот. Переменная «Семейное положение» может принимать четыре различных значения: холост, женат, разведен или вдовец. частота , выражены в тысячах, и для каждого значения дана относительная частота.
В следующем примере мы рассматриваем случай, когда наблюдения сгруппированы в классы. В этой таблице представлена классификация 3014 человек в зависимости от их доходов. в 1955 году .

Доход	Частота		
	ни	Относительный	Относительный кумулятивный
Меньше, чем 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

Доход	Частота		
	ни	Относительный	Относительный кумулятивный
5000–7499	703	0,2332	0,8112
7500–9999	277	0,0919	0,9031
10000 или более	292	0,0969	1,0000
Общий	3014	1.0000	

Из этой таблицы видно, что 57,8% из этих 3014 человек имеют доход меньше 5000 долларов.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Частота
- Распределение частоты

Machine Translated by Google

Гальтон, Фрэнсис

Гальтон, Фрэнсис родился в 1822 году недалеко от Бирмингема, Англия, в семье интеллектуалов: Чарльз Дарвин приходился ему двоюродным братом, а его дедушка был членом Королевского общества.

Его интерес к науке впервые проявился в области географии и метеорологии. Избран членом Королевского общества в 1860 году он начал интересоваться генетики и статистических методов в 1864 г. Гальтон был близким другом Пирсона, К. Действительно, Пирсон пришел к нему на финансовую помощь когда Гальтон основал журнал «Биометрика». Гальтон «Евгенический архив» объединилась с биометрической лабораторией Пирсона в Университетском колледже Лондона и стала известная как «лаборатория Гальтона». Он умер в 1911 году, оставив после себя более 300 публикаций, в том числе 17 книг. по статистическим методам, связанным с регрессией анализ и концепция корреляции, оба из которых приписываются ему.

Некоторые из основных работ и статей Гальтона Ф:

1869 г. Наследственный гений: исследование его законы и последствия. Макмиллан, Лондон.

1877 г. Типичные законы наследственности. Природа, 15, 492–495, 512–514, 532–533.

1889 г. Естественное наследство. Макмиллан, Лондон.

1907 Вероятность, основание евгеники. Генри Фруд, Лондон.

1908 Воспоминания о моей жизни. Метуэн, Лондон.

1914–1930 Пирсон, К. (редактор) Жизнь, письма и труды Фрэнсиса Гальтона. Издательство Кембриджского университета, Кэмбридж

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коэффициент корреляции

Регрессивный анализ

Гамма-распределение

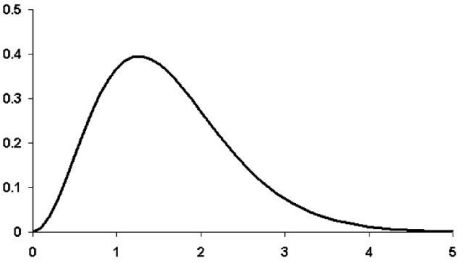
Случайная величина X следует за гаммой распределение с параметром α, если его плотность функция определяется

$$e(x) = \frac{\beta\alpha}{(\alpha)} x\alpha \quad 1 \cdot e \quad \beta x \quad ,$$

если $x > 0$, $\alpha > 0$, $\beta > 0$,

где (α) — гамма-функция.

Machine Translated by Google



Гамма-распределение, α = 2, β = 1

Стандартная форма гамма-распределения получается, если положить β = 1, что дает

$$e(x) = \frac{1}{(\alpha)} x^{\alpha} \quad 1 \cdot e \quad x,$$

если $x > 0$, $a > 0$.

Гамма-функция () появляется часто в статистической теории. Это определяется:

$$(\alpha) = \int_0^{\tau\alpha} \quad 1 e \quad t\,dt.$$

В статистике нас интересуют только значения $a > 0$.

Интегрируя по частям, получаем:

$$(\alpha + 1) = \alpha(\alpha).$$

С

$$(1) = 1 \quad ,$$

у нас есть

$(\alpha+1) = \alpha!$ для каждого положительного целого числа α.

ИСТОРИЯ

Согласно Ланкастеру (1966), это непрерывное распределение вероятностей было предложено Лапласом П.С. (1836).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение гамма-распределения определяется как:

$$E[X] = \beta \frac{1}{} \cdot \frac{(\alpha + 1)}{(\alpha)} \text{ "=" } \frac{\alpha}{\beta}.$$

С

$$E\ X2 = \beta2 \frac{1}{} \cdot \frac{(\alpha + 2)}{(\alpha)} \text{ "=" } \frac{\alpha(\alpha + 1)}{\beta2},$$

дисперсия равна:

$$\mathrm{Var}(X) = E\ X2 \quad (E[X])^{2}$$

$$\text{"=" } \frac{\alpha(\alpha + 1) \quad \alpha2}{\beta2}$$

$$\text{"=" } \frac{\alpha}{\beta2}.$$

Распределение хи-квадрат является особым случаем гамма-распределения, где $\alpha = и \beta = \frac{1}{2}$, и n - количество градусов свободы распределения хи-квадрат.

Гамма-распределение с параметром $\alpha = 1$ дает экспоненциальное распределение:

$$f(x) = \beta \cdot e \quad \beta x \quad .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение хи-квадрат
- Непрерывное распределение вероятностей
- Экспоненциальное распределение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Ланкастер, HO: Предшественники хи-квадрата сына груши. Ау ст. Дж Стат. 8, 117–126 (1966)

Лаплас, PS de: Аналитическая теория вероятности, доп. к 3-му изд. Курсье, Париж(1836 г.)

Гаусс, Карл Фридрих

Гаусса, Карл Фридрих родился в 1777 г. Брауншвейг в Германии. Он быстро стал известный астроном и математик и до сих пор считается равным Архимеда и Ньютона как одного из величайших математиков всех времен.

Machine Translated by Google

Он получил докторскую степень в 1799 году, а затем работал в Хелмстедском университете. В 1807 году он переехал в Геттинген, где стал директором лаборатории. Остаток жизни он провел в Геттингене, где и умер в 1855 году.

Его вклад в науку, особенно в физику, имеет большое значение. В статистике его работы касались теории оценивания, и метод наименьших квадратов, и применение нормального распределения к задачам, связанным с ошибками измерения, возникли у него.

Некоторые из основных работ и статей Gauss, CF:

1803–1809 гг. Disquisitiones de elementis ellipticis pallidis. Верке, 6, 1–24.

1809 г. Theoria motus corporum coelestium. Werke, 7. (английский перевод 1963 г. Дэвиса, CH, опубликовано Dover, Нью-Йорк).

1816 г. Bestimmung der Genauigkeit der Beobachtungen. Верке, 4, 109–117. 1821, 1823 и 1826 гг.

obnoxiae, части 1, 2 и доп. Верке, 4, 1–108.

1823 Anwendungen der Wahrschein lichkeitsrechnung auf eine Aufgabe der praktischen Geometrie. Верке, 9, 231–237.

1855 Méthode des Moandres Carrés: Воспоминания о Combinaison des Наблюдения. французский перевод работа Gauss, CF by Bertrand, J. (санкционированная самим Gauss, CF). Мале-Башелье, Париж

1957 Работа Гаусса (1803–1826). О теории наименьших квадратов. Английский перевод Trotter, HF Technical Отчет № 5, Исследовательская группа статистических методов, Принстон, НьюДжерси.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Де Муавр, Абрахам
- Теорема Гаусса – Маркова
- Нормальное распределение

Теорема Гаусса – Маркова.

Г

Теорема Гаусса-Маркова постулирует, что, когда распределение вероятности ошибки неизвестно в линейной модели, то среди всех линейных несмещенных оценок для параметров линейной модели, оценка, полученная методом наименьших квадратов, является оценкой, минимизирующей дисперсию. Предполагается, что математическое ожидание каждой ошибки равно нулю и все они имеют одинаковую (неизвестную) дисперсию.

ИСТОРИЯ

Гаусса, Карл Фридрих представил доказательство этой теоремы в первой части своей работы «Теория комбинирования есть наблюдение». um erroribus minimis obnoxiae» (1821 г.). Марков Андрей Андреевич переоткрыл эту теорему в 1900 году. Вариант теоремы Гаусса – Маркова, написанный в современной нотации, был предоставлен Graybill в 1976 году.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим линейную модель

$$Y = X \cdot \beta + \varepsilon \quad ,$$

где

Y - вектор наблюдений $n \times 1$,

X - матрица $n \times p$ независимых переменных, которые считаются фиксированными,

Machine Translated by Google

β — вектор p × 1 неизвестных параметров, ε — вектор n × 1
случайных ошибок.

Если распределение вероятностей ошибок
неизвестно, но выполняются следующие условия:

1. Математическое ожидание E[ε] = 0, 2. Дисперсия
Var(ε) = σ2 ·In, где In
— единичная матрица, 3.
Матрица X имеет полный ранг, тогда
оценка β,

$$\hat{\beta} = XX^{-1}XY,$$

полученный с помощью метода наименьших квадратов,
является линейной оценкой без смещения β, которая имеет
наименьшую дисперсию

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Оценщик

Наименьших квадратов

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Gauss, CF: Theoria Combinationis Observationum
Erroribus Minimis Obnoxiae, части 1, 2 и доп.
Верке 4, 1–108 (1821, 1823, 1826)

Грейбилл, Ф.А.: Теория и приложения линейной
модели. Даксбери, Северный округ, Массачусетс
(Уолдсворт и Брукс/Коул, Пасифик-Гроув,
Калифорния) (1976)

Рао, Ч.Р.: Линейный статистический вывод и его
приложения, 2-е изд. Уайли, НьюЙорк (1973)

Обобщенный обратный

Обобщенная обратная аналогична обратной
неособой квадратной матрице, но используется
для матрицы любой размерности и ранга.
Обобщенный обратный используется при решении
систем линейных уравнений.

ИСТОРИЯ

Похоже, что Фредгольм (1903) был первым, кто
рассмотрел понятие обобщенного обратного. Мур
определил единую обобщенную инверсию в своей
книге «Общий анализ» (1935 г.), опубликованной
после его смерти. Однако его работа не
использовалась до 1950-х гг., когда она испытала
всплеск интереса из-за применения обобщенной
обратной задачи и к задачам, связанным с
наименьшими квадратами.
В 1955 г. Пенроуз, повторно используя и расширяя
работу Бьерхаммара, опубликованную в 1951 г.,
показал, что обобщенная обратная матрица Мура
является единственной матрицей G,
удовлетворяющей
следующим четырем
уравнениям: 1. A = A ·
G · A; 2. Г = Г · А · Г; 3. (А
· G) = A · G; 4. (Г · А) = Г · А.

Эта уникальная обобщенная инверсия известна
как инверсия Мура-Пенроуза и обозначается A+.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Тогда ×mmatrixGявляется обобщенной обратной
матрицей A размера m × n, если

$$A = A \cdot Г \cdot A .$$

Матрица G единственна тогда и только тогда, когда
m = n и A несингулярна.

Наиболее распространенное обозначение,
используемое для обобщенной обратной матрицы A, — это A .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Обратная матрица

Наименьших квадратов

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бьерхаммар, А.: Прямоугольные взаимные матрицы
со специальной ссылкой на геодезию

Machine Translated by Google

тиковые расчеты. Бык. Геод. 20, 188–220 (1951)

Додж, Ю, Маджумдар, Д.: Алгоритм поиска обобщенных обратных наименьших квадратов для моделей классификации с произвольными шаблонами. Дж. Стат. вычисл. Симул. 9, 1–17 (1979)

Фредхольм, Дж: Sur une classe d'equations fonctionnelles. Акта Математика. 27, 365–390 (1903)

Мур, ЕН: Общий анализ, часть I. Мем. Являются. Филос. Soc., Филадельфия, Пенсильвания, стр. 147–209 (1935).

Пенроуз, Р.: Обобщенная обратная для матриц. проц. Камб. Филос. соц. 51, 406–413 (1955)

Рао, Ч.Р.: Заметка об обобщенной обратной матрице с приложениями к задачам математической статистики. Дж. Рой. Стат. соц. сер. Б 24, 152–158 (1962)

Rao, CR: Исчисление обобщенных обратных матриц. I. Общая теория. Санкхья А29, 317–342 (1967)

Rao, CR, Mitra, SK: Generalized Inverses ofMatrices andits Applications. Wiley, Нью-Йорк (1971).

Обобщенная линейная регрессия

Расширение модели линейной регрессии для настройки более гибких базовых предположений, которые требуются для линейной регрессии. Например, вместо предположения, что ошибки должны иметь одинаковую дисперсию, мы могли бы иметь следующие формы: • Гетероскедастичность: ошибки, наблюдаемые в модели (также называемые остатками), могут иметь разные дисперсии среди наблюдаемых. ций (или среди разных групп наблюдений);

• Автокорреляция: может быть некоторая корреляция связь между ошибками в различных ент наблюдения.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим общую модель множественной линейной регрессии:

$$Yi = \beta 0 + \sum_{j=1}^{p-1} \beta j Xij + \varepsilon i, \qquad я = 1,...,n,$$

где

Yi — зависимая переменная, Xji, j = 1,...,p 1 — независимые переменные, βj, j = 0,...,p 1 — оцениваемые параметры, εi — случайная не наблюдаемая ошибка срок.

В матричной форме запишем:

$$Y = X\beta + \varepsilon ,$$

где:

Y — вектор (n × 1) наблюдений, связанных с зависимой переменной (n наблюдений), β — вектор (p × 1) оцениваемых параметров.

ε — вектор (n × 1) ошибок,

и

$$X = \begin{matrix} 1 & X11 & \dots & X1(p-1) \\ \vdots & \vdots & & \vdots \\ 1 & Xn1 & \dots & Xn(p-1) \end{matrix}$$

Machine Translated by Google

– матрица (n × p), связанная с независимыми переменными.

В обычной регрессионной модели гипотезы, используемые для ошибок εi , обычно следующие: •

Каждое математическое ожидание равно нулю(другими словами, наблюдения не смещены), • Они независимыи неверны.

поздно,

• их дисперсия V(εi) постоянна и равна σ2 (мы предполагаем, что ошибки одинакового размера).

Если выбранная модель является подходящей, распределение остатковдолжно подтверждать эти гипотезы (см. дисперсионный анализ). Если мы обнаружим, что это не так, или если мы получим дополнительную информацию, указывающую на корреляцию ошибок, то мы можем использовать обобщенную линейную регрессиюдля получения более точной оценки параметров β0,..., βp 1 . Затем мы вводим дисперсионно-ковариационную матрицу V(ε) ошибок (размера n × n), которая определяется следующим уравнением для строки i и столбца j: • V(εi), если i = j (дисперсия в εi), • Cov εi, εj (ковариация между εi и εj), если i = j.

Часто эту матрицу необходимо оценить, но сначала рассмотрим простейший случай, когда V (ε) известно. Мы используем обычный термин для дисперсии σ2 : V(ε) = σ2V.

Оценка вектора β при известном V

Преобразовав нашу общую модель в обычную регрессионную модель, можно оценить параметры обобщенной модели, используя:

$$\hat{\beta} = X V \quad 1X \qquad {}^{1} X V\text{-}1Y .$$

Мы можем доказать, что: 1. Оценка не смещена; то есть:

$E(\hat{\beta}) = \beta$.

2. Проверяется следующее равенство:

$$V\hat{\beta} = \hat{}\sigma 2\, X V \quad 1X \qquad {}^{1},$$

$$\text{с } \sigma\hat{} \quad {}^{2} \text{ "=" } \frac{1}{\text{п - р}} \varepsilon V \quad 1\varepsilon .$$

3. $\hat{\beta}$ —наилучшая (имеющая наименьшую дисперсию) несмещенная оценка β; этот результат известен как обобщенная теорема Гаусса–Маркова.

Оценка вектора β, когда V не известно

Когда дисперсионно-ковариационная матрица не известно, используемая модель называется обобщенной регрессией. Первый шаг состоит в выражении V как функции параметра θ, что позволяет нам оценить V , используя $\hat{V} = V(\hat{\theta})$. Тогда оценки параметров получаются заменой V на $\hat{V}$ в приведенной выше формуле:

$$\hat{\beta} = (X \hat{V} \quad 1X) \quad 1X \hat{V} \quad 1Y \cdot$$

Преобразование V в трех типичных случаях а) Гетероскедастичность

Мы предполагаем, что наблюдения различны. анс $w_i \frac{\sigma 2}{}$, где w_i —вес, который может быть разным для каждого i = 1,...,n. Так:

$$V = \sigma 2 \begin{matrix} \frac{1}{w1} & 0 \ldots 0 \\ 0 & \frac{}{1\ \text{п2}} \quad \ldots 0 \\ \vdots & \vdots \quad \ldots \quad \vdots \\ 0\,0 \ldots & \frac{}{1\ \text{вн}} \end{matrix} .$$

Делая

$$Yw = WY$$

$$X\text{в} = BX$$

$$\varepsilon w = W\varepsilon \quad ,$$

Machine Translated by Google

где W —следующая квадратная матрица порядка n:

$$Вт = \begin{matrix} \overline{w1} & 0 & 0 & \dots & 0 \\ 0 & w2 & \dots & 0 & \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \dots & вн & \end{matrix},$$

такое, что W W = V 1, получаем эквивалентное альтернативная модель:

$$V(\varepsilon w) = V(W\varepsilon) = WV(\varepsilon)W = \sigma 2WVW$$
$$= \sigma 2In,$$

где In – единичная матрица размерности n. Поскольку дисперсии всех ошибок в этой новой модели одинаковы, мы можем применить метод наименьших квадратов и получить вектор оценок:

$$\hat{\beta}ш = (XwXw) \quad 1XwYw$$
$$= (XWWX) \quad 1XWWY$$
$$= (XV \quad 1X) \quad 1XV \quad 1Y,$$

поэтому вектор оценочных значений для Y = W 1Yw:

$$\hat{Y} = W \quad 1\hat{Y} = ж$$
$$W \quad 1WX(X\,V \quad 1X) \quad 1X\,V \quad 1Y = X\hat{\beta}w.$$

б) Гетероскедастичность по группам Предположим, что наблюдения разбиваются на три группы (размера n1, n2, n3) с дисперсиями σ2 σ2, и σ2 3 соответственно. Так:

$$B = \begin{matrix} \sigma 2 1 B 1 & On2\times n1 & On3\times n1 \\ On1\times n2 & \sigma 2 \quad 2 In2 Вкл & 3\times n2 \\ On1\times n3 & On2\times n3 & \sigma 2 \quad 3 In3 \end{matrix},$$

где Inj —единичная матрица порядка nj, а Oni×nj —нулевая матрица порядка ni ×nj.

Дисперсия σ2 1, σ2 2, σ2 3 можно оценить путем выполнения обычных регрессий для различных групп наблюдений и оценки σ2 с использованием S2 (см. множественную линейную регрессию).

в) Автокорреляция первого порядка
Мы предполагаем, что существует корреляция между ошибки в различных наблюдениях. В частности, мы предполагаем автокорреляцию первого порядка; то есть ошибка εi для наблюдения i зависит от ошибки εi 1 для предыдущего наблюдения. Тогда мы имеем:

$$\varepsilon i = \rho\varepsilon i \quad 1 + ui, \qquad я = 2,\dots,n,$$

где к пользовательскому интерфейсу применяются обычные гипотезы, используемые для ошибок регрессионной модели.
В этом случае дисперсионно-ковариационная матрица ошибок может быть выражена следующим образом:

$$V(\varepsilon) = \sigma 2V$$
$$= \sigma 2\,1 \frac{1}{\rho 2}$$
$$\begin{matrix} 1 & \rho & \rho 2 \dots \rho n & 1 \\ \rho & 1 & \rho \dots \rho n & 2 \\ \vdots & \rho & \dots\dots & \vdots \\ \vdots & \vdots & \dots \rho & \\ \rho n & 1\ \rho n & 2 \dots \rho\ 1 & \end{matrix}$$

Мы можем оценить V, заменив следующую оценку ρ в матрице V:

$$\hat{\rho} = \frac{\sum_{i=2}^{н} eiei \quad 1}{\sum_{я=2}^{н} e2_{я-1}} \text{(эмпирическая корреляция)},$$

где ei —оцениваемая ошибка (остатки) при обычной регрессии.
Мы можем проверить автокорреляцию первого порядка путем выполнения теста Дарбина-Ватсона на остатках (как описано выделе для последовательной корреляции).

Machine Translated by Google

ПРИМЕРЫ

Поправка на гетероскедастичность, когда данные представляют собой средства

Для проверки эффективности новой добавки на рост цыплят 40 цыплят разделили на пять групп (размером n1, n2,...,n5) и при разных дозах дополнения. Полученные данные о росте сведены в следующую таблицу:

Группа я	Количество цыплята ни	Иметь ввиду масса И	Иметь ввиду доза Си
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

Поскольку переменные представляют собой средние значения, переменная ность ошибки εi равна $\frac{\sigma^2}{ни}$.

Матрица дисперсии-ковариации ошибок затем:

$$V(\varepsilon) = \sigma^2 \begin{matrix} \frac{1}{12} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} \end{matrix} = \sigma^2 V .$$

Тогда обратная матрица V 1 имеет вид:

$$B\text{-}1 = \begin{matrix} 12 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{matrix} ,$$

а оценки параметров β0 и β1 в модели

$$Yi = \beta 0 + \beta 1 Xi + \varepsilon i$$

являются:

$$\hat{\beta} 0 = 1,23$$

$$\hat{\beta} 1 = 0,502 .$$

Этот результат аналогичен результату, полученному с использованием метода взвешенных наименьших квадратов.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Теорема Гаусса – Маркова
- Модель
- Множественная линейная регрессия
- Нелинейная регрессия
- Регрессивный анализ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бишоп, Ю.М.М., Финберг, С.Е., Холланд, PW: Дискретный многомерный анализ: теория и практика. Массачусетский технологический институт Пресс, Кембридж, Массачусетс (1975)

Грейбилл, Ф.А.: Теория и приложения Линейная модель. Даксбери, Северный округ, Массачусетс (Уолдсворт и Брукс/Коул, Пасифик-Гроув, Калифорния) (1976)

Рао, Ч.Р.: Линейный статистический вывод и Его приложения, 2-е изд. Уайли, Нью Йорк (1973)

Поколение случайных чисел

Генерация случайных чисел включает в себя создание серии чисел без узнаваемых шаблонов или закономерностей, т. е. случайный. Эти случайные числа часто затем используется для моделирования и решать задачи, которые трудно или невозможно решить аналитически, используя метод Монте Метод Карло.

Machine Translated by Google

ИСТОРИЯ

Первыми работами по генерации случайных чисел были работы фон Неймана, Джона (1951), которые ввели метод «средних четырех квадратов». Лемер (1951) ввел простой метод конгруэнтности, который используется до сих пор.

Sowey опубликовал три статьи, содержащие секретную библиографию по генерации и тестированию случайных чисел (Sowey, ER 1972, 1978 и 1986).

В 1996 г. Доджи Ядола предложили естественный метод генерации случайных чисел из цифр после запятой в числе π.

Некоторые десятичные дроби числа π представлены в Приложении А.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Существует несколько способов генерации случайных чисел. Первый заключается в вытягивании нот, пронумерованных от 0 до 9, из урну по одному, а затем положить их обратно в урну.

Некоторые физические системы могут использоваться для генерации случайных чисел: например, колесо рулетки. Полученные числа можно быть перечислены в таблице случайных чисел. Эти таблицы полезны, если требуется несколько симуляций. для разных моделей, чтобы проверить, какая из них лучше. В этом случае один и тот же ряд случайных чисел может быть

повторно использован.

Некоторые компьютеры используют электронные схемы или электромеханические средства для получения последовательности случайных чисел. Поскольку эти списки не воспроизводимы, фон Нейман, Джон и Лемер, Д.Г. предложили методы, которые являются полностью детерминированными, но дают серию чисел, которые выглядят случайными. Такие числа известны как псевдослучайные числа.

Эти псевдослучайные числа рассчитываются по заранее определенной алгебраической формуле. Например, метод Лемера позволяет вычислить серию псевдослучайных чисел x0, x1,...,xn,... из

$$xi = a \cdot xi \quad 1 \text{ (по модулю m)}, \qquad я = 1, 2, \ldots,$$

где x0 = b и a, b, m заданы.

Выбор a, b, m влияет на качество выборки псевдослучайных чисел.

Независимость между псевдослучайными числами диктует, что длина цикла, то есть количество чисел, сгенерированных до перезапуска одной и той же последовательности, должна быть достаточно большой. Для получения максимально длинного цикла необходимо выполнение следующих критериев:

- a не должно делить m, -
x0 не должно делить m, - m должно быть большим.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Если для моделирования требуются наблюдения, полученные из неравномерного распределения, существует несколько методов, которые могут генерировать любой закон из равномерного распределения. Некото
эти методы применимы ко многим дистрибутивам; другие более специфичны и используются для конкретного дистрибутива.

ПРИМЕРЫ

Используем метод Лемера для вычисления первые псевдослучайные числа серии, созданной с a = 33, b = 7 и m = 1000:

x0 = 7

x1 = 7 · 33 (mod1000) = 231 x2 = 231 · 33 (mod1000) = 623 x3 = 623 · 33 (mod1000) = 559

и т. д.

Machine Translated by Google

Ряд чисел ri, для которых распределение
существенно не отличается от
равномерное распределение получается
деление мхi на m:

$$ри = \frac{xi}{м}, \qquad я = 1, 2,...$$

Следовательно: r0 = 0,007

г1 = 0,231

г2 = 0,623

г3 = 0,559

и т. д.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Кендалл, Морис Джордж
- Метод Монте-Карло
- Случайное число
- Моделирование
- Равномерное распределение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Додж, Ю: Генераторы естественных случайных чисел, Int. Статистическое обозрение, 64, 329–344 (1996).

Рипли, Б.Д.: Мысли о псевдослучайных числах генераторы чисел. Дж Вычисл. заявл. Мат. 31, 153–163 (1990)

Генетическая статистика

В генетической статистике анализ генетических данных осуществляется с использованием методов и концепций классическая статистика, а также производная из теории случайных процессов.

Генетика изучает наследственные признаки и случайные вариации; он используется для помогите объяснить трансформизм, на практике домен, на улучшение юнитов.

ИСТОРИЯ

Генетика и статистика связаны более 100 лет. Сэр Гальтон, Фрэнсис был очень заинтересованы в изучении генетики человека, и особенно в евгенике. методов, которые можно было бы использовать для улучшения генетическое качество человеческой популяции, которое очевидно, требует знания наследственности. Гальтон также изобрел коэффициенты регрессии и корреляции, чтобы использовать их в качестве статистические инструменты при исследовании генетики. Позднее эти методы были развиты Пирсоном, Карлом. В 1900 году работа Менделя был открыт заново и стал классическим математические и статистические инструменты для инвестиций Генетика Менделя была выдвинута Сэр Фишер, Рональд Эйлмер, Райт, Сьюэлл и Холдейн, JBS в период 1920–1950 гг.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Генетическая статистика используется в таких областях, как как биоматематика, биоинформатика, биология, эпидемиология и генетика. Стандартные методы оценки и проверки гипотез используется для оценки и проверки генетических параметры. Теория случайных процессов используется для изучения единиц эволюции субъекта популяции, принимая во внимание учитывают случайные изменения частот генов.

ПРИМЕРЫ

Закон Харди-Вайнберга является центральной теоретической моделью, используемой в популяционной генетике. Сформулировано независимо английскими математик Харди, Годфри Х. и Немецкий врач Вайнберг В. (1908 г.), Уравнения этого закона показывают, что когда генетические вариации впервые появляются у некоторых особей в популяции, они не исчезают. при наследственной передаче, а вместо сохраняется в будущих поколениях в том же пропорции, соответствующие законам Менделя.

Machine Translated by Google

Закон Харди–Вайнберга позволяет нам при определенные условия, чтобывыч ислить генотипические частотыиз аллельных частот.
Под генотипом мыпонимаем генетические свойства особи, полученные при наследственной передаче от родители личности. Например, однояйцевые близнецыимеют одинаковый генотип. Ан аллель – возможный вариант гена, построенный из нуклеотидной цепи. Там прежде всего, один и тот же ген может встречаться во многих различные аллельные формывданной популяции. У человека, если каждый ген представленыдвумя аллелями, состоящими из одинаковых нуклеотидов, тогда особь гомозиготна по этому гену, и если аллели имеют разный состав, тогда человек гетерозиготен по этому ген.
Воригинальной версии закон говорит, что если ген контролируется двумя аллелями A и B, происходящие с частотами p и q = 1 p соответственно в популяции в поколениит, тогда частотыгенотипов в поколении t+1 определяются следующим уравнением:

$$p2AA + 2pqAB + q2BB.$$

Основная гипотеза состоит в том, что размер популяция достаточно велика, чтобысвести к минимуму вариации выборки; также не должно быть ни отбора, ни мутации, ни миграции (приобретения/утратыаллеля), и последовательные поколения должныбыть дискретными (без скрещивания поколений).
Чтобыполучить генотип AA , человек должен получают по одному аллелютипа A от обоих родителей. Если этот процесс случайный (т. е. гипотеза независимость), это событие произойдет с вероятностью:

$$P (AA) = pp = p2 .$$

Та же логика применима и к вероятности получение генотипа BB:

$$P (BB) = qq = q2 .$$

Наконец, для генотипа AB возможныдва случая. возможно: лицо, полученное Аот их отец и Б от матери или наоборот, так:

$$P (AB) = pq + qp = 2pq.$$

Таким образом, в идеальной популяции Пропорции Фарди – Вайнберга даны. к:

AA АБ ВВ

p2 2pq q2 .

Эту ситуациюможно обобщить на ген. со многими аллелями A1, A2, ..., Ak. частота встречаемости гомозигот (AiAi) равно:

$$f (AiAi) = p2 \quad я , \quad я = 1,...,к,$$

и частота гетерозигот (AiAj) происходит равно:

$$f AiAj = 2pipj ,$$

$$i = j, i = 1,...,k, j = 1,...,k.$$

Частота раллеля A в поколении t+1 также может представлять интерес. Путем простого подсчета имеем:

$$p = \frac{2п2Н + 2пдН}{2Н} = p2 + pq$$

знак равно п · (p + q) знак равно п · (1 - q + q) знак равно п .

Частота аллеля A в поколении t+1 идентичен частоте этого аллеля в предыдущем поколении и в начальном поколении.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Данные

Эпидемиология

Оценка

Г

Machine Translated by Google

Проверка гипотезы

Независимость

Параметр

Население

Статистика

Стохастический процесс

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Гальтон, Ф.: Типичные законы наследственности. проц. Рой. Инст. Великобритания 8, 282–301 (1877 г.) (перепечатано в: Stigler, SM (1986) The История статистики: измерение Неопределенность до 1900 года. Belknap Press, Кембридж, Массачусетс, с. 280)

Гальтон, Ф.: Соотношения и их мера. замечание, в основном из антропологических данные. проц. Рой. соц. Лонд. 45, 135–145 (1888)

Харди, Г.Х.: Менделевские пропорции в смешанное население. Наука 28, 49–50 (1908)

Смит, САВ: Статистика генетики человека. В: Коц, С., Джонсон, Н.Л. (ред.) Энциклопедия статистических наук, том. 3. Уайли, НьюЙорк (1983)

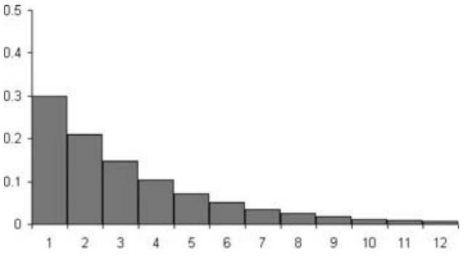
Геометрическое распределение

Рассмотрим процесс, включающий k+1 испытаний Бернулли с вероятностью успеха p и вероятность отказа q.

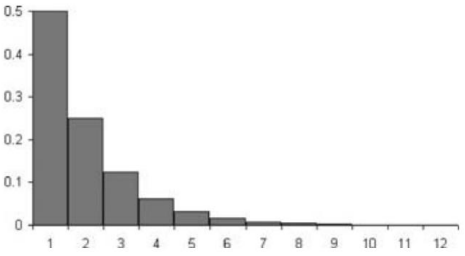
Случайная величина X, соответствующая количество отказов k, которые происходят, когда процесс повторяется до первого успеха, следует геометрическому распределению с параметром p, и поэтому мы имеем:

$$P(X = k) = p \cdot q^k .$$

Геометрическое распределение является дискретным распределение вероятностей.



Геометрическое распределение, p = 0,3, q = 0,7



Геометрическое распределение, p = 0,5, q = 0,5

Это распределение называется «геометрическим». потому что последовательные члены описанной выше функции вероятности образуют геометрическую прогрессию с отношением q.

Геометрическое распределение является особым случай отрицательного биномиального распределения где r = 1, что означает, что процесс продолжается до первого успеха.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение геометрического распределения по определению равно:

$$E[X] = \sum_{x=0} x \cdot P(X = x)$$

$$= \sum_{x=0} x \cdot p \cdot q^x$$

$$= \frac{q}{p} .$$

Действительно:

$$E[X] = \sum_{x=0} x \cdot q^x \cdot p$$

Machine Translated by Google

$$= p \cdot \sum_{x=0} x \cdot q^x \quad 1$$

$$= 0 + p\,1 + 2 \cdot q + 3 \cdot q2 + \ldots .$$

Отсюда имеем: $1+2\cdot q+3\cdot q2+\ldots = (1 \quad q)2$ $\frac{1}{}$ с:

$$д + д2 + д3 + \ldots \quad "=" \quad \frac{}{д\,1 - д}$$

$$д + д2 + д3 + \ldots \quad "=" \quad \frac{д}{1 - д\,1}$$

$$1 + 2 \cdot д + 3 \cdot д2 + \ldots \quad "=" \quad \frac{}{(1 - д)^2} \cdot$$

Дисперсия геометрического распределения по определению равна:

$$\mathrm{Var}(X) = E\,X2 \quad (E[X])^2$$

$$"=" \sum_{x=0} x2 \cdot P(X = x) - \left(\frac{д}{п}\right)^2$$

$$"=" \sum_{x=0} x2 \cdot p \cdot дx - \left(\frac{д}{п}\right)^2$$

$$= p \cdot q(1 + 22 \cdot q + 32q2 + \ldots)$$

$$\left(\frac{д}{п}\right)^2$$

$$"=" \frac{д}{п2} .$$

Действительно, сумма $1+22\cdot q+32q2+\ldots = \frac{1+q}{p2}$, с:

$$д\,1 + 2 \cdot д + 3 \cdot д2 + \ldots$$

$$"=" \quad \frac{д}{2\,(1 - д)}$$

$$1 + 22 \cdot q + 32 \cdot q2 + \ldots \; 1 + q$$

$$"=" \quad \frac{2\,(1}{q)} \quad "=" \quad \frac{1 + д}{п2} .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Геометрическое распределение относительно часто используется в метеорологических моделях. Это также используется в стохастических процессах и теории очередей ожидания.

Геометрическое распределение также может быть выражено следующем виде:

$$P(X = k) = p \cdot qk \quad 1 ,$$

где случайная величина X представляет собой количество задач, необходимых для достижения первого успеха (включая сам успех).

Г

ПРИМЕРЫ

Жребий брошен, и мы хотим знать вероятность того, что «шестерка» выпадет впервые при четвертом броске.

Таким образом, мы имеем:

$p = \frac{1}{6}$ (вероятность того, что выпадет шестерка)

$д = \frac{5}{6}$ (вероятность того, что шестерка не выпадет)

$k = 3$ (так как четвертый бросок удачный, то имеем три неудачи)

Следовательно, вероятность такова:

$$P\,(X = 3) = 6 \quad \frac{1}{} \cdot \left(\frac{5}{6}\right)^3 = 0{,}0965 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение Бернулли
- Дискретное распределение вероятностей
- Отрицательное биномиальное распределение

Среднее геометрическое

Среднее геометрическое определяется как десятый корень произведения n неотрицательных чисел.

ИСТОРИЯ

По словам Дросбеке, Ж-Ж и Tassi, Ph. (1990), среднее геометрическое и среднее гармоническое были введены Джевонсом, Уильямом Стэнли в 1874 г.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x1, x2,...,xn —набор из n неотрицательных величин или n наблюдений, связанных с количественной переменной X. Среднее геометрическое G этого набора равно:

$$G = \sqrt[n]{x1 \cdot x2 \cdot \cdots \cdot xn}.$$

Среднее геометрическое иногда выражается в логарифмической форме:

$$\text{пер}(G) = \frac{\sum_{i=1}^{n} \text{пер}(xi)}{n},$$

или

$$G = \text{опыт}\left(\frac{1}{n} \cdot \sum_{i=1}^{n} \text{пер}(xi)\right).$$

Заметим, что логарифм среднего геометрического набора положительных чисел есть среднее арифметическое логарифмов этих чисел (или взвешенное среднее арифметическое в случае сгруппированных наблюдений).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

На практике среднее геометрическое в основном используется для расчета среднего значения группы отношений или, в частности, среднего значения группы индексов.

Как и среднее арифметическое, среднее геометрическое учитывает каждое наблюдение индивидуально. Однако он уменьшает влияние выбросов на среднее значение, поэтому его иногда предпочитают арифметическому

иметь ввиду.

Одним из важных аспектов среднего геометрического является то, что оно применимо только к положительным числам.

ПРИМЕРЫ

Вот пример использования среднего геометрического: в 1987, 1988 и 1988 гг.

В 1989 году бизнесмен получает прибыль в размере 10 000, 20 000 и 160 000 евро соответственно. Основываясь на этой информации, мы хотим определить среднюю скорость, с которой эти прибыли увеличиваются.

С 1987 по 1988 год прибыль удвоилась, а с 1988 по 1989 год —в восемь раз. Если мы просто посчитаем среднее арифметическое этих двух чисел, то получим

$$x = \frac{2 + 8}{2} = 5,$$

и мы заключаем, что в среднем прибыль ежегодно увеличивалась в пять раз. Однако если мы возьмем прибыль в 1987 году (10 000 евро) и ежегодно умножим ее на пять, мы получим прибыль в 50 000 евро в 1988 году и 250 000 евро в 1989 году.

Эти две прибыли слишком велики по сравнению с реальной прибылью.

С другой стороны, если мы вычислим среднее значение, используя среднее геометрическое этих двух увеличений вместо среднего арифметического, мы получим:

$$G = \sqrt{2 \cdot 8} = \sqrt{16} = 4,$$

и мы можем правильно сказать, что в среднем прибыль увеличивалась в четыре раза ежегодно. Применяя эту среднюю норму роста к первоначальной прибыли в 10 000 евро в 1987 г., мы получаем 40 000 евро в 1988 г. и 160 000 евро в 1989 г. Даже если прибыль в 1988 г. слишком высока, она меньше предыдущего результата, а прибыль за 1989 г. теперь правильно.

Чтобы проиллюстрировать использование формулы для среднее геометрическое в логарифмической форме, найдем теперь среднее значение следующих индексов:

112, 99, 105, 96, 85, 100.

Machine Translated by Google

Рассчитываем (используя таблицу логарифмов):

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

Получаем тогда:

$$\log(G) = = 1{,}9963 \cdot \frac{11{,}9777}{6}$$

Следовательно, G = 99,15.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Среднее арифметическое
Гармоническое среднее
Иметь ввиду
Мера центральной тенденции

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Droesbeke, JJ, Tassi, PH: Histoire de la Statistique. Издания «Que sais-je?» Presses universitaires de France, Париж(1990)

Джевонс, В. С.: Принципы науки: трактат о логике и научных методах (два тома). Макмиллан, Лондон (1874 г.)

Геометрический
Среднеквадратичное отклонение

Геометрическое стандартное отклонение набора количественных наблюдений является мерой дисперсии. Он соответствует отклонениюнаблюдений от геометрической

иметь ввиду.

ИСТОРИЯ

См. оценку L1 .

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x1, x2,...,xn —набор из n наблюдений, связанных с количественной переменной X. Геометрическое стандартное отклонение, обозначаемое σg, вычисляется следующим образом:

$$\text{журнал } \sigma_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\text{журнал } x_i - \text{журнал } G)^2}.$$

где $G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$ —геометрия среднее значение наблюдений.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

оценка L1
Мера рассеивания
Среднеквадратичное отклонение

Геостатистика

Геостатистика —это приложение статистики к задачам геологии и гидрологии. Геостатистика естественным образом обрабатывает пространственные данные с известными координатами. Пространственные иерархические модели следуют принципу: моделировать локально, анализировать глобально. Простые условные модели строятся на всех уровнях иерархии (локальное моделирование). В результате получается совместная модель, которая может быть очень сложной, но анализ возможен (глобальный анализ). Геостатистика охватывает набор, используемый в теории, а также методы статистические приложения, используемые для анализа и прогнозирования распределения значений переменной в пространстве и (в конечном итоге) во вре

ИСТОРИЯ

Термин «геостатистика» впервые был использован Хартом (1954) в географическом контексте , чтобы

Machine Translated by Google

подчеркнуть применение конкретных статистических методов к наблюдениям, охватывающим региональное распространение. Первые геостатистические концепции были сформулированы Матероном (1963) в Ecole des Mines de Paris в для оценки запасов майнера al по пространственно распределенным данным (наблюдениям).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Геостатистика используется в самых разных областях, включая прогнозирование (осадков, пористости грунта, концентраций тяжелые металлы) и при обработке изображений со спутников.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Пространственные данные

Пространственная статистика

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Аттея, О., Дюбуа, Ж-П., Вебстер, Р.: Геостатистический анализ загрязнения почвы в швейцарской Юра. Окружающая среда. Загрязн. 86, 315–327 (1994)

Кресси, NAC: Статистика пространственных данных. Уайли, НьюЙорк (1991)

Гувертс, П.: Геостатистика природных Оценка ресурсов. Оксфорд Универ Сити Пресс, Оксфорд (1997)

Исаакс, Э.Х., Шривастава, РМ: Введение введение в прикладную геостатистику. Оксфорд Университетское издательство, Оксфорд (1989)

Journel, A., Huijbregts, CJ: Статистика Mining Geo. Академик, НьюЙорк (1981)

Матерон, Г.: Теория переменных региональы и другие приложения. Масссон, Париж (1965)

Oliver,MA,Webster,R.:Кригинг: amethod интерполяции для географической информации система. Междунар. Дж. Геогр. Инф. Сист. 4(3), 313–332 (1990)

Панатье, Ю: Вариовин 2.2. Программное обеспечение для анализа пространственных данных в 2D. Спрингер, Берлин Гейдельберг НьюЙорк (1996)

Джини, Коррадо

Джини Коррадо (1884–1965) родился в Матта ди Ливенца, Италия. Хотя он закончил права в 1905 году он прошел курсы математики и биологии во время учебы. Его последующие наклонности как к науке, так и к Статистика привела его к тому, что он стал временным профессором статистики в Университете Кальяри в 1909 г., а в 1920 г. он стал председателем статистики того же университета. В 1913 г. он начал преподавать в Падуанском университете и поступил на кафедру статистики в Uni Версити Рима в 1927 году. В период с 1926 по 1932 год он также был Президент Центрального статистического института (ИСТАТ). Он основал два журнала. Первый, «Метрон» (основан в 1920 г.) —международный статистический журнал, второй из Род (основан в 1934 г.), является журналом Итальянский комитет по изучениюпроблем народонаселения. Вклад Джини в область статистика в основном касается средних значений и изменчивость, а также ассоциации между случайными величинами. Он также кон отдал дань уважения оригинальным работам по экономике, социологии, демографии и биологии.

Некоторые основные работы и статьи Джини, Коррадо:

1910 Indici di concentrazione e di depen денза. Атти делла III Риунионе делла

Machine Translated by Google

Societa Italiana per il Progresso delle Scienze, R12, 3–210.

1912 г. Изменчивость и изменчивость. Студия Economico-giuridici, Università di Cagliari, III 2a, R12, 211–382.

Индекс Джини

Индекс Джини является наиболее часто используемым индексом неравенства; он позволяет нам измерить степень неравенства в распределении доходов для данного населения. Графически индекс Джини представлен площадью поверхности между линией под углом 45 и Кривая Лоренца (графическое представление совокупного процента от общего дохода по сравнению с совокупным процентом населения, получающего этот доход, где доход увеличивается слева направо). Разделив площадь поверхности между линией под углом 45° и кривой Лоренца (площадь A на следующей диаграмме) на общую площадь поверхности под линией (A + B), мы получим коэффициент Джини:

$$ИГ = \frac{A}{A + B}.$$

Коэффициент Джини равен нулю, если нет неравенство в распределении доходов и значение 1, если распределение доходов совершенно неравномерно, поэтому 0 IG 1. Индекс Джини —это просто коэффициент Джини, умноженный на 100, чтобы преобразовать его в проценты. Следовательно, чем ближе индекс к 100%, тем более неравномерно распределение доходов. В развитых странах значения индекса Джини составляют около 40% (см. пример).

ИСТОРИЯ

Коэффициент Джини был изобретен в 1912 году. итальянским статистиком и демографом Джини Коррадо.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Общая площадь поверхности под линией под углом 45 (A + B) равно 1 и поэтому мы можем написать: 2 ,

$$ИГ = \frac{A}{A + B} "=" \frac{A}{\frac{1}{2}} = 2A$$

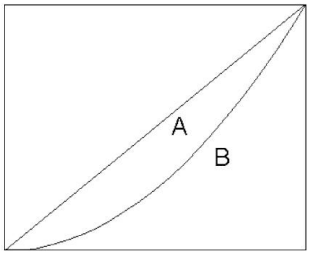
$$= 2 \quad \frac{1}{2} \quad B \quad IG = 1 \quad 2B \qquad (1)$$

Следовательно, чтобы рассчитать значение коэффициента Джини, нам нужно вычислить B (общая поверхность под кривой Лоренца).

Г

а) индекс Джини: дискретный случай

Для дискретной кривой Лоренца, где распределение доходов расположено в порядке возрастания (x1 ... xn), кривая Лоренца имеет следующий профиль:



Для вычисления B необходимо разделить поверхность под кривой Лоренца на ряд многоугольников (где первый многоугольник —треугольник, а остальные —трапеции). Поверхность треугольника:

$$\frac{1}{2н} \frac{x1}{Икс}.$$

Поверхность j-й трапеции:

$$\frac{1}{2н} \sum_{я=1}^{j-1} \frac{xi}{Икс} + \sum_{я=1}^{Дж} \frac{xi}{Икс}$$

$$"=" \frac{1}{2nX} \sum_{2}^{j-1} xи + xдж ,$$

Machine Translated by Google

с j = 2,...,n. Тогда у нас есть

$$B = \frac{1}{2n} \cdot \frac{x1}{\text{Икс}} + n \sum_{j=2} \frac{1}{2nXj} \sum_{i=1}^{j} (x_i + x_j),$$

где:

x1,...,xn — доходы; X — общий доход; x1 + ... + xn — это общая сумма выручка, в %.

После построения этого уравнения получаем:

$$B = \frac{1}{2nX} \cdot X + 2 \cdot \sum_{i=1}^{n} (n - i + 1)x_i . \quad (2)$$

Используя уравнение (1), мы можем получить индекс Джини:

$$IG = 1 - 2B = 1 - \frac{2}{2nX} X + 2 \cdot \sum_{i=1}^{n} (n - i + 1)x_i$$

$$1 = 1 + \frac{1}{n} \cdot 2 \cdot \frac{\sum_{i=1}^{n} (n - i + 1)x_i}{nX}$$

$$1 = 1 + \frac{1}{n} \cdot \frac{2}{n2\lambda(x)} \sum_{i=1}^{n} ix_n - i+1 ,$$

где $\lambda(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$ = средний доход распределения.

Коэффициент Джини может быть переформулирован во многих отношениях. Мы могли бы использовать, например:

$$IG = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_i - x_j}{2n2\lambda(x)} .$$

б) Индекс Джини: непрерывный случай. В этом случае B соответствует интегралу от 0 до 1 функции Лоренца. Поэтому нам нужно только подставить этот интеграл в уравнение. (1):

$$IG = 1 - 2B = 1 - 2 \int_0^1 L(P)\,dP ,$$

где $P = F(y) = \int_0^y f(x)\,dx$ – за процент населения с доходами меньше y. Например,

$$IG = \int_0^1 1 - (1 - P)^{\frac{\alpha - 1}{\alpha}} dP$$

$$= 1 - \frac{\alpha}{2\alpha - 1}$$

$$= 1 - 2\,1 - 2\alpha - 1 \frac{\alpha}{} .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Индекс Джини (или коэффициент) является параметром, наиболее часто используемым для измерения степени неравенства в распределении доходов. Его также можно использовать для измерения разницы различия в безработице между регионами.

ПРИМЕРЫ

В качестве примера расчета коэффициента Джини (или индекса) возьмем дискретную случай.

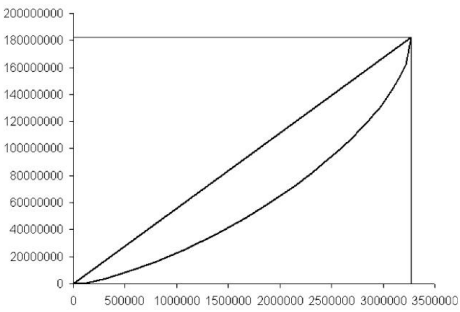
Имеем: IG = 1 2B

Напомним, что в этом случае мы находим поверхность B, вычисляя сумму ряда многоугольников под кривой Лоренца.

Данные о доходах отдельного населения за 1993–1994 гг. приведены в следующей таблице вместе с площадями, соответствующими интервалам доходов:

Machine Translated by Google

Доход населения		Площадь поверхности
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



Итак, поверхность B равна:

$$B = \frac{1}{2nX} \cdot X + 2 \cdot \sum_{i=1}^{n} (n - i + 1) \cdot x_i$$

$$= 0{,}315246894 .$$

Тогда мы имеем:

$$ИГ = 1 - 2B = 1 - 2 \cdot 0{,}315246894$$

$$= 0{,}369506212 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Джини, Коррадо
- Простой порядковый номер
- Равномерное распределение

Г

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Гринвальд, Д. (ред.): Энциклопедия экономики. Economica, Париж, стр. 164–165 (1984)

Флюкигер, Ю: Социально-экономический анализ des différances cantonales de chômage. Le chômage en Suisse: баланс и перспективы. Форум Хелвет. 6, 91–113 (1995)

Тест на пригодность

Выполнение критерия согласия на образце позволяет нам определить, наблюдаемое распределение соответствует определенному распределениювероятностей(например, нормальное распределение или распределение Пуассона). Это позволяет выяснить, является ли наблюдаемая выборка была взята из популяции, которая следует этому распределению

ИСТОРИЯ

См. тест на соответствие хи-квадрат и Тест Колмогорова-Смирнова.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Цель теста на пригодность состоит втом, чтобы определить, был ли наблюдаемый образец взятыиз совокупности , которая следует определенному распределениювероятностей.

Machine Translated by Google

Используемый процесс обычно включает некоторые проверка гипотезы:

Нулевая гипотеза H0 : Ф= Ф0

Альтернативная гипотеза H1 : F = F0

где F —неизвестная функция распределения основной популяции, а F0 – предполагаемая функция распределения. Тест на пригодность включает сравнение эмпирическое распределение с предполагаемым распределение. Мы отвергаем нулевую гипотезу, если эмпирическое распределение недостаточно близко к предполагаемому распределению Точный правила, определяющие, является ли нулевая гипотеза отклоняется или принимается в зависимости от типа используемый тест.

ПРИМЕРЫ

Существует множество тестов согласия, в том числе критерий согласия хи-квадрат и тест Колмогорова-Смирнова.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Хи-квадрат критерия согласия
- Проверка гипотезы
- Тест Колмогорова – Смирнова

Госсет, Уильям Сили

Госсет, Уильям Сили, более известный как псевдоним «Студент», родился в Кентербери, Англия, в 1876 году. математика и химия в New College, Оксфорде, он начал работать пивоваром в GuinnessBreweriesinDublinin 1899. Guinness был бизнесом, который благоприятствовал исследованиям, и сделал свои лаборатории доступными для пивовары и его химики. Действительно, в 1900 г. открыл «Исследовательскую лабораторию Гиннесса», в которой работал один из величайших химиков примерно в то время Гораций Браун в качестве директора. Работы, выполненные в этой лаборатории была связана с качеством и стоимостью много сортов ячменя и хмеля. Именно в этой среде Госсет развивал проявил интерес к статистике. В Оксфорде Госсет изучал математику, поэтому к нему часто обращались его коллеги из Гиннесса, когда возникали определенные математические проблемы. Это привело его к изучению теории ошибок, и он консультировался Пирсон К., с которым он познакомился в июле 1905 г. Эта тема. Основная трудность, с которой столкнулся Госсет во время его работы были небольшие размеры выборки. Он поэтому решил попытаться найти подходящий метод для обработки данных из эти маленькие образцы. В 1906–1907 годах Госсет провел год в сотрудничестве с Пирсоном К. в лаборатории Пирсона в Университетском колледже Лондона. попытки разработать методы, связанные с вероятная ошибка среднего. В 1907 году Госсет был назначен ответственным за Экспериментальная пивоварня Гиннесса, и он использовал таблицу Стьюдента, которую он ранее получил экспериментально, чтобы определить лучший сорт ячменя для использования для пивоварения. Гиннесс, опасаясь публикации коммерческую тайну, позволил Госсету публиковать свою работу только под псевдонимом «Ученик» или «Студент», так он и выбрал латер.

Он умер в 1937 году, оставив важные работы, которые были опубликованы под псевдонимом из «Студент».

Работы Госсета оставались в значительной степени проигнорированными для много лет, и немногие ученые использовали его стол в стороне от исследователей в Книге рекордов Гиннесса пивоварни и на экспериментальной станции Ротамстеда (где работал Фишер, Р.А. как статистик).

Machine Translated by Google

Некоторые из основных работ и статей Госсета В.С.:

1907 г. Об ошибке счета гемоцитометром. Биометрика 5, 351–360.

1908 г. Вероятная ошибка среднего. Биометрика 6, 1–25

1925 Newtables для проверки значимости наблюдений. Метрон 5, 105–120.

1942 Пирсон, Э.С. и Уишарт, Дж (редакторы) Студенческий сборник статей (предисловие Макмаллена Л.). Издательство Кембриджского университета, Кембридж (выпущено отделом биометрии Университетского колледжа Лондона).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Студенческое распределение
- Ученический стол
- Студенческий тест

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Fisher-Box, J.: Guinness, Gosset, Fisher и небольшие образцы. Стат. науч. 2, 45–52 (1987)

Макмаллен, Л., Пирсон, Э.С.: Уильям Сили Госсет, 1876–1937. В: Пирсон, Э.С., Кендалл, М. (ред.) Исследования по истории статистики и вероятностей, том. И. Гриффин, Лондон (1970)

Плакетт, Р. Л., Барнард, Джорджия (ред.): Студент: Статистическая биография Уильяма Сили Госсета (на основе работ Э. С. Пирсона). Кларендон, Оксфорд (1990)

Греко-латинские квадраты

См. Греко-латинский квадратный дизайн.

Греко-латинский квадратный дизайн

Греко-латинская квадратная схема — это схема эксперимента, в которой экспериментальные единицы сгруппированы тремя различными способами. Получается путем наложения двух латинских квадратов одинакового размера.

Если каждая латинская буква ровно один раз совпадает с греческой буквой, два латинских квадрата ортогональны. Вместе они образуют греко-латинскую квадратную форму.

В этом дизайне каждое обращение (латинская буква) появляется только один раз в каждой строке, один раз в каждом столбце и один раз с каждой греческой буквой.

ИСТОРИЯ

Построение греко-латинского квадрата было начато Эйлером, Леонгардом (1782).

Книга Фишера, Рональда Эйлмера и Йейтса, Франка (1963) дает греко-латинские таблицы от порядка 3 до порядка 12 (не включая порядка шести). Книга Денеса и Кидуэлла (1974) также содержала всестороннюю информацию о греко-латинских квадратных рисунках. Доджи Шах (1977) рассматривают случайоценки при отсутствии данных.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Греко-латинские квадраты используются для уменьшения влияния трех источников систематических ошибок.

Существуют греко-латинские квадраты для любого размера n, кроме n = 1, 2 и 6.

ПРИМЕРЫ

Мы хотим протестировать три разных типа бензина. Для этого у нас есть три водителя и три автомобиля. Однако тест слишком сложен, чтобы проводить его за один день: нам нужно провести эксперимент в течение трех дней.

В этом случае использование водителей и транспортных средств свыше

Machine Translated by Google

три отдельных дня могут привести к систематическим ошибкам. Поэтому, чтобы устранить эти ошибки, мы строим экспериментальный план греко-латинского квадрата. В этом случае каждый вид бензина (обработка) будет тестироваться только один раз с каждым водителем, один раз с каждым транспортным средством и один раз в день.

В этой греко-латинской квадратной схеме мы обозначаем различные типы бензина латинскими буквами A, B и C, а разные дни — буквами α, β и γ.

Транспорт	Драйверы 1	2	3
1	Aα	Bβ	Cγ
2	Bγ	Cα	Aβ
3	Cβ	Aγ	Bα

Затем дисперсионный анализ покажет нам, после устранения влияния линий (транспортных средств), столбцов (водителей) и греческих букв (дней) существует ли значительная разница между типами бензина.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- План экспериментов
- Латинские квадратные узоры

ССЫЛКИ

Dénes, J., Keedwell, AD: Латинские квадраты и их приложения. Академик, НьюЙорк (1974)

Додж, Ю, Шах, К.Р.: Оценка параметров в латинских квадратах и греко-латинских квадратах с отсутствующими наблюдениями. Коммун. Стат. Теория А6(15), 1465–1472 (1977)

Эйлер, Л.: Recherches sur une nouvelle espèce de carrés magiques. Верх. Zeeuw. Генерал Ветен. Флиссинген 9, 85–239 (1782 г.)

Фишер, Р.А., Йейтс, Ф.: Статистические таблицы для биологических, сельскохозяйственных и медицинских Исследования, 6-е изд. Хафнер (Макмиллан), НьюЙорк (1963)

Графическое представление

Графические представления охватывают широкий спектр методов, которые используются для уточнения, интерпретации и анализа данных путем нанесения точек и сегментов линий, поверхностей и других геометрических форм или символов.

Назначение графика — быстрая визуализация набора данных. Например, он должен четко иллюстрировать общее поведение исследуемого явления и выделять любые важные факторы. Его можно использовать, например, как средство перевода или заполнения частотной таблицы.

Следовательно, графическое представление является формой представления данных.

ИСТОРИЯ

Концепция построения точки в координатном пространстве восходит, по крайней мере, к древним грекам, но нам пришлось ждать до работ Декарта, Рене, чтобы математики исследовали эту концепцию.

Согласно Royston, E. (1970), немецкому математику по имени Crome, AW был одним из первых, кто использовал графическое представление в статистике. Первоначально он использовал его в качестве учебного пособия.

В своих работах Geographisch-statistische Darstellung der Staatskräfte (1820) и Ueber die Grösse und Bevolkerung der sämtlichen Europäischen Staaten (1785) Кром использовал различные типы графического представления, в том числе круговую диаграмму.

Royston, E. (1970) также цитирует Playfair, W., чья работа The Commercial and Polit

В атласе (1786 г.) также упоминались различные

Machine Translated by Google

графические изображения, особенно линия диаграмма. Playfair был очень заинтересован в балансе между народной торговли и иллюстрированный его исследования с использованием различных графиков, таких как гистограмму и круговую диаграмму .

Термин «гистограмма» впервые был использован для время Пирсона, Карла и Герри (1833 г.) по-видимому, был одним из первых, кто использовал линейный график; в своем «Опыте о статистике морали Франции», опубликованном в Париже в 1833 году, он описал частоты преступлений по их признакам; эта учеба представляет собой одно из первых применений частотного распределения.

Шмид, Кальвин Ф. (1954) иллюстрирует и описывает различные типы графических представлений в своем Handbook of Graphic Презентация.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Тип используемой графики зависит от тип данных , которые должны быть представлены, характер изучаемой переменной, а цель изучать:

- Количественный график особенно полезен для представления качественных категориальных переменных. К такой графике относятся круговая диаграмма, линейная диаграмма и пиктограмма.
- Частотный график позволяет нам представить (дискретное или непрерывное) частотное распределение количественного переменная. К такой графике относится его тограмма и схема стебля и листа .
- Декартова графика, использующая систему оси перпендикулярны друг другу и пересекаются в точке, известной как «начало». Такая система координат называется «картезианский».

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Таблица частот
- Количественный график

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кром, AFW: Ueber die Grösse und Bevölkerung der sämtlichen Europä ischen Staaten. Вейганд, Лейпциг (1785 г.)

Кром, AFW: Географически-статистика Darstellung der Staatskräfte. Вейган, Лейпциг (1820 г.)

Fienberg, SE: Графический метод в статистике . Являюсь. Стат. 33, 165–178 (1979)

Герри, AM: Essai sur la statistique моральный дух де ла Франс. Крошар, Париж(1833 г.)

Playfair, W.: Коммерческий и политический Атлас. Playfair, Лондон (1786 г.)

Ройстон, Э .: Заметка об истории графическое представление данных. В: Грушевый сьн, Э .С., Кендалл, М. (ред.) Исследования в История статистики и вероятностей, том. Я. Гриффин, Лондон (1970)

Шмид, К. Ф.: Справочник по графическому представлению Рональд Пресс, НьюЙорк (1954)

Г

Machine Translated by Google

Хайек, Ярослав

Хайек, Ярослав родился в 1926 году в Поде Брады, Богемия. Инженер-статистик по профессии, он получил докторскую степень в 1954 году. С 1954 по 1964 год он работал научным сотрудником в Институте математики.

Чехословацкой академии наук. Затем он поступил в Карлов университет в Праге, где был профессором с 1966 г. до своей преждевременной смерти в 1974 г. Основные работы Хайека Дж. касаются теории вероятностей выборки и теории ранговых тестов. В частности, он разработал асимптотическую теорию статистики линейных рангов. Он был первым, кто применил концепцию инвариантности к теории ранговой проверки.

Некоторые основные работы и статьи Хайека, Ярослав:

1955 Некоторые ранговые распределения и их приложения. Кас. Пешт. Мат. 80, 17–31 (на чешском языке); перевод в (1960) Выбирать. Перевод Мат. Стат. Вероятно., 2, 41–61.

1958 г. Некоторые вклады в теорию вероятностной выборки. Бык. Инст. Междунар. стат., 36, 127–134.

1961 Некоторые расширения Вальда–Вулфовица–Нётер. Анна. Мат. Стат. 32, 506–523.

1964 г. Асимптотическая теория отбраковочной выборки с различными вероятностями из конечной популяции. Анна. Мат. Стат. 35, 1419–1523.

1965 г. Распространение теста Колмогорова-Смирнова на регрессионные альтернативы. В: Нейман, Дж. и Лекам, Л. (ред.) Бернулли-Байес-Лаплас: Материалы международного семинара, 1963. Springer, Берлин, Гейдельберг, НьюЙорк, стр. 45–60.

1981 Дюпак, В. (ed) Выборка из конечной совокупности. Марсель Деккер, НьюЙорк.

Гармоническое среднее

Гармоническое среднее n наблюдений равно определяется как n, деленное на сумму инверсий всех наблюдений.

ИСТОРИЯ

См. среднее геометрическое.

Соотношение между средним гармоническим, средним геометрическим и средним арифметическим описано Mitrinovic, DS (1970).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x1, x2,...,xn —n ненулевых величин или n наблюдений, относящихся к количественному

Machine Translated by Google

переменная X. Гармоническое среднее H этих n величин рассчитывается следующим образом:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Если $\{x_i\}_{i=1,\ldots,n}$ представляет собой конечный ряд положительных чисел, мы утверждаем, что:

$$\min_i x_i \quad H \quad G \quad \bar{x} \quad \max_i x_i.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Среднее гармоническое редко используется в статистике. Тем не менее, иногда это может быть полезно, например, в следующих случаях:

- Если набор инвестиций инвестируется по разнице различные процентные ставки, и все они дают одинаковый доход, уникальная ставка, по которой все капитала, связанного с этими инвестициями необходимо инвестировать, чтобы получить тот же доход, что и совокупность инвестиций. равно среднему гармоническому для индивидуальных скоростей.
- Допустим, у нас есть группа различных материалов, и каждый материал можно купить. по указанной цене за количество материала (где цена за количество может быть разной для каждого материала). Затем мы покупаем определенное количество каждого материала, тратя одинаковое количество денег на каждом. В этом случае средняя цена за количество по всем материалов задается средним гармоническим цен за количество для всех материалов.
- Одно из свойств гармонического среднего состоит в том, что он в значительной степени нечувствителен к выбросам, которые имеют гораздо большие значения, чем другие данные. Например, рассмотрим следующие значения: 1, 2, 3, 4, 5 и 100. Здесь среднее гармоническое равно 2,62, а среднее арифметическое равно 19,17. Однако гармоническое среднее гораздо более чувствительно к выбросам, когда они имеют гораздо меньшие значения, чем остальные данные. Таким образом, для наблюдения 1, 6, 6, 6, 6, 6, получаем H = 3,27, тогда как среднее арифметическое равняется 5,17.

ПРИМЕРЫ

Три инвестиции, каждая из которых дает одинаковую прибыль доход имеют следующие процентные ставки: 5%, 10% и 15%.

Среднее гармоническое дает процентную ставку при котором весь капитал должен быть инвестируется для того, чтобы произвести тот же общий доход в виде трех отдельных инвестиций:

$$H = \frac{3}{\frac{1}{5} + \frac{1}{10} + \frac{1}{15}} = \frac{3}{\frac{11}{30}} = 8{,}18\%.$$

Отметим, что этот результат отличается от среднее арифметическое 10% (5 + 10 + 15)/3.

Представитель покупает три партии кофе, каждый разного сорта (качества), на 3, 2 и 1,5 евро за кг соответственно. Он покупает 200 евро каждого сорта.

Тогда средняя цена за кг кофе получается путем деления общей стоимости на общую купленное количество:

$$\text{средняя цена} = \frac{\text{Общая стоимость}}{\text{общее количество}} = \frac{3 \cdot 200}{66{,}66 + 100 + 133{,}33} = 2.$$

Это соответствует среднему гармоническому цены на три разных сорта кофе:

$$\text{средняя цена} = \frac{3}{\frac{1}{3} + \frac{1}{2} + \frac{1}{1{,}5}} = \frac{6}{3} = 2.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Среднее арифметическое
- Среднее геометрическое

Machine Translated by Google

Иметь в виду

Мера центральной тенденции

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Митринович, Д.С. (совместно с Васичем, П.М.):
Аналитические неравенства. Springer, Берлин
Гейдельберг, НьюЙорк (1970)

Шляпа Матрица

Матрица шляпы—это матрица, используемая в регрессионном и дисперсионном анализе. Он определяется как матрица, которая преобразует значения из наблюдаемой переменной в оценки получен методом наименьших квадратов. Поэтому при проведении линейной регрессии в матричной форме, если Yˆ —вектор формируется из оценок, рассчитанных по параметры наименьших квадратов, а Y —вектор наблюдений, относящихся к зависимой переменной, то Yˆ задается вектором Y, умноженным на H, т. е. Yˆ = HY преобразуется из Y в Y's.

ИСТОРИЯ

Матрица шляпы H была введена Тьюки, Джоном Уайлдером в 1972 году. В статье Хоаглина, Д.К. и Уэлша, Р.Э. (1978) приводятся свойства матрицы H, а также множество примеров ее применения.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим следующую модель линейной регрессии:

$$Y \text{ знак равно Икс} \cdot \beta + \varepsilon .$$

где

Y —(n × 1) вектор наблюдений над зависимой переменной;

X – (n × p)-матрица независимых переменных (независимых переменных p);

ε - (n × 1) вектор ошибок, и;

β —(p × 1) вектор оцениваемых параметров.

Оценка βˆ вектора β определяется выражением

$$\beta\hat{} = X \cdot X^{-1} \cdot X \cdot Y .$$

и мы можем вычислить оценочные значения Yˆ для Y, если мы знаем βˆ:

$$Y\hat{} = X \cdot \beta\hat{} = X \cdot X \cdot X^{-1} \cdot X \cdot Y .$$

Затем матрица H определяется следующим образом:

$$H = X \cdot X \cdot X^{-1} \cdot X .$$

В частности, диагональный элемент hii будет определяться следующим образом:

$$hii = xi \cdot X \cdot X^{-1} \cdot x_i .$$

где xi —i-я строка X.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Матрица H, позволяющая получить n оценок зависимой переменной из n наблюдений, представляет собой идемпотентную юсимметричную квадратную юматрицу порядка n. Элемент (i, j) этой матрицы измеряет влияние j-е наблюдение над i-м прогнозируемым значением. В частности, диагональные элементы оценивают влияние наблюдений на правильную соответствующие оценки зависимых переменных. Значение каждого диагонального элемента матрицы H находится в диапазоне от 0 до 1. Записывая H = (hij) для i, j = 1,...,n, имеем соотношение:

$$\text{хии} = \text{ч}\,\frac{2}{\text{II}} + \sum_{\substack{\text{я}=1 \\ \text{я}=\text{j}}}^{\text{н}} h2_{\text{и}} .$$

который получен на основе идемпотентного характера H; другими словами, H = H2.

ЧАС

Machine Translated by Google

Поэ тому $tr(H) = \sum_{i=1}^{n} h_{ii} = p$ = колич е ство параме трыдля оце нки.

Матрица H использ у е тся для опре де ле ния точ е к воз де йствия в ре г рессионном анализ е .

ПРИМЕ РЫ

Рассмотрим сле ду ющу ютабл ицу , г де Y — з ависимая пе ре ме нная , связ анная с не з ависимой пе ре ме нной X:

Икс	Д
50	6
52	8
55	9
75	7
57	8
58	10

Моде л ь простой л ине йной ре г ре ссии з аписывае тся в матрич ной форме сле ду ющим образ ом:

$$Y = X \cdot \beta + \varepsilon .$$

г де

$$Y = \begin{matrix} 6 \\ 8 \\ 9 \\ 7 \\ 8 \\ 10 \end{matrix}, \quad X = \begin{matrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 75 \\ 1 & 57 \\ 1 & 58 \end{matrix} .$$

ε —(6 × 1) ве ктор ошибок, β —(2 × 1) ве ктор параме тров.

Нахо дим матрицу H, использ у я ре з у л ьтат:

$$H = X \cdot (X \cdot X)^{-1} \cdot X .$$

Пу те м пошаг овых матрич ных вы чис ле ний пол у ч ае м:

$$(X \cdot X)^{-1} = \frac{1}{2393} \cdot \begin{matrix} 20467 & 347 \\ 347 & 6 \end{matrix}$$

$$= \begin{matrix} 8,5528 & 0,1450 \\ 0,1450 & 0,0025 \end{matrix} .$$

и наконе ц:

$$H = X \cdot (X \cdot X)^{-1} \cdot X$$

$$= \begin{matrix} 0,32 & 0,28 & 0,22 & -0,17 & 0,18 & 0,16 \\ 0,28 & 0,25 & 0,21 & -0,08 & 0,18 & 0,16 \\ 0,22 & 0,21 & 0,19 & 0,04 & 0,17 & 0,17 \\ 0,17 & 0,08 & 0,04 & 0,90 & 0,13 & 0,17 \\ 0,18 & 0,18 & 0,17 & 0,13 & 0,17 & 0,17 \\ 0,16 & 0,16 & 0,17 & 0,17 & 0,17 & 0,17 \end{matrix} .$$

Заме тим, например, ч то ве с у1 , испол ьз ованный при оце нке у^1 , раве н 0,32. Зате м мы прове ря е м, ч то трасса Не равна 2, дру г ими сл овами, она равна кол ич е ству параме тров моде л и .

$$tr(H) = 0,32 + 0,25 + 0,19 + 0,90 + 0,17 + 0,17 = 2 .$$

ДАЛЬНЕ ЙШЕ Е Ч ТЕ НИЕ

- Точ ка рыч ага
- Матрица
- Множе ственная л ине йная ре г ре ссия
- Ре г ре ссивный анал из
- Простая л ине йная ре г ре ссия

ССЫЛ КИ

Belsley, DA, Kuh, E., Welsch, RE: Диаг ностика ре г ре ссии. У айл и, НьюЙорк, стр. 16–19 (1980).

Hoaglin, DC, Welsch, RE: Матрица шл япы в ре г ре ссии и ANOVA. Явл якось. Стат. 32, 17–22 (и испр авл е ние на 32, 146) (1978)

Tukey, JW: Не которые г рафич е ские и пол у г рафич е ские дисп л е и. В: Бэ нкрофт, Т.А. (ре д.) Статистич е ские статьи в ч е сть Джорджа У . Сне де кор. Из дате л ьство г осу дарстве нного у ниве рсите та Айовы, Э ймс, Айова, стр. 293–316 (1972).

Гистограмма

Г истог рамма пре дстав л яе т собой г рафич е ское пре дстав л е ние распре де л е ния данны х , которые бы л и

Machine Translated by Google

сгруппированы в классы. Он состоит из ряда прямоугольников и представляет собой тип частотной диаграммы. Каждое значение данных сортируется и помещается в соответствующий интервал класса. Количество значений данных в каждом интервале классов определяет частота (или относительная частота) этого интервала класса.

Каждый прямоугольник на гистограмме представляет класс данных. Ширина прямоугольника соответствует ширине интервала классов, а поверхность прямоугольника представляет вес класса.

ИСТОРИЯ

Термин гистограмма был впервые использован Карлом Пирсоном в 1895 году.

Также см. графическое представление.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Первый шаг в построении гистограммы состоит в представлении данных в виде таблицы частот.

Для этого необходимо, чтобы интервалы классов были установлены, а значения данных отсортированы и размещены в классах, что дает возможность вычислить частоты занятий.

Затем интервалы классов и частоты добавляются в таблицу частот.

Затем мы используем таблицу частот для построения гистограммы. Разобьем горизонтальную ось гистограммы на интервалы, причем ширина этих интервалов соответствует ширине интервалов классов. Затем на гистограмме рисуем прямоугольники.

Ширина каждого прямоугольника равна ширине класса, которому он соответствует. Высота прямоугольника такова, что площадь поверхности прямоугольника равна относительной частоте соответствующего класса.

Сумма площадей поверхности прямоугольника gles должен быть равен 1.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Гистограммы используются для представления набора данных в визуальной форме, которую легко понять. Они позволяют визуализировать и извлекать определенные общие характеристики (такие как типичные значения, диапазон или форма данных).

Просмотр формы гистограммы может позволить нам обнаружить модель вероятности, за которой следуют данные (нормальное распределение, логарифмически-нормальное распределение и т. д.).

Также возможно обнаружить непредвиденное поведение или аномальные значения с помощью гистограммы.

Этот тип графического представления чаще всего используется в экономике, но поскольку это чрезвычайно простой способ визуализации набора данных, он используется и во многих других областях.

Мы также можем проиллюстрировать относительную частоту на гистограмме. В этом случае высота каждого прямоугольника равна относительной частоте соответствующего класса, деленной на длину интервала классов. В этом случае, если мы суммируем площади всех прямоугольников на гистограмме, мы получим единицу.

ПРИМЕРЫ

В следующей таблице приведены необработанные данные о среднегодовом количестве осадков в 69 городах США. Мы будем использовать эти данные для установления таблицу частот, а затем соответствующую гистограмму.

Среднегодовое количество осадков в 69 городах США (в дюймах)

Моб. 67.0	Шик. 34,4	ст.л. 35,9	
Сакр. 17.2	Лу и. 43.1	Ат.К. 45,5	
Вашингтон 38,9	Де тр. 31,0	симв. 42,7	
Бойсе 11,5	КС 37,0	Кол. 37,0	
Wich. 30.6	Конц. 36.2	Прил. 42,8	
Бост. 42,5	НьюЙорк 40,2	Дал л. 35,9	
Джек. 49.2	Клев. 35.0	Норф. 44,7	
Рено 7.2	Питт.	36.2	Чей. 14,6

Machine Translated by Google

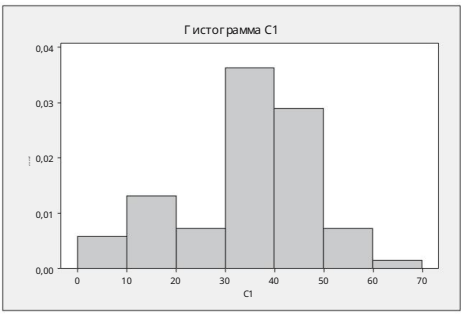
Бафф.	36.1 Нэш. 46,0 ЛР		48,5
Цинк.	39,0 Берл.	32,5 Харт.	43,4
Фил.	39,9 Млв.	29.1 Атл.	48,3
Мемп. 49.1 Фен. 7.0 Инди.			38,7
СЛ	15.2 Денв. 13.0 Порт.		40,8
Чар.	40,8 Майами 59,8 Дул.		30,2
июнь	54.7 Пер.	35,1 G.Fls 15,0	
СФ	20,7 Н.Ор. 56,8 Альбк		7,8
Джек.	54,5 ССМ 31,7 Рал.		42,5
Спок. 17,4 ЛА		14.0 Омаха 30.2	
Ок.К. 31.4 Вильм. 40,2 Альб.			33,4
полковник	46.4 Дростопочтенный. 22.9 Бизм. 16.2		
Эль-Пасо 7.8 D.Mn. 30.8 Порт.			37,6
S-Tac 38.8 Балт.		41,8 S.Fls	24,7
СК	59,2 Мн-СП 25,9 Хстн. 48,2		

Источник: Бюро переписи населения США (1975 г.), Статистическое резюме Соединенных Штатов, перепись населения США, Бюро, Вашингтон, округ Колумбия.

Эти данные могут быть представлены следующей таблицей частот:

Класс	Частота	Относительная частота
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
Общий	69	1.000

Теперь мы можем построить гистограмму:



Горизонтальная ось разделена на классы, и в этом случае относительные частоты равны задается высотами прямоугольников, потому что все классы имеют одинаковую ширину.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение частоты
- Полигон частот
- Таблица частот
- Графическое представление
- оживает
- Схема стебля и листа

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Додж, Ю: Некоторые трудности, связанные с непараметрической оценкой функции плотности. Дж Оффик. Стат. 2 (2), 193–202 (1986)

Фридман, Д., Пизани, Р., Пурвес, Р.: Статистика. Нортон, НьюЙорк (1978)

Пирсон, К.: Вклад в математическую теорию эволюции. II: перекос изменение однородного материала. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, Кембридж, стр. 41–112 (1948). Впервые опубликовано в 1895 г. в Филосе. Транс. Рой. соц. Лонд. сер. А 186, 343–414

Тест на однородность

Один вопрос, который часто необходимо учитывать при анализе категорийных данных, получ ченных для многих групп — это однородность групп. Другими словами, нам нужно выяснить, есть ли существенные различия между этими группами по одной или многим качественным категориальным переменным. Тест на однородность может показать существенны ли различия или нет.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Мы рассматриваем критерий однородности хи-квадрат здесь, который является особым типом хи-квадрата тест.

Пусть I будет числом рассматриваемых групп и J количество рассматриваемых категорий.

Позволять:

$$ни. = \sum_{j=1}^{Дж} nij$$ соответствует размеру группы я;

ндж= ний; я

$$п = \sum_{я=1}^{я} \sum_{j=1}^{Дж} nij$$ —общее количество наблюдения

nij —эмпирическая частота (т. е. количество наблюдаемых явлений), соответствующих группе i и категории j;

mij – теоретическая частота, соответствующая группе i и категории j, которая, предполагая однородность среди группы, равно:

$$мидж = \frac{ни. \cdot ньойорк}{н}.$$

Если представить данные в виде таблицы сопряженности с I строками и J столбцами, мы можем вычислить ni. которые способствуют сумма элементов строки i и nj , которые вклад в сумму все х элементов столбца j.

Мы рассчитываем

$$x2 "=" \sum_{я=1}^{я} \sum_{j=1}^{Дж} \frac{нидж- мидж2}{Мидж}.$$

а критерий однородности хи-квадрат равен выражается следующим образом: мы отвергаем гипотеза однородности (при значимости уровне 5%), если значение $\chi 2_c$ больше, чем значение распределения $\chi 2$ (хи-квадрат) с (J 1) · (I 1) степенями свободы.

Примечание: здесь мы исходили из того же количество единиц тестируется для каждой комбинации группы и категории. Однако мы можем хотите протестировать различное количество единиц для dif разные комбинации. В этом случае, если мы имеем пропорции pij единиц для группы i и категории j достаточно заменить mij на ni. · пиж.

ПРИМЕРЫ

Мы рассматриваем исследование, проведенное в области фармацевтики и касающееся 100 человек, страдающих определенным заболеванием. В для проверки действия медицинского лечения случайным образом было выбрано 100 человек. Половина из них была помещена в контрольную группу и получала плацебо. Остальные пациенты получали медикаментозное лечение. Тогда число здоровых людей в каждом группа находилась под наблюдением в течение 24 часов после администрирование лечения. Результаты представлены в следующей таблице.

Наблюдаемый частота	Здоровый на 24 часы	Нет здоровый	Общий
Плацебо	2	48	50
Уход	9	41	50

Теоретические частоты получаются при условии, что общее состояние здоровья было бы одинаковым для обеих групп, если бы лечение не применялось. В этом случае получаем m11 = m21 = = 5,5 и $\frac{50 \cdot 11}{100}$ м12 = м22 = $\frac{50 \cdot 89}{100}$ = 44,5.

Мы вычисляем значение $\chi 2$, сравнивая теоретические частоты с наблюдаемыми частоты:

$$x2 "=" \frac{(2 \quad 5{,}5)^2}{5{,}5} + \frac{(48 \quad 44{,}5)^2}{44{,}5} + \frac{(9 \quad 5{,}5)^2}{5{,}5} + \frac{(41 \quad 44{,}5)^2}{44{,}5}$$

$$= 5{,}005.$$

Machine Translated by Google

Если затем обратиться к таблице распределения χ^2 для одной степени свободы, то получим значение χ^2 = 3,84 для уровня значимости 5%, $_{0,05}$ что меньше рассчитанного нами значения χ^2 = 5,005. Делаем вывод, что группы неоднородны и лечение эффективно.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Анализ категорийных данных
- Дисперсионный анализ
- Категориальные данные
- Распределение хи-квадрат
- критерий хи-квадрат
- Частота

ССЫЛКА

См. анализ категорийных данных.

Хотеллинг, Гарольд

Гарольд Хотеллинг (1895–1973) считается одним из пионеров в области экономической математики в период 1920–1930 годов. Он представил многомерный критерий T^2, анализ главных компонент и канонический корреляционный анализ.

Он учился в Вашингтонском университете, где в 1919 году получил степень бакалавра журналистики. Затем он перешел в Принстонский университет. В 1924 году он получил там докторскую степень по математике. В том же году он начал преподавать в Стэнфордском университете. Его приложения математики к социальным наукам сначала касались журналистики и политологии, а затем он сосредоточил свое внимание на населении и прогнозировании.

В 1931 г. он перешел в Колумбийский университет, где активно участвовал в создании его статистического факультета. В течение Во время Второй мировой войны он проводил статистические исследования для военных. В 1946 году он был нанят Университетом Северной Каролины в Чапел-Хилл для создания там отдела статистики.

Некоторые основные работы и статьи Hotel ling, Harold:

1933 Анализ комплекса статистических переменных с главными компонентами.
Дж Образ. Psychol., 24, 417–441 и 498–520.

1936 Отношение между двумя наборами переменных.
Биометрика 28, 321–377.

Хубер, Питер Дж

Хубер, Питер Дж родился в Волене (Швейцария) в 1934 г. Он блестяще показал себя во время учебы и защиты докторской диссертации по математике в Федеральной политехнической школе Цюриха, где получил Серебряную медаль за научное качество своей диссертации. Он работал профессором математической статистики.

тики в Федеральной политехнической школе Цюриха. Затем он переехал в США и работал в самых престижных университетах (Принстон, Йель, Беркли) в качестве приглашенного профессора. В 1977 году ему было присвоено звание профессора Прикладная математика в Массачусетском технологическом институте. Он является членом престижной Американской академии искусств и наук, Общества Бернулли и Национального научного фонда США, в которых иностранные члены крайне редки.

С момента публикации его статьи «Надежная оценка параметра местоположения» в 1964 году он считается основателем надежная статистика.

Machine Translated by Google

Хубер, Питер Дж получил звание доктора HonorisCausa из Невшательского университета в 1994.

Некоторые основные работы и публикации Хубер, Питер Дж:

1964 г. Надежная оценка местоположения параметр. Анна. Мат. Стат. 35, 73–101.

1968 Надежная статистика процедуры. SIAMCBMS-NSF Рег. конф. сер. заявл. мат.,

1981 Надежная статистика. Уайли, НьюЙорк.

1995 Надежность: где мы сейчас? Стью вмятина, Том 1, 75–86.

Гипергеометрическое распределение

Гипергеометрическое распределение описывает вероятность успеха, если серия объектыберутся из населения (которое содержит некоторые объекты, которые представляют собой отказ в то время как другие представляют собой успех), без замена.
Поэтому он используется для описания случайного эксперимент, вкотором возможнытолько два результата: «успех» и «неудача».
Рассмотрим множество N событий, в которых M «успехов» и N M «неудачи».
случайная величина X, соответствующая количество успехов, полученных, если мынарисуем n события без замены подчиняются гипергеометрическому распределениюс параметрами N, M и n, обозначаемые через H(N, M, n).
Гипергеометрическое распределение является дискретным распределением вероятностей.
Количество способов, которыми n событий могут быть из N событий равно:

$$C_{H}^{n} = \binom{H}{n} = \frac{H!}{H! \cdot (n \quad N)!}.$$

Количество элементарных событий зависит на X и есть:

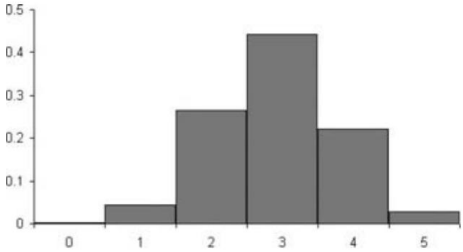
$$C_{M}^{x} \cdot C_{H-M}^{n-x}.$$

что дает следующую функциювероятности:

$$P(X = x) = \frac{C_{M}^{x} \cdot C_{H-M}^{n-x}}{C_{H}^{n}},$$

для x = 0, 1,...,n

(где $C_{в}^{u} = 0$, если u < v по соглашению).



Гипергеометрическое распределение, N = 12, M = 7, n = 5

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим случайную величину X = X1 + X2 + ... + Xn, где:

$$C_i = \begin{cases} 1, & \text{если i-й рисунок удался} \\ 0, & \text{если i-й рисунок провален} \end{cases}.$$

В этом случае распределение вероятностей Ci это:

Ci	0	1
n (Ci)	$\frac{H - M}{H}$	$\frac{M}{H}$

Таким образом, ожидаемое значение Xi дано к:

$$E[C_i] = \sum_{j=1}^{2} x_j P(x_i = x_j) = 0 \cdot \frac{H - M}{H} + 1 \cdot \frac{M}{H} = \frac{M}{H}.$$

Machine Translated by Google

Используя тот факт, что X = X1+X2+...+Xn, у нас есть:

$$E[X] = \sum_{i=1}^{n} E[Ci] = \sum_{i=1}^{n} \frac{M}{N} = n \frac{M}{N}.$$

Дисперсия Xi по определению:

$$Var(Xi) = E X_i^2 - (E[Ci])^2$$
$$= \frac{M}{N} - \left(\frac{M}{N}\right)^2$$
$$= \frac{M(N-M)}{N^2}.$$

Так как Xi, i = 1, 2,...,n зависят случайные величины, ковариация должна учитывать при расчете отклонение X._

Вероятность того, что Xi и Xj (i = j) оба успеха равны:

$$P(Xi = 1, Xj = 1) = \frac{M(M-1)}{N(N-1)}.$$

Если положить V = Xi · Xj, значения V и ассоциированные вероятности:

B	0	1
p(B)	$1 - \frac{M(M-1)}{N(N-1)}$	$\frac{M(M-1)}{N(N-1)}$

Таким образом, ожидаемое значение V равно:

$$E[V] = 0 \cdot \left(1 - \frac{M(M-1)}{N(N-1)}\right) + 1 \cdot \frac{M(M-1)}{N(N-1)}.$$

Ковариация Xi и Xj по определению равна равно:

$$Cov(Xi, Xj) = E[Xi \cdot Xj] - E[Xi] \cdot E[Xj]$$
$$= \frac{M(M-1)}{N(N-1)} - \left(\frac{M}{N}\right)^2$$
$$= \frac{M(N-M)}{N^2(N-1)}$$
$$= -\frac{1}{N-1} Var(Kси).$$

Теперь мы можем вычислить дисперсию X:

$$Var(X) = \sum_{i=1}^{n} Var(Kси) + 2 \sum_{j=1}^{n} \sum_{i<j} Cov(Xi, Xj)$$
$$= n\,Var(Xi) + n(n-1)Cov(Xi, Xj)$$
$$= n\,Var(Xi) - \frac{n-1}{N-1} Var(Kси)$$
$$= n\,Var(Xi) - \frac{N-n}{N-1}$$
$$= n \frac{M(N-M)}{N^2} \frac{N-n}{N-1}$$
$$= \frac{N-n}{N-1} n \frac{M}{N}\left(1 - \frac{M}{N}\right).$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Гипергеометрическое распределение часто используется в контроле качества.

Предположим, что производственная линия производит N продукции, которые затем отправляются на проверку. Из него берется выборка размера n. партию продукции и количество бракованной продукции в этом образце. Можно использовать это, чтобы получить (путем логического вывода) сведения о вероятном общем количестве бракованная продукция во всей партии.

ПРИМЕРЫ

В коробке 30 предохранителей, 12 из них дефектный. Если мы возьмем пять предохранителей наугад, вероятность того, что ни один из них не исправен, равна равно:

$$P(X = 0) = \frac{C_M^x C_{N-M}^{n-x}}{C_N^n} = \frac{C_{12}^0 C_{18}^5}{C_{30}^5}$$
$$= 0{,}0601.$$

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Распределение Бернулли

Биномиальное распределение

Дискретное распределение вероятностей

Гипотеза

Статистическая гипотеза —это утверждение относительно распределения (распределений) одной или нескольких случайных величин. Это может касаться параметров данного распределения или распределения вероятностей изучаемой совокупности .

Справедливость гипотезы проверяется путем проверки гипотезы на наблюдениях, собранных для выборки изучаемой популяции.

При проверке гипотезы о распределении вероятностей изучаемой совокупности гипотеза о том, что изучаемая совокупность следует заданному распределению вероятностей, называется нулевой гипотезой. Гипотеза, утверждающая, что популяция не следует заданному распределению вероятностей, называется альтернативной гипотезой (или противоположной гипотезой).

Если мы проводим проверку гипотез о параметрах распределения , то гипотеза о том, что изучаемый параметр равен заданному значению, называется нулевой гипотезой. Гипотеза, утверждающая, что значение параметра отличается от данного заданного значения, называется альтернативной гипотезой.

Нулевая гипотеза обычно обозначается H0 , а альтернативная гипотеза —H1.

ИСТОРИЯ

При проверке гипотезы гипотеза, подлежащая проверке, называется нулевой гипотезой. Мы обязаны термином «нольь» Фишеру Р.А.

(1935). Вводя эту концепцию, он упомянул известную проблему дегустации чая, когда дама утверждала, что может определить по вкусу, что было налито ей в чашку первым —молоко или чай. Гипотеза, которую нужно было проверить, заключалась в том, что на вкус абсолютно не влиял порядок, в котором был заварен чай.

Первоначально нулевая гипотеза обычно понималась как означающая, что конкретное лечение не оказывает никакого эффекта или что не было никаких различий. связь между эффектами различных видов лечения ментыы.

В настоящее время нулевая гипотеза, в отличие от альтернативной гипотезы, в основном используется для обозначения того, что гипотеза нуждается в проверке . См. также проверку гипотез.

ПРИМЕРЫ

Многие задачи связаны с повторением эксперимента с двумя возможными результатами.

Одним из примеров этого является пол новорожденного ребенка. В данном случае нас интересует соотношение мальчиков и девочек в данной популяции. Рассмотрим р, долю девочек, которую мы хотели бы оценить по наблюдаемой выборке . Чтобы определить, одинаковы ли пропорции новорожденных мальчиков и девочек, мы делаем статистическую гипотезу, что р =

$$\frac{1}{2}.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Альтернативная гипотеза

Дисперсионный анализ

Проверка гипотезы

Нулевая гипотеза

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: План экспериментов.

Оливер и Бойд, Эдинбург (1935)

Machine Translated by Google

Проверка гипотезы

Проверка гипотезы—это процедура, которая позволяет нам (в зависимости от определенных правил принятия решений) подтвердить исходную гипотезу, называемую нулевой гипотезой, или отклонить эту нулевую гипотезу в пользу альтернативной гипотезы.

ИСТОРИЯ

Теория проверки гипотез развивалась в ходе исследования. Первые шаги были сделаны, когда стали появляться работы, в которых обсуждалась значимость (или незначимость) группы наблюдений. Некоторые примеры таких работ относятся к восемнадцатому веку, в том числе работы Арбутнотта, Дж (1710 г.), Бернулли, Даниэля (1734 г.) и Лапласа, Пьера Симона де (1773 г.). Эти работы чаще можно было увидеть в девятнадцатом веке, например, работы Гаваретта (1840 г.) и Эджворта, Фрэнсиса Ю (1885 г.).

Развитие проверки гипотез происходило параллельно с теорией оценивания. Проверка гипотез, по-видимому, впервые была разработана исследователями экспериментальных наук и управленческой сферы. Например, тест Стьюдента был разработан Госсетом Уильямом Сили во время его работы в Guinness.

Нейман, Ерзянд Пирсон, Эгон Шарп разработали математическую теорию проверки гипотез, которую они изложили в статье, опубликованной в 1928 году в журнале «Биометрика». Они первыми осознали, что рациональный выбор, который необходимо сделать во время проверки гипотезы, должен быть между нулевой гипотезой, которую мы хотим проверить, и альтернативной гипотезой. Вторая фундаментальная статья по теории проверки гипотез была опубликована в 1933 г. теми же математиками, где они также различали ошибку I рода и ошибку II рода.

Работы, ставшие результатом сотрудничества Neyman, J. и Pearson, ES, описаны в Pearson (1966) и в биографии Neyman, опубликованной Reid (1982).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Проверка гипотез выборки обычно включает следующие этапы: 1. Сформулируйте гипотезы: • Нулевая гипотеза H0, • Альтернативная гипотеза H1.

2. Определить уровень значимости α теста.

3. Определить распределение вероятностей, соответствующее выборочному распределению

4. Рассчитайте критическое значение нулевой гипотезы и выявести область отклонения или область принятия.

5. Установите правила принятия решений:

- Если статистика, наблюдаемая в выборке, находится в допустимой области, мы не отвергаем нулевую гипотезу H0;
- Если статистика, наблюдаемая на сам ple расположены в области отклонения, мы отвергаем нулевую гипотезу H0 для альтернативной гипотезы H1.

6. Примите решение принять или отклонить нулевую гипотезу на основе наблюдаемой выборки.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Наиболее распространенные типы проверки гипотез описаны ниже.

1. Проверка гипотезы выборки. Мы хотим проверить, совпадает ли значение параметра θ совокупности с предполагаемым значением. Гипотезы будут следующие:

$$H0 : \theta = \theta 0 ,$$

$$H1 : \theta = \theta 0 .$$

Machine Translated by Google

где θ0 – предполагаемое значение
неизвестный параметр θ.

2. Проверка гипотезы на двух выборках :
цель в данном случае – выяснить,
две популяции, обе описанные
тот или иной параметр отличается. Пусть θ1
и θ2 — параметры, описывающие популяции 1 и 2
соответственно. Мы можем тогда
сформулируйте следующие гипотезы:

$$H0 : \theta1 = \theta2 ,$$

$$H1 : \theta1 = \theta2 .$$

или

$$H0 : \theta1 - \theta2 \text{ знак равно } ,$$

$$0\ H1 : \theta1 - \theta2 \text{ знак равно } 0 .$$

3. Проверка гипотез более двух
образцы: Что касается теста, проведенного на двух
образцы, проводится проверка гипотез
более чем на двух образцах для определения
отличаются ли эти популяции,
на основе сравнения одного и того же параметра
из всех тестируемых популяций.
В этом случае проверьте следующую гипотезу
сэс:

$$H0 : \theta1 = \theta2 = \cdots = \theta k ,$$

H1 : значения θi (i = 1, 2,...,k)
не все идентичны.

Здесь θ1,..., θk — неизвестные параметры
популяций, k — количество сравниваемых
популяций.

В теории проверки гипотез и на практике мы можем
различать два типа
тестов: параметрический тест и непараметрический
тест.

Параметрические тесты

Параметрический тест — это тест гипотезы,
предполагающий определенную форму для каждого из
распределения, связанные с лежащими в их основе
популяциями. Этот случай применим, например, когда
эти популяции имеют нормальное распределение.
Яичко студента является примером параметрического
теста. В этом тесте сравниваются средние значения двух
нормально распределенные популяции.

Непараметрический тест

Непараметрический тест — это проверка гипотезы
где нет необходимости указывать
параметрическая форма распределения
основное население.
Примеров такого типа много
тест, в том числе знаковый тест, тест Уилкоксона
тест, подписанный тест Уилкоксона, тест Манна-
тест Уитни, тест Краскела-Уоллиса и
тест Колмогорова-Смирнова.

ПРИМЕРЫ

Примеры параметрической проверки гипотез см. в
биномиальном тесте, тесте Фишера или тесте
Стьюдента. Примеры непараметрической проверки
гипотез см. у Колмогорова.
тест Смирнова, тест Крускала–Уоллиса,
критерий Манна-Уитни, критерий Вилкоксона,
подписанный тест Уилкоксона и знаковый тест.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Область приема
- Альтернативная гипотеза
- Непараметрический тест
- Нулевая гипотеза
- Односторонний тест
- Параметрический тест
- Область отторжения
- Выборочное распределение
- Уровень значимости
- Двусторонний тест

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Арбетнотт, Дж: Аргумент в пользу божественного
провидения, взятый из постоянной регулярности
наблюдается при рождении обоих полов. Фи
лос. Транс. 27, 186–190 (1710)

Machine Translated by Google

Бернулли, Д.: Что является причиной физического наклона планет (...). Рек. Отчет о предметах. Приз акад. Рой. наук. 3, 95–122 (1734)

Эджворт, FY: Методы статистики. Юбилейный том Королевского статистического общества, Лондон (1885 г.)

Гаварре, Ж.: Principes généraux de statistiquemédicale. Beché Jeune & Labé, Париж(1840 г.)

Лаплас, PS de: Mémoire sur l'inclinaison moyenne des orbites des comètes. Мем. акад. Рой. наук. Париж, 7, 503–524 (1773 г.)

Леманн, Э.Л.: Проверка статистических гипотез, 2-е изд. Уайли, НьюЙорк (1986)

Нейман, Дж, Пирсон, Э.С.: Об использовании и интерпретации некоторых критериев тестирования для целей статистического вывода, Части I и II. Биометрика 20A, 175–240, 263–294 (1928).

Нейман, Дж, Пирсон, Э.С.: К проблеме наиболее эффективных тестов статистики. кал гипотезы. Филос. Транс. Рой. соц. Лонд. сер. А 231, 289–337 (1933)

Пирсон, Э.С.: История Неймана-Пирсона: 1926–1934. В: Дэвид, Ф.Н. (ред.) Исследовательские статьи по статистике: Festschrift для Дж Неймана. Уайли, НьюЙорк (1966)

Рид, К.: Нейман —из жизни. Спрингер, Берлин Гейдельберг, НьюЙорк (1982)

Machine Translated by Google

Идемпотентная матрица

См. матрицу .

Единичная матрица

См. матрицу .

Заболеваемость

См. уровень заболеваемости.

Уровень заболеваемости

Частота заболевания определяется как число новых случаев заболевания появляющиеся в течение определенного периода среди особи популяции . Это понятие аналогично на «потоку».

Уровень заболеваемости определяется как относительный размер населения и время; это значение выражается отношением к числу лиц к продолжительности.

Уровень заболеваемости - это уровень заболеваемости , разделенный на по количеству людей, подверженных риску заболевания.

ИСТОРИЯ

См. показатель распространенности.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Формально имеем:

$$\text{уровень заболеваемости} = \frac{\text{риск}}{\text{продолжительность наблюдения}} .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Точный показатель заболеваемости, рассчитанный в населения и в течение определенного периода будет интерпретировать как показатель существования эпидемии в этой популяции.

ПРИМЕРЫ

Рассчитаем заболеваемость молочной железы рак у 150 000 женщин наблюдается в течение 2 лет, согласно следующей таблице:

	Число из случаи	Число из предметы	Заболеваемость ставка / 100000 / год
Группа	[А]	[Б]	$[\frac{А}{2Б}]$
Не защищенный	40	35100	57,0
Пассивный курильщики	140	55500	126,1
Активный курильщики	164	59400	138,0
Общий	344	150000	114,7

Числитель заболеваемости равен то же, что и у риска: количество новых

Machine Translated by Google

случаев (или новых случаев) рака молочной железы за 2 года наблюдения среди группы предпочтения. Знаменатель заболеваемости получается путем умножения числа лиц в группе предпочтения на продолжительность наблюдения (здесь 2 года). Обычно ставка выражается как 1 год.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Атрибутивный риск
- Предотвратимый риск
- Причина и следствие в эпидемиологии
- Шансы и отношение шансов
- Уровень распространенности
- Относительный риск
- Риск

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Лилиенфельд, А.М., Лилиенфельд, Д.Э.: Основы эпидемиологии, 2-е изд. Кларендон, Оксфорд (1980)

МакМахон, Б., Пью, Т. Ф.: Эпидемиология: принципы и методы. Литтл Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Морабия, А.: L'Epidémiologie Clinique. Издания «Que sais-je?». Presses Universitaires de France, Париж (1996)

Неполные данные

См. отсутствующие данные.

Независимость

Понятие независимости может иметь два значения: первое касается событий случайного эксперимента. говорят о двух событиях быть независимым, если реализация второго события не зависит от реализации первого.

Независимость между двумя событиями A и B можно определить с помощью вероятностей: если вероятность события A при реализации B одинакова с вероятностью A при нереализации B, то A и B называются независимыми.

Второе значение слова независимость касается отношения, которое может существовать между случайными величинами. Независимость двух случайных величин можно проверить ed с тестом Кендалла и тестом Спирмена тест.

ИСТОРИЯ

Понятие независимости имплицитно использовалось задолго до того, как был установлен формальный набор аксиом вероятности. Согласно Майстрову, Л.Е. (1974), Кардано уже использовал правило умножения вероятностей (P (A B) = P (A) · P (B)). Майстров Л.Е. также упоминает, что понятия независимости и зависимости между событиями были хорошо знакомы Паскалю, Блезу, де Ферма, Пьеру и Гюйгенсу, Христиану.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Математически мы можем сказать, что событие A не зависит от события B, если

$$P(A) = P(A \mid B),$$

где P(A) — вероятность реализации события A, а P(A|B) — условная вероятность события A как функция реализации B, что означает вероятность того, что A произойдет, зная, что B произошло.

Используя определение условной вероятности, предыдущее уравнение можно записать как:

$$P(A) = \frac{P(A \quad B)}{P(B)}$$

Machine Translated by Google

или

$$P(A) \cdot P(B) = P(A \cap B),$$

который дает

$$P(B) = P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

С помощью этого простого преобразования мы показали, что если A не зависит от B, то и B не зависит от A. Будем говорить, что события A и B независимы.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

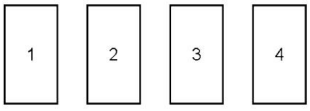
Если два события A и B независимы, то они всегда совместимы. Вероятность того, что события A и B произойдут одновременно, может быть определена путем умножения индивидуальных вероятностей двух событий:

$$P(A \cap B) = P(A) \cdot P(B).$$

Но если A и B зависимы, они могут быть совместимы или несовместимы. Абсолютного правила не существует.

ПРИМЕРЫ

Рассмотрим коробку с четырьмя купюрами, пронумерованными от 1 до 4, из которых мы будем делать два последовательных рисунка, не помещая первую заметку обратно в коробку, прежде чем рисовать вторую заметку.



Демонстрационное пространство может быть представлено следующими 12 парами:

$$= \{(1, 2), (1, 3), (1, 4), (2, 1), (2, 3), (2, 4), (3, 1), (3, 2), (3, 4), (4, 1), (4, 2), (4, 3)\}.$$

Рассмотрим следующие два события:

A = «вытащите заметку № 3 на первом чертеже»

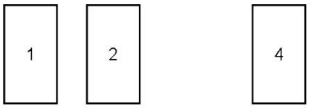
$$= \{(3, 1), (3, 2), (3, 4)\}$$

B = «вытащите заметку № 3 на втором чертеже»

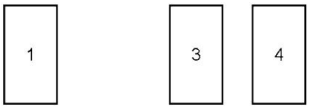
$$= \{(1, 3), (2, 3), (4, 3)\}$$

Являются ли эти два события зависимыми или независимыми? Если они независимы, то вероятность события B должна быть идентична вероятности события A независимо от результата события A.

Если произойдет A, то в поле для второго рисунка останутся следующие примечания:



Тогда вероятность того, что произойдет B, равна нулю, потому что купюры № 3 больше нет в коробке. С другой стороны, если A не выполнено (например, на первом чертеже была нарисована примечание № 2), то на втором чертеже в рамке будут следующие примечания:



Вероятность того, что B будет выполнено, равна $\frac{1}{3}$. Таким образом, мы можем заключить, что вероятность события B зависит от события A или что события A и B зависимы.

Мы можем проверить наш вывод следующим образом: если A и B независимы, то должно быть проверено следующее уравнение:

$$P(A \cap B) = P(A) \cdot P(B).$$

В нашем примере $P(A) = \frac{3}{12}$ и $P(B) = \frac{3}{12}$.

Machine Translated by Google

Пересечение между A и B равно нулю.
Поэтому у нас есть

$$P(A \cap B) = 0 = P(A) \cdot \frac{1}{16} = P(B).$$

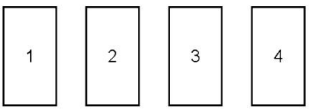
Если равенство не подтверждается, то события A и B не являются независимыми. Поэтому они зависимы.

Рассмотрим теперь аналогичный эксперимент, но положить нарисованную записку обратно в коробку, прежде чем второй рисунок.

Выборочное пространство этого нового случайного эксперимент состоит из следующих 16 пары:

$$= \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4), (4, 1), (4, 2), (4, 3), (4, 4)\}.$$

В этих условиях второй рисунок будет производиться из коробки, идентичной той, что первый рисунок:



События A и B содержат следующие результаты:

$$A = \{(3, 1), (3, 2), (3, 3), (3, 4)\},$$

$$B = \{(1, 3), (2, 3), (3, 3), (4, 3)\}.$$

Вероятность A, как и B, равна 4
fore $\frac{4}{16}$.
Пересечение между A и B
задано событием $A \cap B = \{(3, 3)\}$; его вероятность 1
способность $\frac{1}{16}$, поэтому мы имеем

$$P(A \cap B) = \frac{1}{16} = \Pi(A) \cdot \Pi(B).$$

Поскольку $P(A \cap B) = P(A) \cdot P(B)$, мы можем заключить, что событие B не зависит от события A, т.е. или что события A и B независимы.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Критерий независимости хи-квадрат
- Совместимость
- Условная возможность
- Зависимость
- Событие
- Коэффициент ранговой корреляции Кендалла
- Вероятность
- Коэффициент ранговой корреляции Спирмена
- Тест на независимость

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Майстров, Л.Е.: Теория вероятностей - очерк истории (перевод и редакция Коца С.). Академик, НьюЙорк (1974)

Независимая переменная

В регрессионной модели переменные, которые считается влияющим на зависимые переменные или объясняющим вариации последних, все независимые переменные.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Зависимая переменная
- Множественная линейная регрессия
- Регрессивный анализ
- Простая линейная регрессия

Индекс

В самом общем определении порядковый номер — это значение, представляющее относительное изменение переменной между двумя определенными значениями. периоды (или ситуации).

Простые порядковые номера следует отличать от составных порядковых номеров.

Machine Translated by Google

- Простые порядковые номера описывают относительное изменение одной переменной.
- Составные порядковые номера позволяют описать одним числом сравнение набора значений, которые принимают несколько переменных в определенной ситуации с относительно набора значений одних и тех же переменных в эталонной ситуации.

Эталонная ситуация определяет основу номер индекса. Мы можем сказать, например, что для отчетного (базисного) 1980 года некоторый индекс имеет значение 120 в 1982.

ИСТОРИЯ

Первые исследования с использованием индексов относятся к началу 18 века.

В 1707 г. англиканский епископ Флитвуд предпринял исследование эволюции цен между 1440 и 1700 гг. В 1460 г. был установлен существенный пункт, согласно которому любой допущенный член должен был покинуть колледж, если его состояние превышало 5 фунтов стерлингов в год. Флитвуд хотел знать, можно ли, принимая во внимание эволюцию цен, сдержать такое обещание три столетия спустя.

Он рассмотрел четыре продукта — пшеницу, мясо, напитки и одежду — и изучил эволюцию цен на них. В конце своей работы он пришел к выводу, что 5 фунтов (£) в 1440–1460 годах имели ту же стоимость, что и 30 фунтов (£) в 1700 году.

Позже, в 1738 г., Дюто К. исследовал уменьшение стоимости денег в работе, озаглавленной «Политические размышления о финансах и торговле». Для этого он изучил доходы двух королей, Людовика XII и Людовика XV. Два короля имели следующие доходы:

Людовик XII: 7650000 фунтов стерлингов в 1515 году.

Людовик XV: 100000000 фунтов стерлингов в 1735 году.

Чтобы определить, кто из суверенов имел наибольший располагаемый доход, Дюто, К. отмечал различные цены, рассматриваемые даты на определенное количество товаров: курицу, кролика, голубя, стоимость дневного труда, и т. д.

Измерение Дюто было следующим:

$$я = \frac{P1}{P0},$$

где P1 и P0 – суммы отмеченных цен на соответствующие даты. Обратите внимание, что этот порядковый номер не обдумывается. Дюто пришел к выводу, что богатство Людовик XII был ниже, чем Людовик XV.

Уменьшение стоимости денег также было предметом исследования Карли, Джан Ринальдо. В 1764 году профессор астрономии из Падуэ изучал эволюцию цен с момента открытия Америки.

Он взял в качестве эталонов три продукта — зерно, вино и растительное масло — изучил цены на эти продукты в период 1500–1750 гг.

Измерение Карли было следующим:

$$я = \frac{1}{н} \cdot \frac{P1}{P0}.$$

Полученное числовое значение является арифметическим средним значением относительных цен на каждый наблюдаемый продукт.

В 1798 году была опубликована работа под названием «Отчет о некоторых усилиях по установлению стандарта веса и меры», подписанная сэром Джорджем Шакбургом Эвелином. Этот предшественник индексов определяет индекс цен, охватывающий период с 1050 по 1800 год, путем введения понятий базисного года (1550) и относительных цен. Он также

Machine Translated by Google

включил стоимость некоторых услуг в свой наблюдения.

Это Лоу, Джозеф, которого следует увидеть, согласно Kendall, MG (1977), как истинный отец индексных чисел. Его работа, опубликованная в 1822 г., называлась « Современное состояние Англия лечила многие проблемы относительно создание индексных номеров. Среди прочего, он ввел понятие цены. обдумывание вразные даты (P1 и P0) по количества, наблюдаемые в первый день (Q0). Измерение Лоу следующее:

$$я = \frac{(П1 \cdot K0)}{(P0 \cdot Q0)} .$$

Позже этот порядковый номер станет известен как индекс Ласпейреса.

Во второй половине XIX века статистики добились значительных успехов в этой области. В 1863 году с «Серьезным падением стоимости обнаруженного золота и его социальные последствия Сет Форт, Джевонс, В.С. изучали тип предназначен для использования при создании индекс. Он рекомендовал использовать среднее геометрическое. Между 1864 и 1880 годами три немецких ученых, Ласпейрес, Э. (1864, 1871), Дробиш, М.В. (1871) и Пааше, Х. (1874), работал над эволюцией цен на материальные блага в Гамбурге, не принимая во внимание услуги. Все три являются не обдуманными методами расчета.

Индекс Пааше определяется по следующей формуле:

$$я = \frac{(П1 \cdot K1)}{(P0 \cdot Q1)} .$$

Palgrave, RHI (1886) предложил рассматривать относительные цены по общей стоимости относился хорошо; этот метод приводит к исчислению номер индекса следующим образом:

$$я = \frac{П1 \cdot В1 \cdot \frac{P1}{P0}}{(П1 \cdot K1)} .$$

Это были работы таких статистиков, как И. Фишер, определивший в 1922 г. числовой индекс бер, носящий его имя и составляющий среднее геометрическое индекса Ласпейреса и индекс Пааше:

$$я = \sqrt{\frac{(П1 \cdot Q0)\,(П1 \cdot Q1)}{(P0 \cdot Q0)\,(P0 \cdot Q1)}} ,$$

тогда как в тот же период Маршалл, А. и Эджворт Ф.У. предложил следующую формулу:

$$я = \frac{(P1\,(Q1 + Q0))}{(P0\,(Q1 + Q0))} .$$

Обратите внимание, что в этой формуле это среднее вычисляемых весов.

Современные статистические исследования по-прежнему в значительной степени называют по порядковым номерам, которые были определено здесь.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Расчет порядковых номеров является обычным альный метод описания изменений экономических переменных. Даже если самая актуальная порядковые номера в основном касаются цены переменная и переменная количества, они могут также рассчитываться по ценностям, качествам и т. д. Индексы, как правило, ассоциируются со сферой бизнеса и экономики. широко используется в других сферах деятельности.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Составной порядковый номер
- Индекс Фишера
- Индекс Ласпейреса
- Индекс Пааше
- Простой порядковый номер

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Карли, GR: Del valore и т. Д. В: Opere Scelte di Carli, vol. я, с. 299. Кустоди, Милан (1764 г.)

Дробиш, MW: ÜberMittelgrössen und die Anwendbarkeit derselben auf die Berech

nung des Steigens und Sinkens des Geld werts. Berichte über die Verhandlungen der König. Сакс. Гэс. Висс. Лейпциг.
математика-физ. Классе, 23, 25 (1871 г.)

Джого, К.: Политические размышления о финансах и торговле. Вайян и Николя Прево, Гаага (1738 г.)

Эвелин, сэр GS: отчет о некоторых попытках установить стандарт веса и меры. Фил ос. Транс. 113 (1798)

Флитвуд, В.: Chronicon Preciosum, 2-е изд. Флитвуд, Лондон (1707 г.)

Jevons, WS: Установлено серьезное падение стоимости золота и его социальные последствия Прояснять. Стэнфорд, Лондон (1863 г.)

Кендалл, MG: Ранняя история индексов. В: Кендалл, М., Плакетт, Р.Л. (ред.) Исследования по истории статистики и вероятностей, том. II. Гриффин, Лондон (1977)

Ласпейрес, Э.: Гамбургер Варенпрайз 1850–1863 гг. и Калифорнийская Австралия.

lischen Geldentdeckung seit 1848. Jahrb.
Натл. Стат. 3, 81–118, 209–236 (1864 г.)

Ласпейрес, Э.: Die Berechnung einer mit tleren Waarenpreissteigerung. Ярб.
Натл. Стат. 16, 296–314 (1871 г.)

Лоу, Дж: Современное состояние Англии в отношении сельского хозяйства, торговли и финансов. Келли, Лондон (1822 г.)

Пааше, Х.: Über die Preisentwicklung der letzten Jahre nach den Hamburger Borsen-notirungen. Ярб. Натл. Стат. 23, 168–178 (1874 г.)

Palgrave, RHI: Валюта и стандарт стоимости в Англии, Франции, Индии и т. д. (1886 г.) В: IUP (1969) British Parliamentary Papers: Session 1886: Third Report and Final Report of the Royal Commission on the Depression in Trade and Industry., с протоколом свидетельских показаний и приложениями, Приложение B. Irish University Press, Shannon

Индикатор

Индикатор – это статистический показатель (официальная статистика), целью которого состоит в том, чтобы дать представление о состоянии, поведении и характере изменения в течение определенного периода экономического или политического явления. Индикатор должен информировать об изменениях значений, а также изменения в характере наблюдаемых явлений и должны служить руководством для принятия решений.

производители сион.

Наиболее точно показатель отражает социальный или экономический процесс и предлагает изменения, которые следует произвести. Это тревожный сигнал, привлекающий внимание и служащий призывом к действию (пилотаж). Многие индикаторы, как правило, необходимы для обеспечения глобального отражения данных явлений и той политики, которую нужно оценить. Набор индикаторов предоставляет информацию об одном и том же предмете и называется системой индикаторов.

катеры.

Не вся статистика является индикатором; индикаторы должен:

1. Быть измеримым, что, строго говоря, не означает измеримым.
2. Будьте одномерными (избегайте дублирования с другими показателями).
3. Охватить набор приоритетных задач.
4. Непосредственно оценить эффективность политики.

Machine Translated by Google

5. Покажите пути улучшения.
6. Обратитесь к основным аспектам системы и
быть изменчивой при изменении политической
системы.
7. Будьте поддающимися проверке.
8. Поддерживайте отношения друг с другом.
9. Будьте максимально экономны.
10. Будьте всегда доступны.
11. Быть легко понятым в общем публичном
лицензия.
12. Быть признанным всеми как действительный
и надежный.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Построение индикаторов

Показатели, как и все измерения социальных
явлений, не могут быть точными и исчерпывающими.
построение индикаторов требует философских
соображений (суждение о ценностях)
и технические и научные знания, но это
также основан на логике и часто отражает
политические интересы.

Сбор и выбор индикаторов

Для сбора хороших индикаторов
модель экономического, социального и образовательного
системы нужны. Модель в целом
адаптированная OCDE представляет собой систему юмодель:
в данной среде мы изучаем ресурсы, процессы и
эффекты системы. Но
выбор показателей, объединенных в систему
показателей также зависит от логистической
возможность сбора данных. Это объясняет
большая разница между показателями.
количество показателей, необходимых для хорошего
пилотирование экономической или социальной системы
трудно оценить. Если их слишком много,
индикаторы теряют значение и рискуют не играть
важной роли. Но вместе они
должны представлять оцениваемую систему.
Слишком малое число рискует сделать систему
недействительным и нерепрезентативным.

Интерпретация индикаторов

Интерпретация индикаторов так же важна для системы,
как их выбор, построение и сбор. Имея значение

что индикаторы могут принять за политически
ответственность за финансирование и успех
некоторых проектов должны соблюдаться все меры предосторожности.
принятые для обеспечения интерпретативной целостности,
качество и нейтральность индикатора.

Валидация индикаторов

Прежде чем использовать эту статистику в качестве индикаторов
производительности, качества и здоровья
экономическая, социальная или образовательная система,
они должны быть утверждены. Если значения
индикаторов рассчитываются научно с использованием
математических методов и инструментов, то их
влияние на описываемую систему и их
надежность как основа плана
действия не всегда научно доказаны.
Кроме того, когда используется набор индикаторов,
до сих пор далеко не ясно, как эти индикаторы объединяют свои эффекты, чтобы произвести
наблюдаемые результаты. Методы проверки
Показатели еще не разработаны.

ПРИМЕРЫ

Экономические показатели

К ним относятся валовой внутренний продукт
(ВВП), национальный доход, потребительские цены
индекс, индекс цен производителей, безработица
процентная ставка, общая занятость, долгосрочные и
краткосрочные процентные ставки, экспорт и импорт
произведенных товаров и платежного баланса.

Показатели образования

К ним относятся расходы на обучение,
активность в системе преподавания, продолжительность
обучения, географические различия в доступе к
бакалавриату,

Machine Translated by Google

общий уровень призывников, уровень образования молодежи после окончания школы выпускной, преимущества бакалавриата степень в отношении трудоустройства, обязательное обучение детей от 2 до 5 лет и расходы на обучение непрерывное образование.

Показатели науки и техники

К ним относятся валовые внутренние расходы по исследованиям и разработкам (GDERD), ВРВР на душу населения, ВРВР в процентах от ВВП, общие расходы на НИОКР на персонал, общее количество университетских дипломов, процентное финансирование ВВРНИ предприятиями, процентное финансирование ВВРНИ за счет государства, процентное за границей, а также национальные патентные заявки.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Официальная статистика

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

OCDE: OCDE и индикаторы между нации де l'enseignement. Центр de la Recherche et l'innovation dans l'enseignement, OCDE, Париж(1992)

OCDE: Принципы указателей науки и технологии. OCDE, Париж(1988)

Инерция

Инерция — это понятие, которое широко используется в механика. Чем больше инерция тела вокруг оси, тем больше энергии должно быть используется, чтобы заставить это тело вращаться вокруг этого ось. Масса тела и (в квадрате) расстояния от тела до оси используются при расчете инерции. определение дисперсии можно признать здесь, если «масса» заменяется «частотой», а «расстояние» — «отклонением».

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Предположим, что у нас есть облако из n точек с центром в пространстве размерности p; инерция _ матрица V может быть построена с координатами динаты и частоты точек.

Собственные значения k1,...,kp (записанные в порядке убывания) матрицы V — это инерция, объясняемая осями факториала облако точек. Их сумма, равная следу V, равно полной инерции облака.

Скажем, например, что первая факториальная ось (или ось инерции) объясняет:

$$100 \cdot \frac{k1}{n} \text{ (в \%) инерции}$$

ки

я=1

Первые две оси инерции, например, объяснять:

$$\frac{k1 + k2}{n} \text{ (в \%) инерции } 100 \cdot$$

ки

я=1

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Понятие инерции или, точнее, используется матрица инерции облака точек, в том числе для проведения анализа соответствия между рассматриваемыми точками.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Анализ корреспонденции

Расстояние

собственное значение

Факторная ось

Частота

Матрица инерции

Machine Translated by Google

Матрица инерции

Рассмотрим облако точек в пространстве размеры. Мы предполагаем, что это облако с центром в начале системы отсчета; если нет, то достаточно ввыввать изменение переменных путем вычитания из каждой точки центр облака. Обозначим через A матрица координат, умноженная на квадратный корень из частоты, связанной с каждая точка. Такая матрица для n различных точек будет иметь размерность (n×p). Предполагать что р меньше или равно n; в противном случае мыиспользуем транспонирование матрицыA в качестве нового матрица координат.

Мы получаем матрицу инерции V , умножая A на транспонирование A, A . V — симметричная матрица порядка p. След этой матрицыV в точности равен сумме инерция (или полная дисперсия).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть n —различные точки Xi p компонент

$$Xi = xi1, xi2,\ldots,xip, \qquad я = 1, 2,\ldots,n,$$

каждая точка Xi имеет одну частоту fi (i = 1, 2,...,n).

Если точки не центрированывокруг начала координат, то мыдолжнысделать так, чтобыизменяющиеся переменные соответствовали смещениюк началу координат. центр облака:

$$Икс_{я} = Kси - Y \quad ,$$

где Y = y1, y2,...,yp с

$$y j = \frac{1}{\text{фи}_{я=1}^{н}} \cdot \sum_{я=1}^{н} фи \cdot xij \quad .$$

Общий член матрицыинерции V тогда:

$$Видж = \sum_{к=1}^{н} fk \cdot (xki - yi) \cdot xkj - yj \, .$$

В матричной форме V = A · A, где коэффициенты матрицыA равны:

$$aki = fk \cdot (xki - yi) \, .$$

Суммарная инерция облака точек равна получается путем вычисления следа матрицыинерции V, равной:

$$Itot = tr(V) = \sum_{я=1}^{п} VII$$

$$"=" \sum_{я=1}^{п} \sum_{к=1}^{н} fk \cdot (xki - yi)^2 \, .$$

ОБЛАСТИ ИОГРАНИЧЕНИЯ

Матрица инерции облака точек используется для проведения анализа соответствия между двумя рассматриваемыми точками.

ПРИМЕРЫ

Рассмотрим следующие точки в пространстве три измерения:

(2; 0; 0,5)
(3; 1; 2)
(2; 1; 1)
(-1;-1;-0,5)
(-1;-1;-0,5)
(-1;-1;-0,5)

Есть четыре разные частоты равно 1, 1, 1 и 3.
Центр облака, образованный его точками дан кем-то:

$$y 1 = \frac{2 + 3 - 2 + 3 \cdot (-1)}{6} = 0 \quad ,$$

$$y 2 = \frac{1 - 1 + 3 \cdot (-1)}{6} = -0,5 \quad ,$$

$$y 3 = \frac{0,5 + 2 - 1 + 3 \cdot (-0,5)}{6} = 0 \, .$$

Machine Translated by Google

Новые координаты наших четырех разных баллы дадут:

(2; 0,5; 0,5) (3; 1,5; 2) (-2; -0,5; -1) (-1; -0,5; -0,5) частоты 3 ,

из которого вычисляется A — матрица координат, умноженная на корень квадратный из частоты:

$$A = \begin{pmatrix} 2 & 0{,}5 & 0{,}5 \\ 3 & 1{,}5 & 2 \\ 2 & 0{,}5 & 1 \\ -\sqrt{3} & -\frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{pmatrix}$$

и V, матрица инерции:

$$B = A \cdot A = \begin{pmatrix} 20 & 8 & 10{,}5 \\ 8 & 3{,}5 & 4{,}5 \\ 10{,}5 & 4{,}5 & 6 \end{pmatrix}$$

Инерция облака точек равна :

$$Itot = tr(V) = 20 + 3{,}5 + 6 = 29{,}5 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Анализ корреспонденции
Частота
Матрица
След
Транспонировать
Дисперсия

ССЫЛКИ

Lagarde, J. de: Initiation à l'analyse des don nées. Дюно, Париж (1983)

Вывод

Вывод — это форма рассуждения по индукции, основанная на информации, собранной на выборка с точки зрения обобщения информации на совокупность, связанную с этой выборкой.

ИСТОРИЯ

См. выводную статистику .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Выведенный статистика

Выведенный статистика

Выводная или дедуктивная статистика дополняет описательную статистику , поскольку целью большинства исследований является не установление определенного количества показателей для данной выборки, а скорее оценка определенного количества параметров, характеризующих совокупность, связанную с рассматриваемой в Логической статистике гипотезы формулируются в параметрах совокупности . Эти гипотезы затем проверяются на основе наблюдений за репрезентативной выборкой населения.

ИСТОРИЯ

Истоки логической статистики совпадают с истоками теории вероятностей и соответствуют работам Байеса, Томаса (1763 г.), де Муавра, Абрахама (1718 г.), Гаусса, Карла Фридриха (1809 г.) и де Лапласа, Пьера Симона (18 В то время существовало множество исследователей в области логического вывода. тики. Мы упомянем, например, работы Гальтона, Фрэнсиса (1889 г.) в отношении корреляции , а также развитие проверки гипотез , главным образом благодаря Пирсону , Карлу (1900 г.) и Госсету , Уильяму

Machine Translated by Google

Силли, известного как «Студент» (1908). Нейман, Ежи, Пирсон, Эгон Шарп (1928) и Фишер, Рональд Эйлмер (1956) также внесли фундаментальный вклад в развитие статистики логического вывода.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- теорема Байеса
- Описательная статистика
- Оценка
- Проверка гипотезы
- Максимальная вероятность
- Оптимизация
- Вероятность
- Выборка
- Статистика
- Опрос

ССЫЛКИ

Байес, Т.: Эссе по решению проблемы учении о шансах. Филос. Транс. Рой. соц. Лонд. 53, 370–418 (1763). Опубликовано по инициативе Прайса Р. через 2 года после его смерти. Переиздано с биографией Барнарда, Джорджа А., в 1958 г. и в Пирсоне, Э.С., Кендалле, М.Г.: Исследования по истории статистики и вероятностей. Гриффин, Лондон, стр. 131–153 (1970).

Фишер, Р.А.: Статистические методы и наука точный вывод. Оливер и Бойд, Эдинбург (1956)

Гальтон, Ф: Естественное наследование. Макмиллан, Лондон (1889 г.)

Гаусс, CF: Theoria Motus Corporum Целестий. Верке, 7 (1809 г.)

Госсет, С. В. «Студент»: вероятная ошибка среднего. Биометрика 6, 1–25 (1908).

Лаплас, PS de: Аналитическая теория вероятностей. Курсье, Париж (1812 г.)

Муавр, А. де: Доктрина шансов: или Метод расчета вероятности событий в игре. Пирсон, Лондон (1718 г.)

Нейман, Дж, Пирсон, Э.С.: Об использовании и интерпретации некоторых критериев тестирования для целей статистического вывода, Части I и II. Биометрика 20A, 175–240, 263–294 (1928).

Пирсон, К.: О критерии, что данная система отклонений от вероятного в случае коррелированной системы переменных такова, что можно разумно предположить, что она возникла в результате случайной выборки. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, стр. 339–357. Впервые опубликовано в 1900 году в Philos.Mag. (5-я серия) 50, 157-

Внутренний продукт

Скалярное произведение x и y, обозначаемое $x \cdot y$, представляет собой сумму произведений, компонент за компонентом, двух векторов.

Геометрически скалярное произведение двух векторов размерности n соответствует в пространстве n измерений числу, полученному умножением нормы одного на норму проекции другого. Это определение эквивалентно утверждению, что скалярное произведение двух векторов получается произведение их норм и косинусу угла, определяемого двумя векторами.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x и y будут двумя векторами размерности n, заданными их компонентами:

$$x = \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix} \quad \text{и} \quad y = \begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{matrix} .$$

Machine Translated by Google

Скалярное произведение x и y, обозначаемое x · y, определяется формулой:

$$x \cdot y = [x1\ x2\ \ldots xn] \cdot \begin{matrix} y1 \\ y2 \\ \vdots \\ yn \end{matrix},$$

$$x \cdot y = x1 \cdot y1 + x2 \cdot y2 + \ldots + xn \cdot yn,$$

$$x \cdot y = \sum_{i=1}^{n} xi \cdot yi.$$

Скалярное произведение вектора x на самого себя соответствует его квадратной норме (или длине), обозначаемой x^2:

$$x^2 = x \cdot x = \sum_{i=1}^{n} xi^2.$$

С помощью нормы мы можем вычислить скалярное произведение

$$x \cdot y = x \cdot y \cdot \cos\alpha,$$

где α — угол между векторами x и y.

Скалярное произведение двух векторов обладает следующими свойствами:

- Оно коммутативно, т. е. для двух векторов одинаковой размерности x и y $x \cdot y = y \cdot x$. Другими словами, нет никакой разницы, проецируем ли мы x на y или наоборот.
- Мы можем взять отрицательные значения, если, например, проекции двух векторов, на которые мы сделали проекцию, имеют противоположные направления.

ПРИМЕРЫ

Рассмотрим два вектора в трехмерном пространстве, заданные своими компонентами:

$$x = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \quad \text{и} \quad y = \begin{matrix} 0 \\ 2 \\ 1 \end{matrix}.$$

Скалярное произведение x и y, обозначаемое x · y, определяется формулой:

$$x \cdot y = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \cdot \begin{matrix} 0 \\ 2 \\ 1 \end{matrix}$$

$$= [1\ 2\ 3] \cdot \begin{matrix} 0 \\ 2 \\ 1 \end{matrix},$$

$$x \cdot y = 1 \cdot 0 + 2 \cdot 2 + 3 \cdot (-1) = 4 - 3 = 1,$$

$$x \cdot y = 1.$$

Норма x, обозначаемая x, задается квадратным корнем скалярного произведения самого по себе. Таким образом:

$$x \cdot x = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \cdot \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$$

$$= [1\ 2\ 3] \cdot \begin{matrix} 1 \\ 2 \\ 3 \end{matrix},$$

$$x \cdot x = 12 + 22 + 32 = 1 + 4 + 9,$$

$$x \cdot x = 14,$$

откуда $x = \sqrt{14} \approx 3{,}74$.

Вектор x имеет длину 3,74 в трехмерном пространстве.

Точно так же мы можем вычислить норму вектора y. Его дают:

$$y^2 = y \cdot y = \sum_{i=1}^{3} yi^2 = 2^2 + (-1)^2 = 5$$

и равен $y = \sqrt{5}$.

Таким образом, мы можем определить угол α, расположенный между двумя векторами, с помощью соотношения:

$$\cos\alpha = \frac{x \cdot y}{x \cdot y} = \frac{1}{\sqrt{14} \cdot \sqrt{5}} \approx 0{,}1195.$$

Это дает приблизительный угол 83,1 между векторами x и y.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Матрица

Транспонировать

Вектор

Взаимодействие

Термин «взаимодействие» используется при построении методов эксперимента, точнее в факторных экспериментах, когда одновременно может быть изучено определенное количество факторов. Он возникает, когда влияние на зависимую переменную изменения уровня одного фактора зависит от уровней других факторов.

ИСТОРИЯ

Концепция взаимодействия при разработке методологии эксперимента принадлежит Фишеру Р.А. (1925, 1935).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Модель двустороннего классификационного дизайна с взаимодействием выглядит следующим образом:

$$Yijk = \mu + \alpha i + \beta j + (\alpha\beta)ij + \varepsilon ijk\ ,\ i = 1, 2,...,a,$$

$$j = 1, 2,...,b \text{ и } k = 1, 2,...,nij\ ,$$

где Yijk —k-е наблюдение, получающееся лечение ij, μ —общее среднее значение, общее для всех видов лечения, αi —эффект i-го уровня фактора A, βj —эффект j-го уровня фактора B, (αβ)ij —эффект взаимодействия между i-м уровнем фактора A и j-й уровень фактора B, а ijk – экспериментальная ошибка наблюдения Yijk.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Взаимодействие между двумя факторами называемые взаимодействиями первого порядка, между три фактора называются взаимодействием второго порядка ции и так далее.

ПРИМЕРЫ См.

двусторонний дисперсионный анализ.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Дисперсионный анализ

План экспериментов

Эксперимент

Распределение Фишера

Таблица Фишера

Тест Фишера

Двусторонний дисперсионный анализ

ССЫЛКИ

Кокс, Д.Р.: Взаимодействие. Между нар. Стат. Откр. 52, 1–25 (1984)

Фишер, Р.А.: Статистические методы для научных работников. Оливер и Бойд, Эдинбург (1925)

Фишер, Р.А.: План экспериментов. Оливер и Бойд, Эдинбург (1935)

Межквартильный диапазон

Межквартильный размах —это мера дисперсии, соответствующая разнице между первым и третьим квартилями.

Таким образом, он соответствует интервалу, который содержит 50% наиболее центрированных наблюдений распределения.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим Q1 и Q3 как первую и третью четверти распределения. Межквартильный диапазон рассчитывается следующим образом:

$$\text{Межквартильный размах} = Q3 - Q1\ .$$

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Межквартильный размах — это мера изменчивости, не зависящая от числа наблюдений. Более того, на противоположном конце диапазона эта мера менее чувствительна к выбросам.

ПРИМЕРЫ

Рассмотрим следующий набор из десяти наблюдений:

0112223345

Первый квартиль Q1 равен 1, а третий квартиль Q3 равен 3. Таким образом, межквартильный размах равен:

.

наблюдения располагаются в интервале

длина 2, что означает интервал между 1 и 3.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Мера рассеивания
- Квартиль
- Диапазон

Интервал

Интервал определяется двумя границами, a и b. Это так называемые предельные значения интервал.

Интервалы могут быть закрытыми, открытыми, полуоткрытыми и полузакрытыми.

- Закрытый интервал:

  обозначенный [a, b], представляет собой набор x с:

  $$a \leq x \leq b.$$

- Открытый интервал:

  обозначается]a, b[, представляет собой набор x с:

  $$a < x < b.$$

- Интервал, который является полуоткрытым слева или полузамкнутым

  интервалом справа: – обозначается]a, b], представляет набор x с:

  $$a < x \leq b.$$

- Интервал, который является полуоткрытым справа или полузамкнутым

  слева: обозначается [a, b[, представляет набор x с:

  $$a \leq x < b.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение частоты
- Гистограмма
- ожив

Обратная матрица

Для квадратной матрицы A порядка n мы называем обратную матрицу A квадратной матрицей порядка n, так что результат матричного произведения матрицы A и ее обратной (подобно результату обратного умножения A) равен единичной матрице порядка n. Для любой квадратной матрицы обратная матрица, если она существует, единственна. Чтобы иметь обратимую квадратную матрицу A, ее определитель должен быть отличен от нуля.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть A — квадратная матрица порядка n с ненулевым определителем. Мы называем обратную A

единственная квадратная матрица порядка n, обозначаемая A-1, например:

$$A \cdot A^{-1} = A^{-1} \cdot A = B,$$

где In — единичная матрица порядка n. Поэтому мы говорим, что матрица невырожденная. Когда det(A) = 0, т. е. матрица A необратима, говорят, что она вырожденная.

Machine Translated by Google

Условие, что определитель A может не быть нулю объясняется свойствами определителя:

$$1 = \det(I_n) = \det A \cdot A^{-1}$$

$$= \det(A) \cdot \det A^{-1},$$

то есть $\det A^{-1} = \frac{1}{\det(A)}$.

С другой стороны, выражение A^{-1}, из A, получается с помощью $\det(A)^{-1}$ (которое (должен быть определен) и равен:

$$A^{-1} = \frac{1}{i+j \cdot (-1)} \det(A) \cdot \text{Дий},$$

где Dij — определитель порядка n − 1, который мы получаем, удаляя в A (транспозицию A) строку i и столбец j. Dij называется кофактором элемента aij.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Обратная матрица в основном используется в линейной регрессии в матричной форме для определения оценок параметров.
В общем случае можно сказать, что понятие обратной матрицы позволяет решить систему линейных уравнений.

ПРИМЕРЫ

Рассмотрим общий случай квадратной матрицы порядка 2: Пусть $A = \begin{matrix} a & b \\ c & d \end{matrix}$. сдерживание _ minant of A определяется как:

$$\det(A) = a \cdot d - b \cdot c .$$

Если последний не равен нулю, определяем:

$$A^{-1} = \frac{1}{a \cdot d - b \cdot c} \cdot \begin{matrix} d & -b \\ -c & a \end{matrix} .$$

Таким образом, мы можем проверить, что $A \cdot A^{-1} = A^{-1} \cdot A = I_2$ прямым вычислением.

Пусть B — квадратная матрица третьего порядка:

$$B = \begin{matrix} 3 & 2 & 1 \\ 4 & -1 & 0 \\ 2 & 2 & 0 \end{matrix},$$

и давайте определим обратную матрицу B^{-1} матрицы B, если она существует.

Вычислим сначала определитель B, разверну в по последнему столбцу:

$$\det(B) = 1 \cdot [4 \cdot (-2) - (-1) \cdot 2],$$

$$\det(B) = -8 + 2 = -6 .$$

Последний отличен от нуля, мы можем вычислить B^{-1} путем определения Dij, т. е. определителей 2-го порядка, полученных удалением i-й строки и j-го столбца транспонированного Б из Б:

$$Б = \begin{matrix} 3 & 4 & 2 \\ 2 & -1 & 2 \\ 1 & 0 & 0 \end{matrix}$$

$$D11 = -1 \cdot 0 - (-2) \cdot 0 = 0 ,$$

$$D12 = 2 \cdot 0 - (-2) \cdot 1 = 2 ,$$

$$D13 = 2 \cdot 0 - (-1) \cdot 1 = 1 ,$$

$$\vdots$$

$$D33 = 3 \cdot (-1) - 4 \cdot 2 = -11 .$$

Таким образом, мы получаем обратную матрицу B^{-1}, подбирая правильный знак к элементу Dij [то есть (-), если сумма i и j четна, и (+), если она нечетна]:

$$B^{-1} = -\frac{1}{6} \cdot \begin{matrix} 0 & -2 & 1 \\ 0 & -2 & 4 \\ -6 & 10 & -11 \end{matrix}$$

$$= \frac{1}{6} \cdot \begin{matrix} 0 & 2 & 1 \\ 0 & 2 & -4 \\ 6 & -10 & 11 \end{matrix} .$$

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- определитель
- Обобщенный обратный
- Матрица
- Множественная линейная регрессия
- Транспонировать

Инверсия

Рассмотрим перестановку целых чисел 1, 2,...,n. Когда в паре элементов (не обязательно соседних) перестановки первый элемент больше второго, это называется инверсией. Таким образом можно подсчитать в перестановке количество инверсий, то есть количество пар, в которых большое целое число предшествует меньшему.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим перестановку целых чисел 1, 2,...,n, заданную формулами σ1, σ2,..., σn; количество инверсий определяется подсчетом числа пар (σi, σj) с σi > σj.

Чтобы убедиться, что мы ничего не забыли, начнем с того, что возьмем σ1 и посчитаем σi, i > 1, меньшие, чем σ1, а затем возьмем σ2 и посчитаем элементы, расположенные ниже него справа, продолжая таким образом до σn 1; тогда получается число инверсий рассматриваемой перестановки.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Количество инверсий используется для вычисления позднего определителя квадратной матрицы. Обозначение aij используется для элемента матрицы, расположенного на пересечении i-я строка и j-й столбец. Осуществляя произведение a1i · a1j · ple, четность числа ··· · a1r, для экзамена инверсий перестановки ij...r определяет знак отнести к произведению для расчета определителя.

Для перестановки элементов 1, 2,...,n количество инверсий всегда равно целому числу, расположенному между 0 и $\frac{n \cdot (n-1)}{2}$.

ПРИМЕРЫ

Рассмотрим целые числа 1, 2, 3, 4, 5; перестановка этих пяти целых чисел дается, например, как:

$$14523 .$$

Количество инверсий этой перестановки можно определить, записав пары, в которых есть инверсия:

(4, 2)
(4, 3)
(5, 2)
(5, 3),

что дает число инверсий, равное 4.

Для перестановки 1 2 3 4 5 (идентичность на мутацию) мы обнаруживаем, что инверсии нет. Для перестановки 5 4 3 2 1 мы получаем 10 инверсий, потому что каждая пара является инверсией, следовательно, количество инверсий равно

равно количеству комбинаций двух из пяти элементов, выбранных без повторения, что означает:

$$C_5^2 = \frac{5!}{2! \cdot 3!} = 10 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- определитель
- Перестановка

Не правильная вариация

Нерегулярные вариации или случайные вариации составляют один из четырех компонентов времени.

Machine Translated by Google

ряд. Они соответствуют движениям, которые появляются нерегулярно и обычно в течение коротких периодов времени.

Нерегулярные вариации не следуют определенной модели и не предсказуемы.

На практике все компоненты временного ряда, которые нельзя отнести к влиянию

циклические колебания или сезонные колебания или колебания вековой тенденции классифицируются как нерегулярные.

ИСТОРИЯ

Смотрите временной ряд.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть Yt — временной ряд; мы можем описать его с помощью его составляющих: • Вековая тенденция Tt • Сезонные вариации St • Циклические колебания Ct • Нерегулярные вариации It Мы различаем: • Мультипликативную модель: Yt = Tt · St ·Ct ·It. • Аддитивная модель: Yt = Tt + St + Ct + It.

Определив вековую тенденцию, сезонные колебания и циклические колебания, можно скорректировать исходные данные временного ряда по этим трем составляющим. Таким образом, мы получаем значения нерегулярных вариаций в каждый момент времени t с помощью следующих выражений:

$$\frac{Yt}{Ст \cdot Тт \cdot Кт} = It \text{ (мультипликативная модель)}$$

Yt St Tt Ct = It (аддитивная модель).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Изучение компонентов временных рядов показывает, что нерегулярные вариации можно разделить на две категории:

• К наиболее многочисленным можно отнести большое количество мелких причин, особенно погрешностей измерения, провоцирующих вариации малой амплитуды. • Возникшие в результате отдельных случайных событий более крупного масштаба, таких как забастовки, административные решения, финансовые бумы, стихийные бедствия и т. д.

В этом случае, чтобы применить этот метод скользящей средней, мы делаем первую обработку данных, чтобы исправить, насколько это возможно, необработанные данные. данные.

Даже если мы обычно предполагаем, что нерегулярные вариации не вызывают устойчивых вариаций, а только в течение короткого промежутка времени, мы можем представить, что они достаточно сильны, чтобы следовать циклическим или другим флуктуациям.

На практике мы обнаруживаем, что нерегулярные вариации имеют тенденцию иметь не большую амплитуду и следовать нормальному распределению.

Это означает, что часто бывают не большие отклонения и редко большие.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Циклическое колебание
- Нормальное распределение
- Сезонная вариация
- Светская тенденция
- Временная последовательность

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бауэрман, Б.Л., О'Коннелл, Р.Т.: Временные ряды и прогнозирование: прикладной подход. Даксбери, Белмонт, Калифорния (1979)

Харнетт, Д.Л., Мерфи, Дж.Л.: Вводный статистический анализ. Аддисон-Уэсли, Ридинг, Массачусетс (1975)

Machine Translated by Google

Метод складного ножа

При точечных оценках по доверительному интервалу часто бывает трудно оценить погрешность или стандартную ошибку используемой оценки. тор. Причины обычно сводятся к двум пунктам: (1) отсутствие информации о форме распределения теоретических данных.

2) сложность теоретической модели, если гипотеза считается слишком сильной. Метод складного ножа позволяет уменьшить биал при численной оценке.

стандартной ошибки, а также достоверности.

интервал. Обычно это делается путем создания искусственных подвыборок из данных полной выборки.

ИСТОРИЯ

Метод складного ножа был разработан Кенуем (1949, 1956) и Джоном Уайлдером Тьюки (1958). В настоящее время это наиболее широко используемый метод благодаря компьютерам, которые могут генерировать большое количество данных за очень короткое время.

Создание наименования складной ножприписывается Тьюки, Дж В. (1958). Название этого метода произошло от ножа разведчика с

множество лезвий, готовых к использованиюв различных ситуациях.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x1,...,xn —количество независимых наблюдений случайной величины. Мы

хотят оценить параметр θ населения . Техника складного ножа заключается в следующем: оценка θˆ θвычисляется по исходной выборке . Последние разбиваются на J групп, где m = их (для упрощения задачи и предположим, $\frac{н}{Дж}$ наблюдения в каждом из что n кратно J).

Оценка θ по методу складного ножа представляет собой среднее значение псевдозначений:

$$\hat{1}\,\theta^{-} = \frac{}{Дж} \cdot \sum_{j=1}^{Дж} j\,,_$$

где $\hat{\theta}_j$ называется псевдозначением и определяется как:

$$j_ = J \cdot \hat{\theta} \quad (J \quad 1) \cdot \hat{\theta} \qquad (к)$$

а оценки θˆ (j) (j = 1,...,J) вычисляются так же, как θˆ, но с удалением j-й группы.

Оценка дисперсии определяется _ _

к:

$$S2 = Var(\hat{}\,\theta)^{-} = \sum_{j=1}^{Дж} \frac{(\hat{\theta}_{Дж} — \theta)^{-2}}{(Дж- 1) \cdot Дж} \cdot$$

Machine Translated by Google

Оценщик смещения :

$$B(\hat{\theta})^{-} = (J \quad 1) \cdot (\theta\hat{} \quad (.) \quad \theta)\hat{},$$

где (.) определяется следующим образом:

$$\theta\hat{}(.) \text{ "=" } \frac{1}{Дж} \cdot \sum_{j=1}^{Дж} \theta\hat{}(к).$$

Доверительный интервал при значимости уровень α задается:

$$\hat{\theta}^{-} \quad \frac{S \cdot tn \quad 1,1 \quad \alpha}{\sqrt{н}}; \hat{\theta}^{-} + \frac{S \cdot tn \quad 1,1 \quad \alpha}{\sqrt{н}} \quad .$$

Одним из важных свойств метода складного ножа является тот факт, что мы удаляем из уклон срок заказа $\frac{1}{н}$.

Предположим, что a1, a2,... —функции θ;

мы можем представить как математическое ожидание θˆ:

$$E[\theta\hat{}] = \theta + \frac{a1}{п} + \frac{a2}{п2} + \ldots$$

$$= \theta + + + \frac{a1}{м \cdot Дж} \quad \frac{a2}{(м \cdot Дж)2}$$

и

$$E[\theta\hat{} j] = J \cdot E[\theta\hat{}] \quad (J \quad 1) \cdot E[\theta\hat{}(j)]$$

$$= J \cdot \theta + + + \ldots \frac{a1}{м \cdot Дж} \quad \frac{a2}{(м \cdot Дж)2}$$

$$(J \quad 1) \cdot \theta + + \ldots \quad \frac{a1}{м \cdot (Дж - 1)}$$

$$= J \cdot \theta + \frac{a1}{м} + + \frac{a2}{\ldots m2} \cdot J\, a2 \qquad J \cdot \theta$$

$$+ \theta - \frac{a1}{м} \quad \frac{}{м2 \cdot (Дж \quad 1)\, 1} \quad \ldots$$

$$= \theta \quad a2 \cdot + \ldots \frac{}{m2 \cdot Дж \cdot (Дж \quad 1)}$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Идея метода складного ножа, которая относительно старый, стал современным благодаря информатика. Складной нож генерирует выборки из исходных данных и определяет доверительный интервал при выбранном уровне значимости неизвестной статистики.

В практических случаях группы составляют только наблюдении (m = 1). Оценки θˆ (j) вычисляются удаление j-го наблюдение.

ПРИМЕРЫ

Крупная сеть магазинов хочет определить эволюция его товарооборота между 1950 и 1960.

Для получения правильной оценки возьмем выборка из 8 точек продаж, где мы знаем соответствующие обороты.

Данные обобщены следующем стол:

Торговая точка	Оборот (млн $)	
	1950 г.	1960 г.
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

Общий оборот 8 точек продаж увеличилась с 64,5 в 1950 г. до 86,37 млн долл. в 1960 году.

Таким образом, оценка повышающего коэффициента равно:

$$\theta\hat{} = \frac{86,37}{64,5} = 1,34.$$

Мы хотим рассчитать смещение этой оценки, а также ее стандартное отклонение.

Вычислим θˆ (j) для j от 1 до 8, то есть рассчитаны коэффициенты повышения без данных j-й торговой точки. Мы

Machine Translated by Google

получать:

$$\hat{\theta}_{(1)} = \frac{86{,}37 - 7{,}59}{64{,}5 \quad 6{,}03\,\hat{\theta}} = 1{,}347 \quad ,$$

$$\hat{\theta} = 1{,}358 \ (2) \qquad (6) = 1{,}327$$

$$\hat{\theta} = 1{,}358 \ (3) \qquad \hat{\theta}_{(7)} = 1{,}324$$

$$= 1{,}325 \ (4) \qquad (8) = 1{,}340$$

$$\hat{\theta}(5) = 1{,}337 .$$

Таким образом, мы можем вычислить $\hat{\theta}(.)$:

$$\hat{\theta}(.) = \frac{1}{Дж} \cdot \sum_{j=1}^{Дж} \hat{\theta}_{(к)} = \frac{1}{8} \cdot (10{,}716) = 1{,}339 .$$

Вычислим псевдозначения $\hat{\theta}_{Дж}$. Напомним, что $\hat{\theta}_j$ рассчитывается следующим образом:

$$\hat{\theta}_{Дж} = Дж \cdot \hat{\theta} \quad (Дж \quad 1) \cdot \hat{\theta} \quad (к) .$$

Мы нашли:

$$\hat{\theta}1 = 1{,}281 \qquad \hat{\theta}5 = 1{,}356$$

$$\hat{\theta}2 = 1{,}207 \ \hat{\theta} \qquad \hat{\theta}6 = 1{,}423$$

$$3 = 1{,}210 \qquad \hat{\theta}7 = 1{,}443$$

$$\hat{\theta}4 = 1{,}436 \qquad \hat{\theta}8 = 1{,}333 .$$

Вычислим оценку θ по вал е ту
метод ножа:

$$\hat{\bar{\theta}} = \frac{1}{Дж} \cdot \sum_{j=1}^{Дж} \hat{\theta}_{Дж} = \frac{1}{8} \cdot (10{,}689) = 1{,}336 .$$

Итак, определяем уклон:

$$B(\hat{\theta})^{-} = (Дж \quad 1) \cdot (\hat{\theta} \quad (.) \quad \theta)^{\wedge}$$

$$= (8 - 1) \cdot (1{,}339 - 1{,}34)$$

$$= -0{,}0035 .$$

Стандартная ошибка определяется следующим образом:

$$S = \sqrt{Var(\hat{\theta})^{-}}$$

$$= \sqrt{\frac{\sum_{j=1}^{Дж} (\hat{\theta}_{Дж} - \theta)^{-2}}{(Дж - 1) \cdot Дж}}$$

$$= \sqrt{\frac{0{,}0649}{7 \cdot 8}} = 0{,}034 .$$

Таким образом, мы можем рассчитать доверительный интервал между
val коэффициента увеличения между 1950 и
1960 годами оборота сети. Для уровня значимости α,
равного 5%, критическое значение
Таблица Студенческого для 7 свободы:

$$t_{7,0,975} = 2{,}365 .$$

Доверительный интервал для оценки
повышающий коэффициент:

$$\hat{\bar{\theta}} \pm \frac{S \cdot t_{7,0,975}}{\sqrt{n}} \quad \hat{\bar{\theta}} + \frac{S \cdot t_{7,0,975}}{\sqrt{n}}$$

$$1{,}34 - \frac{0{,}034 \cdot 2{,}365}{\sqrt{8}} ,$$

$$1{,}34 + \frac{0{,}034 \cdot 2{,}365}{\sqrt{8}} ,$$

то есть:

$$[1.312, 1.368].$$

Таким образом, мы можем утверждать, что среднее увеличение
в обороте примерно содержится
от 31% до 37% в течение периода
1950–1960 гг.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Предварительность

- Доверительный интервал
- Оценка
- Оценщик
- Балльная оценка
- Стандартная ошибка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Эфрон, Б.: Складной нож, бутстрап
и другие планы повторной выборки. Общество
по промышленной и прикладной математике,
Филадельфия (1982)

Quenouille, MH: Приблизительные тесты корреляции
во временных рядах. Дж. Рой. Стат. соц.
сер. Б 11, 68–84 (1949)

Machine Translated by Google

Quenouille, MH: Заметки о предвзятости в оценке. Биометрика 43, 353–360 (1956).

Тьюки, Дж. В.: Предвзятость и уверенность в не очень больших выборках. Анна. Мат. Стат. 29, 614 (1958)

Джеффрис, Гарольд

Джеффрис, Гарольд родился в Фатфилде, Англия, в 1891 году. Он изучал математику, физику, физическую химию и геологию в колледже Армстронга в Ньюкасле, где получил степень в 1910 году. Затем он поступил в колледж Св. Иоанна в Кембриджском университете, где он стал стипендиатом в 1914 году и пробыл там 75 лет. Он умер там в 1989 году.

Свою первую статью (о фотографии) он опубликовал в 1910 году. История сохранила, в частности, две его работы:

1939 Теория вероятностей. Оксфордский университет Versity Press, Оксфорд.

1946 г. (совместно с Джеффрисом, Б.С.) Методы математической физики. Издательство Кэмбриджского университета, Кембридж, Макмиллан, Нью-Йорк.

Функция совместной плотности

Предположим, что X1, X2 —непрерывные случайные величины, определенные в одном и том же пространстве выборок, что

$$F(x_1, x_2) = P(X_1 \le x_1, X_2 \le x_2) = \int_{-\infty}^{x_k} \int_{-\infty}^{x_1} f(t_1, t_2)\, dt_1, dt_2$$

для всех x x1, xк. Тогда f (x1, x2) —совместная функция плотности (X1, X2) при условии, что f(x1, x2) 0 и что

$$\int f(x_1 x_2)\, dx_1 dx_2 = 1 \quad .$$

Предположим, что (X1,...,Xk) —непрерывные случайные величины, определенные в одном и том же пространстве выборок, и

$$F(x_1,...,x_k) = P(X_1 \le x_1,...,X_k \le x_k) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f(t_1,...,t_k)\, dt_1 \ldots dt_k$$

для всех x1,...,xk. Тогда f(x1,...,xk) —совместная функция плотности (x1,...,xk), если f(x1,...,xk) 0 и

$$\int \cdots \int f(x_1,...,x_k)\, dx_1 \ldots dx_k = 1 \quad .$$

ССЫЛКИ Найт, К.: Математическая статистика. Чапман и Холл / CRC, Лондон (2000)

Совместная распределительная функция

Совместная функция распределения пары случайных величин определяется как вероятность того, что первая случайная величина примет значение, меньшее или равное первому действительному числу, и что одновременно вторая переменн

Machine Translated by Google

способный принимает значение, меньшее или равное второму действительному числу.

ИСТОРИЯ

См. вероятность.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим две случайные величины X и Y.

Совместная функция распределения определяется следующим образом:

$$F_{XY}(a, b) = P(X \leq a, Y \leq b), \quad -\infty < a, b < \infty.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция распределения
- Предельная функция распределения
- Пара случайных величин

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Хогг, Р.В., Крейг, А.Т.: Введение в математическую статистику, 2-е изд. Макмиллан, НьюЙорк (1965)

Росс, С.М.: Введение в вероятностные модели, 8-е изд. Джон Уайли, НьюЙорк (2006 г.)

Совместная вероятность Функция распределения

Предположим, что X_1, X_2 — дискретные случайные величины, определенные в одном и том же пространстве выборок. Тогда совместная функция распределения вероятностей $X = (X_1, X_2)$ определяется как

$$f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2).$$

Совместная функция распределения вероятностей должна удовлетворять следующим условиям:

$$\sum_{x_1, x_2} f(x_1, x_2) = 1.$$

Множественная совместная функция распределения вероятностей $X = (X_1, \ldots, X_k)$ определяется как $f(x_1, \ldots, x_k) = P(X_1 = x_1, \ldots, X_k = x_k)$, удовлетворяющая

$$f(x_1, \ldots, x_k) = 1 \ (x_1, \ldots, x_k) \cdot$$

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Casella, G., Berger, RL: Статистический вывод. Даксбери, Пасифик-Гроув (2001)

Machine Translated by Google

Кендалл, Морис Джордж

Кендалл, Морис родился в 1907 году в Кеттеринге, графство Нортгемптоншир, Англия. Он изучал математику в колледже Св. Иоанна в Кэмбридже. После окончания факультета математики Вранглер в 1929 году поступил на британскую гражданскую службу в Министерстве сельского хозяйства. Он был избран членом Общества в 1934 году.

В 1937 году работал с Г. Удным Юлем в пересмотре его стандартного статистического учебника, Введение в теориюстатистики. Он также поработайте над коэффициентом ранговой корреляции который носит его имя, тау Кендалла, который в конечном итоге привело к созданиюмонографии по соотношениюRank Cor в 1948 году.

В 1938 и 1939 годах он начал работать вместе с Бернард Бабингтон Смит о проблеме генерации случайных чисел, разработав как один из первых ранних механических устройств для получения случайных цифр и сформулировал ряд тестов, таких как частота тест, серийный тест и покерный тест на статистическую случайность в заданном наборе цифр.

Во время войны ему удалось выпустить один из томов Advanced TheoryofStatistics. в 1943 г. и второй том в 1946 г.

В 1957 году он опубликовал «Многомерный анализ». и в том же году он также разработал, с WR Buckland, Словарь статистических Условия.

В 1953 г. он опубликовал «Аналитику экономических временных рядов», а в 1961 г. Лондонского университета и занял должность управляющий директор консалтинговой компании Scientific Control Systems, а в в том же году начался двухгодичный срокПрезидент Королевского статистического общества.

В 1972 году он стал директором Мирового о Исследование рождаемости, проект, спонсируемый Международный статистический институт и Объединенные Нации. Он продолжил эту работу до 1980 года, когда болезнь вынудила его уйти на пенсию Он был посвящен в рыцари в 1974 году за свои заслуги. теории статистики и получил Медаль мира Организации Объединенных Наций в 1980 г. в знак признания его работы над Всемирным обзором фертильности. Он также был избран членом Британской академии и получил высокую юнаграду Королевского статистического общества, Золотая медаль Гая. На момент его смерти в 1983 году он был почетным президентом Международный статистический институт.

Некоторые основные работы и статьи Кен Далла, Мориса Джорджа:

1938 г. (совместно с Бабингтоном Смитом, Б.) Случайность и числа случайной выборки. Дж Рой. Стат. соц. 101:1, 147–166.

1979 (совместно со Стюартом А.) Продвинутая теория Статистика. Арнольд.

Machine Translated by Google

1957 (совместно с Бакленд, WR) Словарь статистических терминов. Международный al Статистический институт, Гаага, Нидерланды.

Временные ряды 1973 года, Гриффин, Лондон.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коэффициент ранговой корреляции Кендалла

Генерация случайных чисел

Кендалл Ранг Коэффициент корреляции

Коэффициент ранговой корреляции Кендалла (Kendall τ) — непараметрическая мера корреляция.

ИСТОРИЯ

Этот коэффициент ранговой корреляции был ругался еще в начале 20 века Фехнера, Г.Т. (1897 г.), Липпса, Г.Ф. (1906 г.), и Дойхлер, Г. (1914).

Кендалл М.Г. (1938) не только независимо заново открыл его, но и изучил. используя (непараметрический) подход. Его 1970 монография содержит полное подробное изложение теории, а также биографию

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим две случайные величины (X, Y) наблюдается на выборке размера n с n пары наблюдений (X1, Y1), (X2, Y2), ..., (Xn, Yn). Указание на корреляцию между X и Y может быть получено с помощью упорядочив значения Xi в порядке возрастания и путем подсчета числа соответствующих значений Yi, не удовлетворяющих этому порядку. Q будет обозначать количество инверсий среди значений Y, необходимых для получить тот же (возрастающий) порядок, что и значения X.

Поскольку существуют $\frac{n(n-1)}{2}$ различные пары, которые могут формироваться, $0 \le Q \le \frac{2n(n-1)}{2}$; значение 0 получается, когда все значения Yi уже в порядке возрастания, а значение $\frac{n(n-1)}{2}$ достигается, когда все значения Yi обратны порядок Xi, каждая пара должна быть переключена для получения желаемого заказа.

Коэффициент ранговой корреляции Кендалла, обозначается τ, определяется:

$$\tau = 1 - n(n - 1) \frac{4Q}{}.$$

Если все пары расположены в порядке возрастания, то:

$$\tau = 1 - n(n - 1) \frac{4 \cdot 0}{} = 1.$$

Если все пары в обратном порядке, то:

$$\tau = 1 - \frac{4 \cdot \frac{1}{2} \cdot n(n-1)}{n(n-1)} = -1.$$

Эквивалентное определение ранга Кендалла коэффициент может быть задан следующим образом: два наблюдения называются согласующими, если два члена одного наблюдения больше, чем соответствующие члены другого наблюдения. Например, (0,9, 1,1) и (1.5, 2.4) — два совпадающих наблюдения потому что 0,9 < 1,5 и 1,1 < 2,4. Два наблюдения называются несогласующими, если два члены одного наблюдения находятся в противоположных приказ соответствующим членам другой наблюдение. Например, (0,8, 2,6) и (1,3, 2.1) являются двумя противоречивыми наблюдениями, потому что 0,8 < 1,3 и 2,6 > 2,1.

Пусть Nc и Nd обозначают общее количество пары согласующихся и несогласующихся наблюдений соответственно.

Machine Translated by Google

Две пары для которых Xi = Xj и Yi = Yj , не являются ни согласующимися, ни несогласующими и поэтому не учитываются ни в Nc , ни в Nd.

При таком обозначении ранговый коэффициент Кендалла дан:

$$\tau = \frac{2(Nc - Nd)}{n(n-1)}.$$

Обратите внимание, что когда нет пар, для которых Xi = Xj или Yi = Yj, две формулировки тождественно одинаковы. В противоположной ситуации значения, даваемые обеими формулами, могут быть разными.

Проверка гипотезы

Коэффициент ранговой корреляции Кендалла равен часто используется в качестве статистического теста для определения если существует связь между двумя случайными переменные. Тест может быть двусторонним или односторонний тест. Гипотезы: _

A: Двусторонний чехол:

H0: X и Y взаимно независимы.

H1: существует либо положительная, либо отрицательная корреляция между X и Y.

Существует положительная корреляция, когда большие значения X связаны с большими значениями Y , а малые значения X —с малыми значениями Y.

Существует отрицательная корреляция, когда большие значения X , как правило, связаны с малыми значениями Y, и наоборот.

B: Односторонний чехол:

H0: X и Y взаимно независимы.

H1: существует положительная корреляция между X и Y.

C: Односторонний случай:

H0: X и Y взаимно независимы.

H1: существует отрицательная корреляция между X и Y.

Статистический тест определяется следующим образом:

T знак равно Nc - Nd .

Правила принятия решений

Правила принятия решений различаются в зависимости от выдвигаемых гипотез . Вот почему существуют решающие правила A, B и C относительно предыдущих случаев.

Правило принятия решения А

Отклонить H0 на значимом уровне α, если

$$T > t_{n,1-\alpha/2} \quad \text{или} \quad T < t_{n,\alpha/2},$$

где t —критическое значение теста, заданное таблицей Кендалла; в противном случае примите H0.

Правило принятия решения Б

Отклонить H0 на значимом уровне α, если

$$T > t_{n,1-\alpha}.$$

в противном случае примите H0.

Правило принятия решений C Отклонить H0 на значимом уровне α, если

$$T < t_{n,\alpha}.$$

в противном случае примите H0.

Также можно использовать

$$\tau = 1 - n(n-1)/4Q$$

в качестве статистического теста.

Когда X и Y независимо распределены в популяции, точное распределение τ имеет ожидаемое значение, равное нулю, и дисперсию:

$$\sigma_\tau^2 = \frac{2(2n+5)}{9n(n-1)}$$

и очень быстро стремится к нормальному распределению, приближение достаточно хорошее для n 10.

K

Machine Translated by Google

В этом случае для проверки независимости на 5%
уровне, например, достаточно проверить, является ли т
находится за пределами

$$\pm 1{,}96 \cdot \sigma_\tau$$

и отвергнуть гипотезу о независимости, если
так оно и есть.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Коэффициент ранговой корреляции Кендалла равен
используется в качестве проверки гипотезы для
изучения зависимости между двумя случайными величинами. Это
можно считать свидетельством независимости.
В качестве непараметрической корреляционной меры
его также можно использовать с номинальным
или порядковые данные.
Измерение корреляции между двумя
переменные должны удовлетворять следующим требованиям:

1. Его значения находятся в диапазоне от -1 до +1.
2. Существует положительная корреляция между X
   и Y, если значение коэффициента корреляции
   положительный; совершенная положительная
   корреляция соответствует значению+1.
3. Существует отрицательная корреляция между
   X и Y, если значение корреляции
   коэффициент отрицательный; совершенная
   отрицательная корреляция соответствует значению
   из 1.
4. Существует нулевая корреляция между X и
   Y, когда коэффициент корреляции близок
   до нуля; можно также сказать, что X и Y
   не коррелирован.

Коэффициент ранговой корреляции Кендалла имеет
следующие преимущества:

- Данные могут быть нечисловыми
  до тех пор, пока они могут быть классифицированы
  по определенному критерию
- Легко рассчитать.
- Соответствующий статистический тест не
  выдвинуть базовую гипотезу на основе
  форма распределения населения, из которого
  взяты выборки.

В таблице Кендалла приведены теоретические
значения статистики τ ранговой корреляции Кендалла.
коэффициент отношения, используемый в качестве статистического теста
при гипотезе независимости двух
случайные переменные.
AKendalltableможНо найти в Kaarsemaker
и ван Вейнгарден (1953).
Вот образец таблицы Кендалла для n =
4,..., 10 и α = 0,01 и 0,05:

н	a = 00.1	a = 00.5
4	6	4
5	8	6
6	11	9
7	15	11
8	18	14
9	22	16
10	25	19

ПРИМЕРЫ

В этом примере восемь пар реальных близнецов берут
тесты интеллекта. Цель состоит в том, чтобы увидеть, есть ли
независимость между тестами одного
кто рожден первым и те, кто родился
родился вторым.
Данные приведены в таблице ниже; в
наивысшие баллы соответствуют лучшим результатам.

Пара близнецов	Первый рожденный Си	Второй рожденный И
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

Затем пары классифицируются по возрастанию порядок для X и определяются согласующиеся и несогласующие пары. Это дает:

Пара близнецов (Си, И)	согласование пары	Разногласия пары
(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
	Нк = 23	Нд = 1

Коэффициент ранговой корреляции Кендалла равен предоставлено:

$$\tau = \frac{2(Нк - Нд)}{п(п - 1)}$$

$$\text{"="} \frac{2(23 - 1)}{8 \cdot 7}$$

$$= 0{,}7857 .$$

Обратите внимание, что, поскольку существует несколько наблюдений для которых Xi = Xj или Yi = Yj значение коэффициента, определяемого:

$$\tau = 1 - = 1 - = \frac{4Q}{п(п - 1)} \quad 0{,}9286 \quad \frac{4 \cdot 1}{56}$$

отличается.

В обоих случаях мы замечаем положительную корреляцию между тестами интеллекта.

Теперь проведем проверку гипотезы:

H0: существует независимость между тесты интеллекта пары близнецов.

H1: существует положительная корреляция между тесты интеллекта.

Мы выбрали значимый уровень $\alpha = 0{,}05$.

Поскольку мы находимся в случае B, H0 отбрасывается, если

$$T > t_{8,0,95} ,$$

где T = Nc Nd и $t_{8,0,95}$ — значение стол Кендалла. Так как T = 22 и $t_{8,0,95}$ = 14, H0 отвергается.

Тогда мы можем сделать вывод, что есть положительный корреляция между результатами разведки Лигентные тесты пары близнецов.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Непараметрический тест
- Тест на независимость

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Дойхлер, Г.: Über die Methoden der Kor Relationsrechnung in der Pädagogik und Психология. Zeitung für Pädagogische Psychologie und Experimentelle Päda gogik, 15, 114–131, 145–159, 229–242. (1914)

Фехнер, Г.Т.: Коллективмасслере. В. Энгельманн, Лейпциг (1897 г.)

Карсемакер Л., ван Вейнгаарден А.: Таблицы для использования ранговой корреляции. Стат. Нирланд. 7, 53 (1953)

Кендалл, М. Г.: Новая мера ранговой корреляции. Биометрика 30, 81–93 (1938).

Кендалл, М. Г.: Методы ранговой корреляции. Гриффин, Лондон (1948)

Липпс, Г.Ф.: Die Psychischen Massmethod en. F. Vieweg und Sohn, Брауншвейг, Германия (1906 г.)

Кифер, Джек Карл

Кифер, Джек Карл родился в Цинциннати, Огайо в 1924 году. Он поступил в Массачусетс. Технологического института в 1942 г., но после 1 год изучения инженерии и экономики он оставил, чтобы взяться за работу, связанную с войной во время Второй мировой войны. Его магистерская диссертация,

Machine Translated by Google

Последовательное определение максимума функции
Гарольд Фриман. Это послужило основой для его статья «Последовательный минимаксный поиск максимум», появившейся в 1953 г. «Труды Американского математического общества». В 1948 году он отправился в Кафедра математической статистики Колумбийский университет, где Абрахам Вальд был выдающимся сотрудником отдела, в который входили Тед Андерсон, Генри Шеффе и Джек Вулфовиц. Он написал докторскую диссертацию в теории принятия решений при Вулфовице и пошел в Корнельский университет в 1951 году вместе с Вулфовицем. В 1973 году Кифер был избран первым Гораций Уайт, профессор Корнельского университета город, должность, которую он занимал до 1979 года, когда он вышел в отставку и поступил на факультет Калифорнийского университета в Беркли. Он умер в 57 лет в 1981 году.

Областью исследований Кифера был дизайн эксперименты. Большинство из его 100 публикаций занимался этой темой. Он также писал документы онтопики в математической статистике в том числе теория принятия решений, стохастическое приближение, теория массового обслуживания, непараметрический вывод, оценка, последовательный анализ и условный вывод.

Кифер был сотрудником Математического института. математическая статистика и американский Statis ассоциации и президент Института математической статистики (1969–1970).

Он был избран в Американскую академию Искусств и наук в 1972 году и в Национальный Академии наук (США) в 1975 г.

Избранные произведения и публикации Джека Карл Кифер:

1987 Введение в статистический вывод. Springer, Берлин Гейдельберг Новый Йорк

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

План экспериментов

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кифер, Дж: Оптимальные экспериментальные планы. Дж Рой. Стат. соц. сер. Б 21, 272–319 (1959)

Л. Д. Браун, И. Олкин, Дж Сакс и Х.П. Винн (ред.): Джек Карл Кифер, Собрание Бумаги. I: Статистический вывод и вероятность (1951–1963). НьюЙорк (1985)

Л. Д. Браун, И. Олкин, Дж Сакс и Х.П. Винн (ред.): Джек Карл Кифер, Сборник статей. II: Статистический вывод и вероятность (1964–1984). НьюЙорк (1985)

Л. Д. Браун, И. Олкин, Дж Сакс и Х.П. Винн (ред.): Джек Карл Кифер, Собрание Бумаги. III: План экспериментов. Новый Йорк (1985)

Колмогоров, Андрей Николаевич

Колмогоров Андрей Николаевич родился в Тамбове в 1903 году.
основоположники современной вероятности. В 1920 г. поступил в МГУ и
изучал математику, историю и металлургию. В 1925 году он опубликовал свою первую статью
вероятность на неравенства частичного
суммы случайных величин, которые стали
основной справочник в области случайных
процессов. Он получил докторскую степень в
1929 г. и опубликовал 18 статей по праву
больших чисел, а также на интуитивной логике.
он был назначен профессором МГУ.
университета в 1931 году. В 1933 году он опубликовал
свою монографию по теории вероятностей.
В 1939 году он был избран членом Академии
мой наук СССР. Он получил

Machine Translated by Google

Ленинская премия 1965 года и орден Ленина шесть раз, а также Премия Лобачевского в 1987 году. Он был избран членом многих других зарубежных академий, включая Румынскую академию наук (1956 г.), Королевское статистическое общество Лондона (1956), Академия Леопольдина Германии (1959 г.), Американской академии искусств и наук (1959 г.), Лондонского математического общества (1959 г.), Американского философского общества (1961 г.), Индийского института статистики (1962 г.), Голландской академии наук (1963 г.), Лондонского королевского общества (1964 г.), Национального Академия Соединенных Штатов (1967 г.) и Французская академия наук (1968).

Избранные основные произведения Колмогорова, Андрей Николаевич:

1933 Grundbegriffe der Wahrschein lichkeitsrechnung. Springer, Берлин Гейдельберг НьюЙорк.

1933 г. – эмпирическое определение una lege di distribuzione. Джорджнале Делл'Институто Итальяно делья Атуари, 4, 83–91 (6.1).

1941 Локальная структура турбулентности в несжимаемые жидкости с очень высокой Число Рейнольдса. Дэн СССР, 30 лет, 229.

1941 г. Рассеяние энергии локально изотропная турбулентность. Докл. акад. наук. СССР, 32, 16–18.

1958 (совместно с Успенским В.А.) К определению алгоритма. (К определениюалгоритма). Успехи Математических наук 13(4):3–28, Американское математическое общество Серия переводов 2(29):217–245, 1963 год.

1961 (совместно с Фоминым С.В.) Мера, Интегралы Лебега и гильбертово пространство. Наташа Артин Брансуик и Алан Джеффри. Академик, НьюЙорк.

1963 О представлении непрерывных функций многих переменных суперпозиция непрерывных функций одной переменной и сложения. Доклады Академии Наук СССР, 114, 953–956, 1957. Английский перевод. Сделки математического общества, 28, 55–59.

1965 Три подхода к количественному определению информации. Проба леммы передачи информации, 1, 1–17. Перевод слова Проблемы передачи информации 1(1), 3–11 (1965).

1987 (совместно с Успенским В.А.) Алгоритмы и случайность. Теория вероятностей и ее применения (Теория вероятностей и ее приложения), 3(32):389–412.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Тест Колмогорова – Смирнова

Тест Колмогорова – Смирнова.

Критерий Колмогорова–Смирнова представляет собой непараметрический критерий согласия и используется для определить, различаются ли два распределения, или отличается ли лежащее в основе распределение вероятностей от гипотетического распределения. Он используется, когда у нас есть два Семплесени из двух популяций, которые могут быть другой. В отличие от теста Манна-Уитни и тест Уилкоксона, целью которого является обнаружить разницу между двумя средними или медианы, тест Колмогорова-Смирнова имеет

Machine Translated by Google

Преимущество рассмотрения функций распределения в совокупности. Критерий Колмогорова-Смирнова также можно использовать в качестве критерия согласия. В этом случае у нас есть только одна случайная выборка, полученная из совокупности, где функция распределения определена и известна.

ИСТОРИЯ

Критерий согласия выборки изобрел Андрей Николаевич Колмогоров (1933) .

Критерий Колмогорова-Смирнова для двух образцов был изобретен Владимиром Иванович ем Смирновым (1939).

В Massey (1952) мы находим таблицу Смирнова для критерия Колмогорова-Смирнова для двух выборок, а в Миллере (1956) мы находим таблицу Колмогорова для критерия согласия.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим две независимые случайные выборки: (X1, X2,...,Xn), выборка размера n из совокупности 1, и (Y1, Y2,...,Ym), выборка размерности m из совокупности 2. Обозначим соответственно F (x) и G (x) их неизвестные функции распределения.

Гипотезы

Гипотезы для проверки следующие :

A: Двусторонний чехол :

H0: F (x) = G (x) для каждого x

H1: F (x) = G (x) или хотя бы одно значение x

B: Односторонний чехол :

H0: F (x) G (x) для каждого x

H1: F (x) > G (x) хотя бы для одного значения x

C: Односторонний случай:

H0: F (x) G (x) для каждого x

H1: F (x) < G (x) хотя бы для одного значения x

В случае A мы выдвигаем гипотезу об отсутствии разницы между распределением

функции этих двух популяций.

Тогда обе популяции можно рассматривать как одну популяцию

В случае B мы выдвигаем гипотезу о том, что функция распределения совокупности 1 меньше , чем функция распределения совокупности 2. Иногда мы говорим, что, как правило, X имеет тенденцию быть меньше , чем Y.

В случае C мы выдвигаем гипотезу о том, что X больше , чем Y.

Обозначим через H1 (x) эмпирическую функцию распределения выборки (X1, X2,...,Xn) и через H2 (x) эмпирическую функцию распределения выборки (Y1, Y2,..., Ym) . Статистический тест определяется следующим образом:

A: Двусторонний случай

Статистический тест T1 определяется как наибольшее расстояние по вертикали между двумя эмпирическими функциями распределения:

$$T1 = \sup_{x} |H1(x) - H2(x)| .$$

B: случай с одним хвостом

Статистический тест T2 определяется как наибольшее расстояние по вертикали, когда H1 (x) больше , чем H2 (x):

$$T2 = \sup_{x} [H1(x) - H2(x)] .$$

C: односторонний случай

Статистический тест T3 определяется как наибольшее расстояние по вертикали, когда H2 (x) больше , чем H1 (x):

$$T3 = \sup_{x} [H2(x) - H1(x)] .$$

Machine Translated by Google

Правило принятия решения

Мы отвергаем H0 на уровне значимости α, если
соответствующий статистический тест (T1, T2 или T3).
больше значения таблицы Смирнова
имея для параметров n, m и 1 α, что
обозначим через tn,m,1 α, т.е. если

$$T1(\text{или } T2, \text{ или } T3) > tn,m,1 \quad \alpha .$$

Если мы хотим проверить качество подгонки
неизвестная функция распределения F(x) случайной
выборки из совокупности с заданной и известной
функцией распределения Fo (x),
тогда гипотезы будут такими же, как те
для тестирования двух образцов, за исключением того, что F (x) и
G (x) заменены на F (x) и Fo (x).
Если H (x) —эмпирическая функция распределения
случайной выборки, то статистические тесты
T1, T2 и T3 определяются следующим образом:

$$T1 = \sup_{\text{Икс}} |Fo(x) \quad H(x)| ,$$

$$T2 = \sup_{\text{Икс}} [Fo(x) \quad H(x)] ,$$

$$T3 = \sup_{\text{Икс}} [H(x) \quad Fo(x)] .$$

Правило принятия решения следующее: отклонить H0 при
уровень значимости α, если T1 (или T2, или T3)
больше, чем значение колмогорова
таблица, имеющая для параметров n и 1 α,
которую мы обозначим через tn,1 α, т.е. если

$$T1(\text{или } T2, \text{ или } T3) > tn,1 \quad \alpha .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Чтобы выполнить тест Колмогорова-Смирнова,
необходимо соблюдать следующее:

1. Оба образца должны быть взяты случайным образом.
от их соответствующих популяций.
2. Должна быть взаимная независимость
между двумя образцами.
3. Шкала измерения должна превосходить наименьший порядковый номер.
4. Для выполнения точного теста случайные
величины должны быть непрерывными; иначе
тест менее точен.

ПРИМЕРЫ

Первый пример относится к Колмогорову.
Тест Смирнова для двух образцов и второго
один для проверки пригодности.
В классе у нас 25 учеников: 15 мальчиков и
10 девочек. Мы проводим тест умственных вычислений,
чтобы увидеть, склонны ли мальчики быть лучше, чем
девушки в этом доме.
Данные представлены в следующей таблице;
наивысшие баллы соответствуют результатам
теста.

Мальчики (См)		Девушки (И)	
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			

Проверяем гипотезу, согласно которой
распределения результатов девушек и
таковые у мальчиков идентичны. Это означает
что совокупность, из которой сделана выборка
из X взятых имеет одинаковую функцию распределения
как совокупность, из которой взята выборка
Y берется. Отсюда нулевая гипотеза:

$$H0 : F(x) = G(x) \text{ для каждого } x .$$

Если здесь применяется двусторонний случай, мы
вычисляем:

$$T1 = \sup_{\text{Икс}} |H1(x) \quad H2(x)| ,$$

где H1 (x) и H2 (x) – эмпирические функции
распределения выборок

K

Machine Translated by Google

(X1, X2,...,X15) и (Y1, Y2,...,Y10),
соответственно. В следующей таблице имеем
классифицировали наблюдения двух выборок в
в порядке возрастания для упрощения вычислений
H1 (x) H2 (x).

Си	И	ч ас1 (икс) — ч ас2 икс)
	7.1	(0 - 1/10 = - 0,1
7.7		1/15 - 1/10 = - 0,0333
	8,6	1/15 - 2/10 = - 0,1333
9.2		2/15 - 2/10 = - 0,0667
10,6		3/15 - 2/10 = 0
	10,7	3/15 - 3/10 = - 0,1
	11.1	3/15 - 4/10 = - 0,2
11.3		4/15 - 4/10 = - 0,1333
12.3		5/15 - 4/10 = - 0,0667
13.3		6/15 - 4/10 = 0
14,0		7/15 - 4/10 = 0,0667
	14.1	7/15 - 5/10 = - 0,0333
15.2		8/15 - 5/10 = 0,0333
15,6		9/15 - 5/10 = 0,1
	15,7	9/15 - 6/10 = 0
16,0		10/15 - 6/10 = 0,0667
16,4		11/15 - 6/10 = 0,1333
17,5		12/15 - 6/10 = 0,2
	17,7	12/15 - 7/10 = 0,1
17,9		13/15 - 7/10 = 0,1667
19,8		14/15 - 7/10 = 0,2333
	20,3	14/15 - 8/10 = 0,1333
	21,0	14/15 - 9/10 = 0,0333
21.1		1 - 9/10 = 0,1
	23,6	1 - 1 = 0

Имеем тогда:

$$T1 = \sup_{\text{Икс}} |H1(x) \quad H2(x)|$$

$$= 0{,}2333 .$$

Значение таблицы Смирнова для n =
15, m = 10 и 1 α = 0,95 равно
t15,10,0,95 = 0,5.
Таким образом, T1 = 0,2333 < t15,10,0,95 = 0,5,
и H0 нельзя отбросить. Это значит, что
существенной разницы в уровне нет
умственных расчетов девочек и мальчиков.
Рассмотрим следующую случайную выборку
размерность 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.
Мы хотим проверить по Колмогорову –
Смирнов свидетельствует, что эта выборка получена
из равномерного распределения. Функция распределения
Равномерное распределение определяется выражением:

$$\text{Фо}(x) = \begin{cases} 0, & \text{если } x < 0 \\ x, & \text{если } 0 \quad x < 1 \\ 1 & \text{иначе} . \end{cases}$$

Тогда нулевая гипотеза H0 выглядит следующим
образом, где F (x) —неизвестная функция
распределения совокупности, связанной с
к образцу :

$$H0 : F(x) = Fo(x) \text{ для каждого } x .$$

Если применяется двусторонний случай, мы вычисляем:

$$T1 = \sup_{\text{Икс}} |Fo(x) \quad H(x)| ,$$

где H (x) —эмпирическое распределение
функция выборки (X1, X2,...,X10).
Вследующей таблице мы классифицируем 10
наблюдений в порядке возрастания, чтобы упростить
вычисление F0 (x) H (x).

Си	Фо (икс)	ч ас (икс)	Фо ()икс— ч ас ()икс
0,134	0,134	0,1	0,134 - 0,1 = 0,034
0,222	0,222	0,2	0,222 - 0,2 = 0,022
0,239	0,239	0,3	0,239 - 0,3 = - 0,061
0,322	0,322	0,4	0,322 - 0,4 = - 0,078
0,523	0,523	0,5	0,523 - 0,5 = 0,023
0,578	0,578	0,6	0,578 - 0,6 = - 0,022
0,695	0,695	0,7	0,695 - 0,7 = - 0,005
0,763	0,763	0,8	0,763 - 0,8 = - 0,037
0,937	0,937	0,9	0,937 - 0,9 = 0,037
0,980	0,980	1,0	0,980 - 1,0 = - 0,020

Machine Translated by Google

Получаем тогда:

$$T1 = \sup_{x} |Fo(x) - H(x)| = 0{,}078 .$$

Значение таблицы Колмогорова для n = 10 и 1 − α = 0,95 равно t10,0,95 = 0,409.

Если T1 меньше t10,0,95 (0,078 < 0,409), то H0 нельзя отклонить. Это означает, что случайная выборка может быть получена из равномерно распределенной совокупности.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Тест на пригодность
- Проверка гипотезы
- Непараметрический тест

ССЫЛКА

Колмогоров, А.Н.: Эмпирическое определение пути распределения. Giornale dell'Instituto Italiano degli Attuari 4, 83–91 (6.1) (1933)

Мэсси, Ф.Дж: Таблица распределения отклонений между двумя кумулятивными значениями выборки. Анна. Мат. Стат. 23, 435–441 (1952)

Миллер, Л.Х.: Таблица процентных пунктов статистики Колмогорова. Варенье. Стат. доц. 31, 111–121 (1956)

Смирнов Н.В.: Оценка отклонения между эмпирическими функциями распределения в двух независимых выборках. (Русский). Бык. Московский ун-т. 2 (2), 3–16 (6.1, 6.2) (1939)

Смирнов Н.В.: Таблица для оценки согласия эмпирических распределений. Энн.Математика. Стат. 19, 279–281 (6.1) (1948)

Таблица Крускала-Уоллиса

В таблице Крускала–Уоллиса приведены теоретические значения статистики H Крускала–Уоллиса, критерий Уоллиса в соответствии с гипотезой о том, что нет разницы между k (k 2) популяциями, которые мы хотим сравнить.

ИСТОРИЯ

См. тест Краскела-Уоллиса.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть k будет количеством выборок, вероятно, разных размеров n1, n2,...,nk. Обозначим через N общее количество наблюдений:

$$H = \sum_{i=1}^{k} n_i .$$

Мы классифицируем N наблюдений в порядке возрастания без учета того, к каким выборкам они принадлежат. Затем мы присваиваем ранг 1 наименьшему значению, ранг 2 — следующему наибольшему значению и так далее до ранга N, который присваивается наибольшему значению. Обозначим через Ri сумму рангов, присвоенных наблюдениям выборки i:

$$R_i = \sum_{j=1}^{n_i} R(X_{ij}), \qquad i = 1, 2, \ldots, k,$$

где Xij представляет собой наблюдение j выборки i, а R(Xij) — соответствующий ранг. Когда много наблюдений идентичны и имеют одинаковый ранг, мы присваиваем им средний ранг (см. тест Краскела-Уоллиса). При отсутствии средних рангов статистический тест определяется следующим образом:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) .$$

О том, что делать, если есть средние ранги, см. Тест Краскела-Уоллиса.

В таблице Краскела-Уоллиса приведены значения статистики H теста Краскела-Уоллиса в случае трех выборок для различных значений n1, n2 и n3 (при n1, n2, n3 5).

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Таблица Краскела-Уоллиса используется для параметрических тесты, использующие ранги и, в частности, тесты с тем же именем.

Когда количество выборок i больше 3, мы можем сделать аппроксимацию значения таблицы Краскела-Уоллиса таблицей хи-квадрата с k 1 степенями свободы.

ПРИМЕРЫ

См. Приложение D.

Пример использования таблицы Краскела-Уоллиса см. в разделе Тест Краскела-Уоллиса.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- таблица хи-квадрат
- Тест Крускала-Уоллиса
- Статистическая таблица

ССЫЛКИ

Kruskal, WH, Wallis, WA: Использование рангов в однокритериальном дисперсионном анализе. Варенье. Стат. доц. 47, 583–621 и опечатки, там же. 48, 907–911 (1952)

Тест Крускала-Уоллиса

Критерий Крускала-Уоллиса — это непараметрический критерий, целью которого является определение того, все ли k популяций идентичны или по крайней мере одна из популяций имеет тенденцию давать наблюдения, отличные от наблюдений других популяций.

Тест используется, когда у нас есть k выборок с k 2, происходящих из k популяций, которые могут быть разными.

ИСТОРИЯ

Тест Крускала-Уоллиса был разработан в 1952 году Крускалом, У. Х. и Уоллисом, Вашингтон.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Данные представлены k выборках. Обозначим через ni размерность выборки i, если i = 1,...,k, а через N общее количество наблюдений:

$$H = \sum_{i=1}^{k} n_i .$$

Мы классифицируем N наблюдений в возрастающем порядке, не принимая во внимание, принадлежат ли они к одним и тем же выборкам. Затем мы присваиваем ранг 1 наименьшему значению, ранг 2 — следующему наибольшему значению и так далее до тех пор, пока N не будет присвоено наибольшему значению.

Пусть Xij будет j-м наблюдением в выборки i, и пусть i = 1,...,k и j = 1,...,ni; тогда мы обозначаем ранг, присвоенный Xij, через R Xij.

Если много наблюдений имеют одинаковое значение, мы присваиваем им средний ранг. Сумма рангов, присвоенных наблюдениям выборки i, обозначается Ri, и мы имеем:

$$Ри = \sum_{j=1}^{ни} R\,Xij , \quad я = 1,...,к.$$

Если средние ранги отсутствуют (или их число ограничено), то статистический критерий определяется следующим образом:

$$H = \frac{12}{H\,(H+1)} \sum_{я=1}^{к} \frac{R2_{я}}{ни} \quad 3\,(N+1) .$$

Если, наоборот, средних рангов много, то необходимо сделать поправку и вычислить:

$$H = \frac{}{1 - \dfrac{g^{\text{З т - тия}}}{i=1\ N3}} ,$$

N, где g — количество групп средних рангов, а ti — размерность i-й такой группы.

Гипотезы

Цель теста Крускала-Уоллиса - определить, идентичны ли все популяции.

Machine Translated by Google

или если хотя бы одна из популяций стремится к
дают наблюдения, отличные от других популяций.
Гипотезы следующие:

H0: Между тек-попом нет никакой разницы
исп.

H1: По крайней мере одна из популяций отличается
от других популяций.

Правило принятия решения

Если имеется 3 выборки, каждая из которых имеет
размерность меньше или равную5, и если нет
средние ранги (то есть, если вычисляется H), то
мыиспользуем таблицу Краскела-Уоллиса для проверки H0.
Правило принятия решения следующее: Мыотвергаем
нулевая гипотеза H0 при значимости
уровне α, если T больше значения
таблица с параметрами ni, k 1 и 1 α,
обозначают hn1,n2,n3,1 α, а если нет точной
таблицыили имеются средние ранги, то
можно приблизительно оценить значение
таблица Крускала–Уоллиса по распределению
хи-квадрата с k 1 степенями свободы
(распределение хи-квадрат), то есть, если:

H > hn1,n2,n3,1 α (таблица Краскела–Уоллиса)

или H > χ2 k 1,1 α (таблица хи-квадрат) .

Соответствующее решающее правило основано на
H > x2 k 1,1 α (таблица хи-квадрата).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Следует соблюдать следующие правила
сделать тест Крускала-Уоллиса:

1. Все образцы должны быть случайными образцами.
взятыи из соответствующих популяций.
2. Помимо независимости внутри
каждой выборки должна быть взаимная
независимость между различными выборками.
3. Шкала измерения должна быть не ниже ordi
нал .

Если тест Крускала-Уоллиста заставит нас отвергнуть
нулевой гипотезы H0, мыможем использовать
критерий Уилкокса для всех выборок, взятых парами, чтобы
определить, какие пары популяций склонных
будь другим.

ПРИМЕРЫ

Готовим картошку на 4-х разных маслах.Хотим
проверить, поглощается ли количество жира
картофеля зависит от типа используемого масла. Мы
провести 5 разных экспериментов с маслом 1, 6
с маслом 2, 4 с маслом 3 и 5 с маслом 4, и
получаем следующие данные:

Тип масла			
1	2	3	4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85		70
	77		

В этом примере количество образцов
равно 4 (k = 4) со следующими соответствующими
размерностями:

n1 = 5 ,

n2 = 6 ,

n3 = 4 ,

n4 = 5 .

Количество наблюдений равно:

H = 5 + 6 + 4 + 5 = 20 .

Мыклассифицируем наблюдения в порядке возрастания
и дайте имранг от 1 до 20 с учетом
учитывать средние ранги. Мыполучаем следующие
Нижняя таблица с рангом наблюдений
вскобках и в конце каждого образца
сумма рангов, присвоенных соответствующей
выборке :

Machine Translated by Google

Тип масла			
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

Мы рассчитываем:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

$$= \frac{12}{20(20+1)} \cdot \left(\frac{372}{5} + \frac{982}{6} + \frac{53{,}52}{4} + \frac{21.52}{5} \right) - 3(20+1)$$

$$= 13{,}64 .$$

Если мы не сделаем корректив, чтобы принять во внимание средние ранги, получаем:

$$H = \frac{\sum_{i=1}^{g} \left(t_i^3 - t_i \right)}{1 - \frac{N^3 - N}{}}$$

$$= \frac{13{,}64}{1 - \frac{8 - 2 + 8 - 2 + 8 - 2}{203 \quad 20}}$$

$$= 13{,}67 .$$

Мы видим, что разница между H и H минимальна.

Гипотезы следующие:

H0: Нет никакой разницы между четырьмя маслами.

H1: Хотя бы одно из масел отличается от других.

Правило принятия решения следующее: Отклонить нулевую гипотезу H0 на уровне значимости α, если

$$H > \chi^2_{k-1,1-\alpha},$$

где $\chi^2_{k-1,1-\alpha}$ это значение из квадратной таблицы на уровне $1-\alpha$ и $k-1 = 3$ степени свободы.

Если мы выберем $\alpha = 0{,}05$, то значение $\chi^2_{3,0,95}$ равно 7,81, а если H больше $\chi^2_{3,0,95}$ (13,64 > 7,81), то гипотезу H0 отвергаем.

Если H0 отклоняется, мы можем использовать процедуру множественных сравнений, чтобы увидеть, какие пары масла разные.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Непараметрический тест
- Тест Уилкоксона

ССЫЛКА

Kruskal, WH, Wallis, WA: Использование званий в однокритериальном дисперсионном анализе. Варенье. Стат. доц. 47, 583–621 и опечатки, там же. 48, 907–911 (1952)

Machine Translated by Google

Л1 оценка

Метод оценки L1 является альтернативой методу наименьших квадратов, используемому в линейной регрессии для оценки параметров модели. Вместо минимизации остаточной суммы квадратов минимизируется сумма остаточных абсолютных значений.

Преимущество оценок L1 заключается в том, что они не так чувствительны к выбросам, как метод наименьших квадратов.

ИСТОРИЯ

Использование метода оценки L1 в очень простой форме восходит к Галилею в 1632 году, который использовал его для определения расстояния между Землей и недавно открытой звездой.

Boscovich, RJ (1757) предложил два критерия для настройки линии:

- Соответствующие суммы положительных и отрицательных остатков должны быть равны, и оптимальная линия будет проходить через центр тяжести наблюдений.
- Сумма остатков по абсолютной величине должен быть минимальным.

Бошкович Р. Дж предложил геометрический метод решения уравнений, выте кающих из его критериев. Аналитическая процедура, используемая для оценки параметров двух критериев Бошкович а, Р. Дж, принадлежит Лапласу, П. С. (1793). Э джворт, штат Флорида (1887) исключили первое условие, заявив, что сумма остатков должна быть равна нулю, и разработали метод, который можно использовать для обработки случая множественной линейной регрессии.

В первой половине 20-го века статистиков не очень интересовала линейная регрессия, основанная на оценке L1. Основными причинами были:

- Численные трудности из-за отсутствие закрытой формулы.
- Отсутствие асимптотической теории для оценки L1 в регрессионной модели.
- Оценка методом наименьших квадратов удовлетворительна для малых выборок, и нет уверенности в том, что оценка L1 лучше.

Начиная с 1970 г. область L1- оценки стала полюсом исследований множества статистиков, было опубликовано множество работ. Упомянем Basset, G. и Koenker, R. (1978), которые разработали асимптотически несмещенную оценку L1. В 1987 г. в Невшателе (Швейцария) состоялся конгресс, посвященный этим проблемам: Первая международная конференция по статистическим данным.

Анализ на основе L1-нормы и родственных методов.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В одномерном (1D) случае метод оценки L1 состоит в оценке

Machine Translated by Google

параметр β0 в модели

$$y_i = \beta_0 + \varepsilon_i,$$

так что сумма ошибок по модулю минимальна:

$$\min_{\beta_0} \sum_{i=1}^{n} |\varepsilon_i| = \min_{\beta_0} \sum_{i=1}^{n} |y_i - \beta_0|.$$

Эта задача сводится к нахождению β_0 так, чтобы $\sum_{i=1}^{n} |y_i - \beta_0|$ ism минимальна. Значение β_0, которое минимизирует эту сумму, это медиана 1D выборка $y_1, y_2, \ldots, y_n$.

В простой модели линейной регрессии

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i,$$

параметры β_0 и β_1, оцениваемые методом L1-оценки, обычно рассчитываются по итерационному алгоритму, используемому для минимизации

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 \cdot x_i|.$$

Используя методы линейного программирования, сегодня можно относительно просто вычислить оценки L1, особенно в модели множественной линейной регрессии.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Проблема оценки L1 заключается в отсутствии общей формулы для оценки параметров, в отличие от ситуации с методом наименьших квадратов. Для набора одномерных данных мерой центральной тенденции, определяемой оценкой L1, является медиана, которая по определению представляет собой значение, минимизирующее сумму абсолютных отклонений.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Ошибка
- Оценка
- Оценщик
- Наименьших квадратов
- Среднее абсолютное отклонение
- Параметр
- Остаток
- Надежная оценка

ССЫЛКИ

Basset, G., Koenker, R.: Асимптотическая теория наименьшей абсолютной регрессии ошибки. Варенье. Стат. доц. 73, 618–622 (1978)

Бошкович, Р. Дж: De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex instanceria etiam sensorum impress sa. Bononiensi Scientiarium et Artium Instituto Atque Academia Commentarii, vol. IV, стр. 353–396 (1757 г.)

Додж, Ю (ред.): Статистический анализ данных на основе L1-нормы и родственных методов. Эльзевир, Амстердам (1987)

Эджворт Ф.Ю: Новый метод сокращения наблюдений, относящихся к нескольким величинам. Фил ос.маг. (5-я серия) 24, 222–223 (1887 г.)

Множитель Лагранжа

Метод множителя Лагранжа — классический метод оптимизации, позволяющий определить локальные экстремумы функции при определенных ограничениях. Он назван в честь итало-французского математика и астронома Жозефа-Луи Лагранжа.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть f(x, y) — целевая функция, которую нужно максимизировать или минимизировать при условии

Machine Translated by Google

g(x, y) = 0. Составим вспомогательную функцию—лагранжиан:

$$F(x, y, \theta) = f(x, y) \quad \theta g(x, y),$$

где θ—неизвестная переменная, называемая Множитель Лагранжа. Мы ищем минимумы или максимумы функции F, приравнивая нулю частные производные по x, y и θ:

$$\frac{F}{x} = \frac{f}{f} \frac{g - \text{знак}}{g - \text{знак}} \frac{\text{равно } 0}{x} \quad x$$

$$\frac{F}{y} = \frac{\text{равно } 0}{} \quad y \quad y$$

$$\frac{}{\theta} \quad F = g(x, y) = 0.$$

Эти три уравнения могут быть решены относительно x, y и θ. Решение обеспечивает точки, в которых функция со связями имеет минимум или максимум. Мы должны решить, есть ли максимум или минимум.

Мы получим максимум, если

$$\frac{2F}{x2} \cdot \frac{2F}{y2} \quad \frac{2F}{x \ y}^2 > 0,$$

$$\frac{2F}{x2} < 0 \text{ и } \frac{2F}{y2} < 0.$$

Мы получаем минимум, если

$$\frac{2F}{x2} \cdot \frac{2F}{y2} \quad \frac{2F}{x \ y}^2 > 0,$$

$$\frac{x2}{x2} \ 2F > 0 \text{ и } \frac{2F}{y2} > 0.$$

Когда

$$\frac{2F}{x2} \cdot \frac{2F}{y2} \quad \frac{2F}{x \ y}^2 \quad 0,$$

тест не работает, и чтобы определить, существует ли минимум или максимум, мы должны исследовать функции, соседние по (x, y).

Метод множителей Лагранжа распространяется на функции от n переменных f (x1, x2,...,xn), на которые наложено k ограничений gj (x1, x2,...,xn) = 0, j = 1, 2,...,k, где k n. Точно так же в двумерном случае мы определяем лагранжиан:

$$F = f(x1, x2,...,xn) \quad \sum_{j=1}^{k} \theta j \cdot gj (x1, x2,...,xn),$$

с параметрами θ1,..., θk—множителями Лагранжа. Мы устанавливаем частные производные, которые приводят к n + k уравнениям с n + k неизвестными параметрами. Они приводят к координатам экстремумов функции (минимум или максимум), если они существуют.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Невозможно применить метод множителей Лагранжа, когда функции не дифференцируемы или когда налагаемые ограничения являются неравенствами. В этом случае мы можем использовать другие методы оптимизации, правильно выбранные из линейного программирования.

ПРИМЕРЫ

Позволять

$$f(x, y) = 5x2 + 6y2 - x \cdot y$$

быть целевой функцией и

$$г(x, y) = x + 2y \quad 24 = 0$$

наложенное ограничение. Формируем лагранжиан

$$F(x, y, \theta) = f(x, y) \quad \theta g(x, y) = 5x2 + 6y2 \quad x \cdot y \quad \theta \cdot (x + 2y \quad 24),$$

и обнуляем частные производные

$$\frac{F}{x} = 10x - y - \theta = 0, \quad (1)$$

Machine Translated by Google

$$\frac{F}{} = 12y - x - 2\theta = 0 \quad y \quad F \quad , \qquad (2)$$

$$\frac{}{\theta} \text{ знак равно Икс} + 2y - 24 \text{ знак равно } 0 . \qquad (3)$$

Удалив θ из (1) и (2):

$$20x - 2y - 2\theta = 0$$

$$x + 12y \quad 2\theta = 0$$

$$21x - 14y = 0$$

$$3x - 2y = 0$$

и заменив 2y на 3x в (3):

$$0 = x + 2y - 24 = x + 3x - 24$$

$$= 4x - 24$$

мы получаем x = 6. Заменив это значение x в (3), получаем y = 9. Критическая точка имеет координаты (6, 9). Мы вычисляем частные производные второго порядка, чтобы проверить, максимум или минимум. Мы получаем:

$$\frac{2F}{x2} = 10 \quad , \quad \frac{2F}{} = 12 \quad , \quad \frac{2F}{x\ y} = \quad 1 \quad ,$$

$$\frac{2F}{x2} \cdot \frac{y2}{y2} \quad 2F \quad \frac{2F}{x\ y}^{2}$$

$$= 10 \cdot 12 - (-1) \qquad 2 = 119 > 0 \quad ,$$

$$\frac{2F}{x2} > 0 \quad , \quad \frac{2F}{y2} > 0 .$$

Отметим, что точка (6, 9) соответствует минимум для целевой функции f при указанных выше ограничениях. Этот минимум дает f (6, 9) = 180 + 486 - 54 = 612.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Линейное программирование
- Оптимизация

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

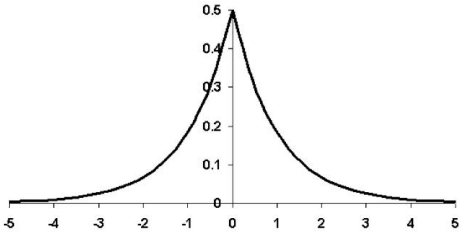
Додж, Ю (2004) Применение оптимизации. Спрингер, Париж

Распределение Лапласа

Случайная величина X следует закону Лапласа распределения, если его функция плотности следующую форму:

$$\phi(x) = 2 \quad \frac{1}{} \cdot \text{опыт} \quad \frac{|x \quad \theta|}{} \quad > 0 \quad ,$$

где – параметр дисперсии, а θ является ожидаемым значением.



Распределение Лапласа, θ = 0, = 1

Распределение Лапласа симметрично вокруг своего ожидаемого значения θ, которое также является мода и медиана распределения.

Распределение Лапласа является непрерывным распределение вероятностей.

ИСТОРИЯ

Это непрерывное распределение вероятностей носит имя своего автора. В 1774 г. Лаплас, П.С. написал фундаментальную статью о симметричных распределениях с целью Описание погрешностей измерения.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение случайной величины X после распределения Лапласа дано к:

$$E[X] = \theta$$

а дисперсия равна:

$$\text{Вар}(X) = 2 \cdot \quad ^{2} .$$

Machine Translated by Google

Стандартная форма функции плотности распределения Лапласа записывается:

$$f(x) = \sigma \frac{1}{2} \cdot \exp\left(-\frac{|x-\mu|}{\sigma}\right), \quad \sigma > 0,$$

с

$$\mu = \theta \text{ и } \sigma = 2 \cdot _____ .$$

Это позволяет написать распределение заданная функция распределения Лапласа к:

$$\Phi(x) = \begin{cases} \frac{1}{2} \text{ опыт } \frac{x-\theta}{_____} & \text{если } x \\ 1 - \frac{1}{2} \text{ опыт } \theta \frac{x}{____} \theta & \text{если } x \end{cases}$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Непрерывное распределение вероятностей

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Лаплас, PS de: Mémoire sur la probabi lite des прич ины par les événements. Мем. акад. Рой. наук. (представлено разными учеными) 6, 621–656 (1774) (или Лаплас, PS de (1891) Œuvres complètes, том 8. Gauthier-Villars, Париж, стр. 27–65).

Лаплас, Пьер Симон Де

Родился во Франции в 1749 году в семье среднего класса. был маркиз Пьер Симон де Лаплас. один из пионеров статистики. Интерес к математике, теоретической астрономии, Вероятность и статистика, его первые публикации появились в начале 1770-х гг. Он стал членом Академии наук г. 1785 г., а также руководил «Долготами». Бюро. В 1799 году он был назначен министром внутренних дел Бонапартом, который в 1806 г. присвоил ему титул графа. Избранный во Французскую академию в 1816 г., он преследовал свою научную работу до самой смерти в Париже, в 1827 году.

Некоторые основные работы и статьи де Лаплас, PS:

1774 Mémoire sur la probabilité des caus es par les événements. Oeuvres com plètes, Vol. 8, стр. 27–65 (1891). Готье-Виллар, Париж.

1781 Mémoire sur les probabilités. Oeu vres complètes, Vol. 9 стр. 383–485. (1891 г.). Готье-Виллар, Париж.

1805 Traité de Mécanique Céleste. Том. 1–4. Дюпро, Париж. Том. 5. Башелье, Париж.

1812 г. Аналитическая теория вероятностей. Курсье, Париж.

1878–1912 Полные произведения. 14 томов. Готье-Виллар, Париж.

Индекс Ласпейреса

Индекс Ласпейреса является составным индексом. число цен, построенное методом взвешенной суммы. Этот индекс представляет собой отношение суммы цен в фактического периода n к сумме цен в базисном периоде 0, причем эти суммы взвешиваются соответствующими количествами базисного периода. Поэтому номер индекса измеряет относительное изменение цены товаров, при этом соответствующие количества считаются неизменными. Индекс Ласпейреса отличается от индекса Паше выбором

Л

Machine Translated by Google

веса: в индексе Пааше взвешивание производится по количеству проданных товаров за текущий период Qn, тогда как в индексе Ласпейреса индекс, это делается по количеству, проданному в отчетный период Q0.

ИСТОРИЯ

Ласпейрес, Этьен, немецкий экономист. статистик французского происхождения, переформулированный номер индекса, разработанный Лоу, который станет известен как индекс Ласпея. в середине 19 века. В течение многих лет 1864–1871 Ласпейрес, над которым работал Этьен. товары и их цены в Гамбурге (он исключенные услуги).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Индекс Ласпейреса рассчитывается следующим образом:

$$B/0 = \frac{Pn \cdot Q0}{P0 \cdot Q0},$$

где Pn и Qn — цены и количества продано в текущем периоде, а P0 и Q0 равны цены и объемы, проданные по ссылке период, выраженный в основе 100:

$$B/0 = \frac{Pn \cdot Q0}{P0 \cdot Q0} \cdot 100 .$$

Сумма относится к рассматриваемому товару. Модель Ласпейреса также может быть использована для рассчитать индекс количества (также называемый индексом объема). В данном случае речь идет о ценах. константы, а величины переменные:

$$B/0 = \frac{Qn \cdot P0}{Q0 \cdot P0} \cdot 100 .$$

ПРИМЕРЫ

Рассмотрим следующую таблицу, указывающую соответствующие цены трех потребительских товаров в отчетный год 0 и в текущем году n, как а также количества, проданные в ссылке год:

Количество проданного товара в 1970 году ($q_{1970}0$) (тысячи)		Цена (евро) 1970 (p_0)	1988 (p_n)
Молоко	50,5	0,20	1,20
Хлеб	42,8	0,15	1,10
Масло	15,5	0,50	2,00

Из следующей таблицы имеем

$$PnQ0 = 138,68 \text{ и } P0Q0 = 24,27 .$$

Продукт	PnK_0	$p_{0_{1970}}0$
Молоко	60,60	10.10
Хлеб	47.08	6,42
Масло	31.00	7,75
Общий	138,68	24.27

Затем мы можем найти индекс Ласпейреса:

$$B/0 = \frac{Pn \cdot Q0}{P0 \cdot Q0} \cdot 100$$

$$\text{"="} \frac{138,68}{24.27} \cdot 100 = 571,4 .$$

Другими словами, согласно Ласпейресу индекс, индекс цен на эти товары имеет увеличился на 471,4% (571,4 100) за период рассматриваемый период.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Составной порядковый номер
- Индекс Фишера
- Индекс
- Индекс Пааше

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Ласпейрес, Э.: Гамбургер Warenpreise 1850–1863 гг. и смерть Калифорнии, Австралии lischen Geldentdeckung seit 1848. Jahrb. Натл. Стат. 3, 81–118, 209–236 (1864 г.)

Machine Translated by Google

Ласпейрес, Э .: Die Berechnung einer mit tleren Waarenpreissteigerung. Ярб. Натл. Стат. 16, 296–314 (1871 г.)

Латинская площадь

Смотрите латинские квадратные узоры.

Латинские квадратные конструкции

Схема латинского квадрата представляет собой расположение t вариантов обработки, каждая из которых повторяется t раз, таким образом, что каждая обработка появляется ровно один раз в каждой строке и каждом столбце схемы. Мы обозначаем римскими буквами методы лечения. Поэтому дизайн называется латинским квадратом. Этот тип дизайна используется для уменьшения систематической ошибки из-за строк (обработок) и столбцов.

ИСТОРИЯ

Согласно Preece (1983), история латинского квадрата восходит к 1624 году. В статистике Fisher, Ronald Aylmer (1925) представили дизайн латинского квадрата.

ПРИМЕРЫ

Фермер имеет в собственности два поля, на которых он хочет выращивать кукурузу.

Он хочет провести эксперимент с четырьмя разными видами кукурузы (обработка). Кроме того, он знает, что его поля не получают то же солнце и влажность. Есть два источника систематических ошибок, учитывая, что солнечный свет и влажность влияют на выращивание кукурузы. Фермер связывает с каждым небольшим участком земли определенный вид кукурузы таким образом, чтобы каждый из них появлялся один раз в каждой строке и столбце

ммм:

A: зерно 1

Б: зерно 2

С: зерно 3

Д: зерно 4 ·

На следующем рисунке показан латинский квадрат 4 × 4, в котором латинские буквы A, B, C и D представляют собой четыре лечения, описанные выше:

ABCD
BCDA
CDAB
DABC

Дисперсионный анализ покажет, существуют ли после исключения эффектов линий (солнце) и столбца (влажность) значительная разница между обработк

Л

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Дисперсионный анализ
План экспериментов

ССЫЛКИ

Фишер, Р.А.: Статистические методы для научных работников. Оливер и Бойд, Эдинбург (1925)

Preece, DA: Латинские квадраты, латинские кубы, латинские прямоугольники и т. д. В: Kotz, S., John son, NL (eds.) Encyclopedia of Statistical Sciences, vol. 4. Уайли, Нью-Йорк (1983)

Закон больших чисел

Закон больших чисел гласит, что среднее значение суммы независимых и одинаково распределенных случайных величин сходится к ожидаемому значению их распределения, когда объем выборки стремится к бесконечности.

Machine Translated by Google

ИСТОРИЯ

Закон больших чисел был впервые установлен Дж. Бернулли в его работе под названием Ars Conjectandi, опубликованной в 1713 г., в рамках эмпирического определения вероятности. Он заявил, что относительная частота события сходится при повторении одинаковых задач (распределение Бернулли) к числу, которое состоит из его вероятности.

Неравенство Бьенеме–Чебышева было первоначально открыто Бьенеме, Дж. (1853 г.), прежде чем несколько лет спустя Чебышев, П.Л. (1867 г.), открыл его совершенно независимо.

Общая версия закона больших чисел приписывается русским математикам.

Ян Хинчин, А. (1894–1959).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Сначала приведем неравенство Бьенеме–Чебышева: Рассмотрим X как случайную величину с математическим ожиданием μ и дисперсией σ2, обе конечными. Для каждого реального ε > 0 существует:

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2},$$

это означает, что вероятность того, что случайная величина X отклонится от своего ожидаемого значения на величину ε своего ожидаемого значения, не может быть больше, чем σ2/ε2. Важность этого неравенства заключается в том, что оно позволяет ограничить величину сходимость наблюдаемого среднего значения случайной величины к ее ожидаемому значению. Рассмотрим n независимых и одинаково распределенных случайных величин X1, X2,...,Xn с математическим ожиданием μ и дисперсией σ2; тогда для каждого ε > 0:

$$\text{п} \quad \frac{X1 + X2 + \ldots + Xn}{\text{н}} \quad \mu > \varepsilon \quad n \quad 0.$$

определенные вероятности в случаях, когда известно только математическое ожидание распределения, в конечном счете его дисперсия. Очевидно, что если само распределение известно, то нет необходимости использовать эти пределы, так как точное значение этих вероятностей можно вычислить. С помощью этого неравенства можно вычислить

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Закон больших чисел подчеркивает тот факт, что сохраняется вероятность, даже если она очень мала, существенного отклонения между эмпирическим средним значением ряда событий и его ожидаемым значением. Именно по этой причине используется термин слабый закон больших чисел. На самом деле можно показать, что это не так и что сходимость является квазиверной. Этот результат, более сильный, чем предыдущий, называется усиленным законом больших чисел и записывается:

$$\text{Плим} \quad X^{-} = \mu = 1 \quad .$$
$$\text{п}$$

ПРИМЕРЫ

Закон больших чисел утверждает, что всегда возможно найти такое значение n, при котором вероятность того, что X^{-} попадает в интервал μ ± ε, сколь угодно велика.
Приведем пример распределения вероятностей с дисперсией σ2 = 1. Выбрав интервал ε = 0,5 и вероятность 0,05, можно записать:

$$\text{п Икс}^{-} - \mu \quad 0{,}5 \text{ знак равно } 0{,}05 .$$

По неравенству Чебышева мы знаем, что

$$P\,X^{-} \quad \mu \quad \varepsilon \quad n\varepsilon 2 \quad \frac{\sigma^2}{} .$$

Таким образом, мы можем установить следующее неравенство:

$$0{,}05 \quad \frac{1}{n \cdot 0{,}25},$$

Machine Translated by Google

из которого:

$$0{,}05 \cdot 0{,}25 \quad 1/n,$$

$$n \quad 80.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Центральная предельная теорема

Теорема сходимости

(Теорема конвергенции)

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бернулли, Дж: ArsConjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Базель (1713 г.)

Bienaymé, IJ: Соображения по поводу аппуи де ла découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. Comptes Rendus de l'Académie des Sciences, Paris 37, 5–13 (1853 г.); переиздано в 1867 г. в журнале Liou ville précédant la preuve des Inégalités de Bienaymé-Cebyshev, Journal de Math ématiques Pures et Appliquées 12, 158–176

Чебышев, П.Л.: Des valeurs moyennes. Дж Матем. Чистый Appl. 12 (2), 177–184 (1867). Publié simultanément en Russe, dans Mat. Сборник 2(2), 1–9

Наименее абсолютный Регрессия отклонения

Метод наименьших абсолютных отклонений (метод LAD) является одной из основных альтернатив методам наименьших квадратов при оценке параметров регрессии.

Цель регрессии LAD состоит в том, чтобы обеспечить надежную оценку.

ИСТОРИЯ

Регрессия с наименьшим абсолютным отклонением была введена Роджером Джозефом Бошковичем примерно за 50 лет до метода наименьших квадратов, в 1757 году. Он использовал эту процедуру, пытаясь согласовать неполе довательные измерения, которые использовались для оценки формы Земли. Пьер Симон де Лаплас применил этот метод 30 лет спустя, но он оставался в тени малейшего

метод квадратов, разработанный Адрианом Мари Лежандром и Карлом Фридрихом Гауссом. Простота исчисления метода наименьших квадратов сделала метод наименьших квадратов гораздо более популярным, чем метод LAD. Тем не менее, в последние годы и с развитием статистических вычислений метод LAD можно легко использовать.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Мы рассматриваем здесь только простую регрессию LAD. случаи.

Рассмотрим простую регрессионную модель: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Метод LAD реализует следующий критерий: выбрать β_0 и β_1 так, чтобы они минимизировали сумму остаточных абсолютных значений, т. е.

$$|e_i| = y_i \quad \beta_0 \quad \beta_1 x_i,$$

где e_i обозначает i-й остаток. Особенность этого метода заключается в том, что в нем нет явной формулы для вычисления β_0 и β_1.

Оценки LAD могут быть вычислены путем применения итеративного алгоритма.

Простой алгоритм LAD

Важным моментом, используемым в алгоритме LAD, является тот факт, что линия регрессии всегда пересекает как минимум две точки данных.

Пусть (x1, y1) будет выбранной нами точкой данных. Ищем наилучшую прямую, проходящую через эту точку. Эта линия пересекает по крайней мере еще одну точку (x2, y2).

Machine Translated by Google

Следующим шагом является поиск наилучшей строки с помощью относительно суммы |ei|, которая пересекает этот время (x2, y2). Эта линия пересекает сейчас по крайней мере еще одна точка (x3, y3). Мы итеративно продолжаем и ищем лучшую линию, которая пересекает

(x3, y3) и так далее. Во время итераций, количество |ei| уменьшается. Если за итерацию получ енная линия идентична предыдущей, то можно сделать вывод, что это

лучшая линия.

Строительство наилучшего пересечения линии ИКС1, у 1) (с точки зрения | эй

Поскольку эта линия пересекает другую точку данных, мы должен найти точку (xk, yk), для которой

линия

$$y(x) = y1 + xk - \frac{yk - y1}{x1}(x - x1),$$

с уклоном

$$\beta 1 = \frac{yk - y1}{xk - x1}$$

и перехватить

$$\beta 0 = y1 \quad \beta 1x1,$$

является лучшим с точки зрения |ei|. Найти

В этой точке мы переименовываем (n 1) кандидата точки (x2, y2),...,(xn, yn) таким образом что

$$\frac{y2 - y1}{x2 - x1} \quad \frac{y3 - y1}{x3 - x1} \quad \ldots \quad \frac{yn - y1}{xn - x1}.$$

Определим $T = \sum_{i=1}^{n} |xi \quad x1|$. Обыскиваемый point(xk, yk) теперь определяется индексом к для которого

$$|x2 - x1| + \ldots + |xk \quad 1 \quad x1| < \frac{T}{2}$$

$$|x2 - x1| + \ldots + |xk \quad 1 \quad x1| + |xk \quad x1| > \frac{T}{2}.$$

Это условие гарантирует, что β1 минимизирует количество

$$\sum_{i=1}^{n} (yi \quad y1) \quad \beta 1 (xi \quad x1)$$

аналогично |ei| для линий регрессии проходящей через (x1, y1). β0 вычисляется таким образом, чтобы линия регрессии пересекала (x1, y1). Мы также можем проверить, что он проходит через точку данных (xk, yk). у нас просто есть переименовать точку (xk, yk) на (x2, y2) и перезапуск.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Могут возникнуть два частных случая, когда один запускает регрессию LAD: неуникальный решение и вырождение. Неуникальный решение означает, что существует как минимум две строки минимизация суммы остаточных абсолютных значений. Вырождение означает, что LAD линия регрессии проходит более двух точки данных.

ПРИМЕРЫ

Возьмем пример с участием 14 северных и страны Центральной Америки.

таблица содержит данные по каждой стране, рождаемость (количество рождений на 1000 человек) вместе с уровнем урбанизации (процент населения, проживающего в городах с более 100 000 жителей) в 1980 г.

Таблица: Уровень урбанизации и рождаемость

Обс. я	Страна	Урбанизация ставка xi	Рождение ставка йи
1	Канада	55,0	16.2
2	Коста-Рика	27,3	30,5
3	Куба	33,3	16,9
4	США	56,5	16,0
5	Сальвадор	11,5	40,2

Machine Translated by Google

Обс.	я	Страна	Урбанизация ставка	xi	Рождение ставка	йи
6		Гватемала	14,2		38,4	
7		Гаити	13,9		41,3	
8		Гондурас	19,0		43,9	
9		Ямайка	33.1		28,3	
10		Мексика	43,2		33,9	
11		Никарагуа	28,5		44,2	
12		Тринидад/ Тобаго	6,8		24,6	
13		Панама	37,7		28,0	
14		Доминиканский Республика	37,1		33.1	

Источник: Биркес и Додж(1993 г.).

Мы пытаемся оценить уровень рождаемости yi как функцию уровня урбанизации xi, используя LAD регрессия, которая заключается в построении LAD регрессионная модель

$$yi = \beta 0 + \beta 1xi + \varepsilon i .$$

Чтобы использовать алгоритм LAD для оценки β0 и β1, на первом шаге мы должны найти линия, пересекающая хотя бы одну точку, например, Канада (выбирается случайным образом). У нас есть (x1, y1) = (55,0, 16,2). Мы вычисляем склоны для всех 13 оставшихся стран. Например, для Коста-Рики получен результат $\frac{30,5 - 16,2}{27,3 - 55,0} = -0,5162$. 13 округ попытки теперь переупорядочены в порядке возрастания в соответствии с вычисленными наклонами, как может быть видно в таблице ниже. Точка (x2, y2) —Мексика, точка (x3, y3) —Никарагуа, и так далее. Далее мы вычисляем количество $T = \sum |xi - 55,0| = 355,9$ и $\frac{T}{2} = 177,95$. Это связано с определением страна k , удовлетворяющая условию

$$|x2 - x1| + \ldots + |xk - 1 - x1| < \frac{T}{2},$$

$$|x2 - x1| + \ldots + |xk - 1 - x1| + |xk - x1| > \frac{T}{2}.$$

С

$$\sum_{j=2}^{8} xj - x1 = 172,5 < \frac{T}{2}$$

и

$$\sum_{j=2}^{9} xj - x1 = 216,0 > \frac{T}{2},$$

можно сделать вывод, что k = 9 и что лучшая линия регрессии проходит через Канаду (x1, y1) и Сальвадор (x9, y9). Следовательно:

$$\beta 1 = \frac{y9 - y1}{x9 - x1} = \frac{40,2 - 16,2}{11,5 - 55,0} = -0,552 ,$$

$$\beta 0 = y1 - \beta 1x1 = 16,2 - (-0,552) \cdot 55,0 = 46,54 ,$$

а линия регрессии LAD:

$$y(x) = 46,54 - 0,552x .$$

Таблица: Расчет для нахождения линии регрессии LAD

Обс.	я	Страна	$\frac{\text{йи} - y1}{\text{хи} - x1}$	|xi - x1|		
2		Мексика	1,5000	11,8		11,8
3		Никарагуа	1,0566	26,5		38,3
4		Домин. Респ.	0,9441	17,9		56,2
5		Гондурас	0,7694	36,0		92,2
6		Панама	0,6821	17,3		109,5
7		Гаити	0,6107	41,1		150,6
8		Ямайка	0,5525	21,9		172,5
9		Сальвадор	0,5517	43,5		216,0
10		Гватемала	0,5447	40,8		256,8
11		Коста-Рика	0,5162	27,7		284,5
12		Тринидад/ Тобаго	0,1743	48,2		332,7
13	США		0,1333	1,5		334,2
14		Куба	0,0323	21,7		355,9

Чтобы улучшить нашу первую оценку, следующий шаг состоит в нахождении наилучшей линии, пересекающей El

Machine Translated by Google

Сальвадор. Мы переименовываем эту точку (x1, y1) = (11,5, 40,2) и упорядочиваем 13 стран в порядке возрастания в соответствии с их наклонами yi 11,5. $\overline{xi}$ 40,2 , и аналогично получаем наилучшую линию регрессии, проходящую как через Сальвадор, так и через США.

Следовательно,

$$= = -0{,}538 \frac{40{,}2 - 16{,}0\ \beta 1}{11{,}5 - 56{,}5} ,$$

$$\beta 0 = 40{,}2 - (-0{,}538) \cdot 11{,}5 = 46{,}38 ,$$

а линия регрессии

$$y(x) = 46{,}38 - 0{,}538x .$$

Третий шаг заключается в поиске наилучшей линии, проходящей через Соединенные Штаты. Поскольку мы находим (аналогично предыдущему способу), что эта линия регрессии проходит через Сальвадор, наш алгоритм останавливается, и линия регрессии y(x) = 46,38 0,538x является линией регрессии LAD нашей задачи.

Заметим здесь, что линия, рассчитанная с использованием метода наименьших квадратов для этих данных, имеет вид

$$y(x) = 42{,}99 - 0{,}399x .$$

Обратите внимание, что с этой линией регрессии Куба, Никарагуа и Тринидад/Тобаго имеют аномально большие остаточные значения. Удалив эти три точки и пересчитав линию регрессии методом наименьших квадратов, мы получим

$$y(x) = 48{,}52 - 0{,}528x ,$$

что очень близко к линии регрессии LAD.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Наименьших квадратов
- Множественная линейная регрессия
- Регрессивный анализ
- Остаток
- Надежная оценка
- Простая линейная регрессия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бирке с Д., ДоджЮ: Альтернативные методы регрессии. Уайли, НьюЙорк (1993)

Наименее значимый Тест на разницу

Критерий наименьшего значимого различия (LSD) используется в контексте дисперсионного анализа, когда F-отношение предполагает отклонение нулевой гипотезы H0, то есть когда разница между средними значениями популяции значительна.

Этот тест помогает выявить популяции, средние значения которых статистически различаются. Основная идея теста состоит в сравнении популяций, взятых попарно. Затем он используется для проведения однофакторного или двустороннего дисперсионного анализа, учитывая, что нулевая гипотеза уже отвергнута.

ИСТОРИЯ

Тест на ЛСД был разработан Фишером и Рональдом Эйлмером (1935), которые хотели узнать, какое лечение оказало значительное влияние на дисперсионный анализ.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если в дисперсионном анализе F-отношение приводит к отклонению нулевой гипотезы H0, мы можем выполнить тест LSD, чтобы определить, какие средства привели к отклонению H0 .
Тест состоит в попарном сравнении средств. В общих чертах стандартное отклонение разности между

среднее значение группы i и среднее значение группы j,

Machine Translated by Google

для i = j равно:

$$c^2_я \left(\frac{1}{ни} + \frac{1}{ндж}\right),$$

где s2 —оценка дисперсии _
внутри групп

$$2_с = \frac{SCI}{N-к} = \frac{\sum_{я=1}^{к} \sum_{Дж=1}^{ни} Yij - Y¯i.^2}{N-к},$$

где SCI —сумма квадратов внутри групп, N —общее количество наблюдений, k —количество групп, Yij —j -е наблюдение группы i, Y¯i. —среднее значение группы i, ni —количество наблюдений в группе i, а nj —количество наблюдений в группе j.

Соотношение

$$\frac{Y¯i. - Y¯j.}{c^2_я \left(\frac{1}{ни} + \frac{1}{ндж}\right)}$$

следует распределению Стьюдента с N k
степени свободы. Разница между парой средних
значительна, когда

$$Y¯ Дж - s2 Y¯i_я \cdot tN - k,1 - \alpha_2 \sqrt{\left(\frac{1}{ни} + \frac{1}{ндж}\right)},$$

где tN k,1 α обозначает значение переменной
Стьюдента с N k степенями свободы для уровня
значимости , установленного на α.
Величина s2 I($\sqrt{\left(\frac{1}{ни} + \frac{1}{ндж}\right)} \cdot tN - k,1 - \alpha_2$ ----
называется ЛСД.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Как и дисперсионный анализ, тест LSD требует
независимых и одинаково распределенных
нормальных переменных с постоянной вариацией.
анс.

ПРИМЕРЫ

Во время варки цельномасляные круассаны впитывают
жир в переменных количествах. Мы хотим
проверить, зависит ли поглощенное количество
от типа содержания жира (животный или растительный
жир). Мы готовим для нашего эксперимента четыре
вида жира и варим по шесть круассанов на сливочном
масле для каждого типа круассанов.
Вот результаты нашего эксперимента:

Тип жира			
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

Мы проводим односторонний дисперсионный анализ,
который позволяет нам проверить нулевую гипотезу.

$$H0 : \mu1 = \mu2 = \mu3 = \mu4,$$

который явно отвергнут.
Поскольку H0 отвергается, мы запускаем тест LSD,
чтобы определить, какое средство вызвало
отклонение H0.
В этом примере количество наблюдений одинаково
для каждой группы (n1 = n2 = n3 =
n4 = n = 6). Следовательно, s2 $\sqrt{\left(\frac{1}{ни} + \frac{1}{ндж}\right)}$
упрощается до $\sqrt{\frac{2c^2_я}{н}}$.
Значение для соответствующего теста LSD равно:

$$\sqrt{\frac{2c^2_я}{н}} \cdot tN - k,1 - \alpha_2 \cdot$$

Результаты дисперсионного анализа
обеспечивают вычислительные элементы, необходимые для

Л

Machine Translated by Google

найти значение $s^2_я$:

$$s^2_я = \frac{SCI}{N-k} = \frac{\sum_{я=1}^{ни} \sum_{Дж=1}^{к} (Yij - \bar{Y}i.)^2}{N-k} = \frac{2018}{20} = 100,9 .$$

Если мы выберем уровень значимости α = 0,05, значение Стьюдента для N k = 20 степеней свободы будет равно:

$$tN\ k,1\ \alpha/2 = t20,0,975 = 2,086 .$$

Следовательно, значение статистики LSD равно:

$$LSD = \sqrt{\frac{2c^2_я}{н}} \cdot t20,0,975 = \sqrt{\frac{2 \cdot 100,9}{6}} \cdot 2,086 = 12,0976 .$$

Разница между парой средних $\bar{Y}i$ и $\bar{Y}j$ существенна, если

$$\bar{Y}i. - \bar{Y}j. \geq 12,0976 .$$

Мы можем обобщить результаты в таблице con получение средних разностей, значений для LSD-статистики и заключение.

Разница	LSD значимый
$\bar{Y}_{.} - \bar{Y}_{2.} = \lvert 72 - 85 \rvert = 13$	12,0976 Да 1.
$\bar{Y}_{.} - \bar{Y}_{3.} = \lvert 72 - 76 \rvert = 4$	12,0976 № 1.
$\bar{Y}_{4.} - \bar{Y}_{1.} = \lvert 72 - 62 \rvert = 10$	12,0976 Нет
$\bar{Y}_{.} - \bar{Y}_{3.} = \lvert 85 - 76 \rvert = 9$	12,0976 № 2.
$\bar{Y}_{2.} - \bar{Y}_{4.} = \lvert 85 - 62 \rvert = 23$	12,0976 Да
$\bar{Y}_{3.} - \bar{Y}_{4.} = \lvert 76 - 62 \rvert = 14$	12,0976 Да

Убедимся, что только разности $\bar{Y}_{1.}$ $\bar{Y}_{2.}$, $\bar{Y}_{2.}$ $\bar{Y}_{4.}$ и $\bar{Y}_{3.}$ $\bar{Y}_{4.}$ являются значительными.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Студенческое распределение
- Ученический стол

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Додж Ю, Томас, Д.Р.: О производительности процедур множественного сравнения не параметрической и нормальной теории.
Санкхья В42, 11–27 (1980)

Фишер, Р.А.: План экспериментов.
Оливер и Бойд, Эдинбург (1935)

Миллер, Р.Г., младший: Одновременный статистический вывод, 2-е изд. Спрингер, Берлин, Хайдельберг, НьюЙорк (1981)

Наименьших квадратов

См. метод наименьших квадратов.

Метод наименьших квадратов

Метод наименьших квадратов заключается в минимизации суммы квадратов невязок. Последние соответствуют квадратам отклонений между расчетными и наблюдаемыми значениями.

ИСТОРИЯ

Метод наименьших квадратов был введен Лежандром, Адрианом-Мари (1805 г.).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим множественную линейную регрессию модель:

$$Yi = \beta 0 + \sum_{j=1}^{p-1} \beta j \cdot Xji + \varepsilon i, \quad я = 1,\ldots,n,$$

где Yi — зависимая переменная, Xji, $j = 1,\ldots,p$ 1 — объясняющие или независимые переменные, βj, $j = 0,\ldots,p$ 1 —

Machine Translated by Google

оцениваемые параметры, а εi – срок ошибки .

Оценка параметров методом наименьших квадратов заключается в минимизации суммы квадратов ошибок (невязок):

$$\text{минимум } f(\beta 0,\ldots,\beta p\ \ 1),$$

где :

$$f(\beta 0,\ldots,\beta p\ \ 1) = \sum_{i=1}^{n} \varepsilon_i^2 .$$

Остаточная сумма квадратов равна:

$$f(\beta 0,\ldots,\beta p\ \ 1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Yi \quad \beta 0 \quad \sum_{j=1}^{p-1} \beta j \cdot Xji \right)^2 .$$

Находим оцениваемые параметры функции, приравнивая нулю частные производные этой функции по каждому параметру этих:

$$\frac{f(\beta 0,\ldots,\beta p\ \ 1)}{\beta 0} = 0 ,$$

$$\frac{f(\beta 0,\ldots,\beta p\ \ 1\ \beta 1)}{} = 0 ,$$

$$\vdots$$

$$\frac{f(\beta 0,\ldots,\beta p\ \ 1)}{\beta p\ \ 1} = 0 .$$

Предположим, у нас есть n наблюдений и что p = 2:

$$(X1, Y1),(X2, Y2),\ldots,(Xn, Yn) .$$

Уравнение , связывающее Yi и Xi, можно определить из линейной модели:

$$Yi = \beta 0 + \beta 1 \cdot Xi + \varepsilon i , \qquad i = 1,\ldots,n.$$

Сумма квадратов отклонений от оценочной линии регрессии равна:

$$f(\beta 0, \beta 1) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} 2 (yi \quad \beta 0 \quad \beta 1 \cdot Xi) \quad \cdot$$

Математически определим β0 и β1 , взяв частные производные функции f по параметрам β0 и β1 и приравняв их нулю

Частная производная по β0:

$$\frac{f}{\beta 0} = 2 \cdot \sum_{i=1}^{n} (Yi \quad \beta 0 \quad \beta 1 \cdot Xi) .$$

Частная производная по β1:

$$\frac{f}{\beta 1} = 2 \cdot \sum_{i=1}^{n} Xi \cdot (yi \quad \beta 0 \quad \beta 1 \cdot Xi) .$$

Следовательно, оценочные значения для β0 и β1, обозначенные как β^ 0 и β^ 1, даны как решение следующих уравнений:

$$\sum_{i=1}^{n} Yi \quad \hat{\beta} 0 \quad \hat{\beta} 1 \cdot Xi = 0 ,$$

$$\sum_{i=1}^{n} Xi \cdot Yi \quad \hat{\beta} 0 \quad \hat{\beta} 1 \cdot Xi \text{ знак равно } 0 .$$

Развивая эти два уравнения, мы получаем:

$$Yi = n \cdot \hat{\beta} 0 + \hat{\beta} 1 \cdot \sum_{i=1}^{n} Си$$

и

$$\sum_{i=1}^{n} (Xi \cdot Yi) = \hat{\beta} 0 \cdot \sum_{i=1}^{n} Xi + \hat{\beta} 1 \cdot \sum_{i=1}^{n} X2_i ,$$

которые мы называем нормальными уравнениями линии регрессии наименьших квадратов.

Л

Machine Translated by Google

Данный

$$X^{-} = \frac{\sum_{i=1}^{n} Ci}{n} \quad \text{и} \quad Y^{-} = \frac{\sum_{i=1}^{n} И}{n},$$

решение нормальных уравнений дает оценки следующих параметров:

$$\hat{\beta}0 = Y^{-} \quad \hat{\beta}1 \cdot X^{-},$$

$$\hat{\beta}1 = \frac{n \cdot \sum_{i=1}^{n}(Ci \cdot Ии) - \sum_{i=1}^{n} Ci \cdot \sum_{i=1}^{n} И}{n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} Ci\right)^2}$$

$$= \frac{\sum_{i=1}^{n} Xi - X^{-} Yi - Y^{-}}{\sum_{i=1}^{n} Kси - X^{-}{}^2}.$$

Оценки методом наименьших квадратов для зависимой переменной:

$$\hat{Y}i = \hat{\beta}0 + \hat{\beta}1 \cdot Xi, \qquad i = 1,\ldots,n$$

или

$$\hat{Y}i = Y^{-} + \hat{\beta}1 \cdot Xi \quad X^{-}, \qquad i = 1,\ldots,n.$$

Согласно этому уравнению линия регрессии МНК проходит через точку $^{-}Y^{-}$ икс, , который называется барицентром или центром тяжести рассеянного облака точек данных.

Точно так же мы можем выразить модель множественной линейной регрессии в терминах векторов и матриц:

$$Y = X \cdot \beta + \varepsilon,$$

где Y — $(n \times 1)$ вектор отклика (зависимая переменная), X — $(n \times p)$ матрица независимых переменных, ε — $(n \times 1)$ вектор члена ошибки, β — $(p \times 1)$ вектор оцениваемых параметров.

Применение метода наименьших квадратов состоит в минимизации следующего уравнения:

$$\varepsilon \cdot \varepsilon = (Y \quad X \cdot \beta) \cdot (Y \quad X \cdot \beta)$$

$$= Y \cdot Y \quad \beta \cdot X \cdot Y \quad Y \cdot X \cdot \beta + \beta \cdot X \cdot X \cdot \beta$$

знак равно $Y \cdot Y - 2 \cdot \beta \cdot X \cdot Y + \beta \cdot X \cdot X \cdot \beta$.

Нормальные уравнения задаются в матричной форме следующим образом:

$$X \cdot X \cdot \hat{\beta} = X \cdot Y.$$

Если $X \cdot X$ обратимо, параметры β оцениваются по формуле:

$$\hat{\beta} = X \cdot X \quad {}^{1} \cdot X \cdot У.$$

ПРИМЕРЫ

См. простую линейную регрессию

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Оценка
- Множественная линейная регрессия
- Регрессивный анализ
- Простая линейная регрессия

ССЫЛКИ

Лежандр, А. М.: Новые методы определения орбит комет.

Курсье, Париж (1805 г.)

Лежандр, Адриан Мари

Родившийся в 1752 году в Париже, французский математик Лежандр, Адриен Мари сменил де Лапласа на посту профессора математики сначала в Военной школе, а затем в Высшей военной школе. Нормальная школа Парижа. Он был заинтересован Э д в теории и практике астрономии и геодезии. С 1792 года входил в состав

Machine Translated by Google

Французская комиссия, ответственная за измерение длины квадранта меридиана от Северного полюса до экватора, проходящего через Париж.

Он умер в 1833 г., но в истории статистики его имя осталось связанным с публикацией его работы в 1805 г., включавшей приложение под названием «О методе наименьших квадратов». Эта публикация была в корне полемики, которая противопоставляла его Гауссу, CF, который соперничал с ним из-за признания открытия этого метода.

Основные работы Лежандра, А.М.:

1805 г. Новые методы определения орбит комет.

Курсье, Париж.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Плакетт, Р.Л.: Исследования по истории вероятностей и статистики. В: Кендалл М., Плакетт Р.Л. (ред.) Открытие метода наименьших квадратов. об. II. Гриффин, Лондон (1977)

Леманн, Эрих

Леманн, Эрих родился в Страсбурге, Франция, в 1917 году и уехал в Соединенные Штаты в 1940 году.

вклад в теориюстатистики. Его научные интересы включают статистическую теорию, историю и философиюстатистики. Он является автором книг «Основные понятия вероятности и статистики», «Элементыконечной вероятности», «Непараметрические параметры: статистические методы, основанные на рангах», «Проверка статистических гипотез» и «Теория точечной оценки». Он является членом Американской академии министра искусств и наук и Национальной академии наук (НАН), будучи избран в НАН в 1978 г. Он был редактор Annals of Mathematical Statis тики и президент Института математики математическая статистика.

Избранные произведения и публикации Эриха Леманн:

1983 Теория точечной оценки, 2-е изд. Уайли, НьюЙорк.

1986 Проверка статистических гипотез. 2-е изд. В: Серия по теории вероятностей и математической статистике. Уайли, НьюЙорк.

1993 (совместно с Ходжесом, Дж Л.) Проверка статистических гипотез. Чепмен и Холл, НьюЙорк.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Оценка

Проверка гипотезы

Уровень значимости

См. уровень значимости.

Точка рычага

В регрессионном анализе мы называем точкой рычага наблюдение i, для которого на оценочное значение отклика $\hat{Y}_i$ влияет значение соответствующей независимой переменной X_i. Понятие точки рычага эквивалентно понятиюрычага в физике. То есть даже небольшая модификация значения наблюдения i может серьезно повлиять на оценку y_i.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Диагональные элементы матрицы H в регрессионном анализе отражают влияние наблюдения на оценку

Machine Translated by Google

ответ. Мы в принципе связываем диагональные элементы hii матрицы H с понятием точки рычага. Поскольку след матрицы H равен числу оцениваемых параметров, мы ожидаем, что hii будет близка к p/n. hii должен быть близок к p/n, если все наблюдения имеют одинаковое влияние на сметы. Если для наблюдения i мы получить hii больше, чем p/n, то наблюдение i считается точкой воздействия. На практике наблюдение i называется точкой высокого рычага, если:

$$\text{привет} > \frac{2p}{n}.$$

Рассмотрим следующую модель линейной регрессии в матричной записи:

$$Y = X \cdot \beta + \varepsilon ,$$

где Y —(n × 1) вектор отклика (зависимая переменная), X —(n × p) матрица независимых переменных, ε —(n × 1) вектор членов ошибки, β —(p ×1) вектор оцениваемых параметров.

Оценка $\hat{\beta}$ вектора β определяется как:

$$\hat{\beta} = XX^{-1}XY .$$

Матрица H определяется:

$$H = XXX^{-1}_{\text{Икс}} .$$

В частности, диагональные элементы hii определяются следующим образом:

$$\text{хи} = \text{хх}\,XX^{-1}_{\text{х я}},$$

где xi —i-я строка X.

Мы называем точкой рычага каждое наблюдение i, для которого:

$$\text{хи} = \text{хх}\,XX^{-1} > 2 \cdot \text{х я}\ \frac{p}{n} .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Huber (1981) ссылается на точки рычага, когда диагональный элемент матрицы $XXX^{-1}X$ больше 0,2, причем критический предел не зависит ни от числа параметров p, ни от числа наблюдений n.

ПРИМЕРЫ

Рассмотрим следующую таблицу, где Y — зависимая переменная, объясняемая независимой переменной X:

Икс	Д
50	6
52	8
55	9
75	7
57	8
58	10

Модель линейной регрессии в матричной записи:

$$Y = X \cdot \beta + \varepsilon ,$$

где

$$Y = \begin{matrix} 6 \\ 8 \\ 9 \\ 7 \\ 8 \\ 10 \end{matrix}, \quad X = \begin{matrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 75 \\ 1 & 57 \\ 1 & 58 \end{matrix},$$

ε —вектор невязки (6 × 1), а β —вектор параметров (2 × 1).

Мы получаем матрицу H:

$$H = XXX^{-1}_{\text{Икс}} .$$

Мы нашли:

$$H = \begin{matrix} 0{,}32 & 0{,}28 & 0{,}22 & -0{,}17 & 0{,}18 & 0{,}16 \\ 0{,}28 & 0{,}25 & 0{,}21 & -0{,}08 & 0{,}18 & 0{,}16 \\ 0{,}22 & 0{,}21 & 0{,}19 & 0{,}04 & 0{,}17 & 0{,}17 \\ 0{,}17 & 0{,}08 & 0{,}04 & 0{,}90 & 0{,}13 & 0{,}17 \\ 0{,}18 & 0{,}18 & 0{,}17 & 0{,}13 & 0{,}17 & 0{,}17 \\ 0{,}16 & 0{,}16 & 0{,}17 & 0{,}17 & 0{,}17 & 0{,}17 \end{matrix} .$$

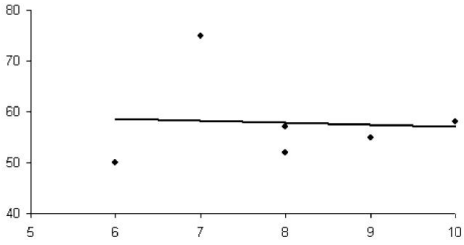
Machine Translated by Google

Запишем диагональ матрицы H на векторе h:

$$h = \begin{matrix} 0{,}32 \\ 0{,}25 \\ 0{,}19 \\ 0{,}90 \\ 0{,}17 \\ 0{,}17 \end{matrix}.$$

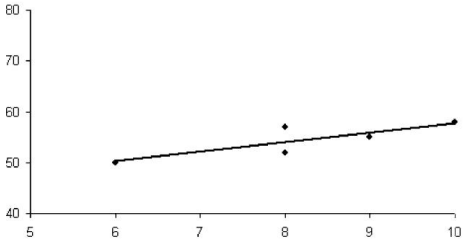
В этом случае имеется два параметра (p = 2) и шесть наблюдений (n = 6), т. е. критическое значение равно $\frac{4}{6} = 0{,}67$.

Сравнивая компоненты h с этим значением, замечаем, что, за исключением наблюдения i = 4, все наблюдения имеют hii < 0,67. Четвертое наблюдение является точкой опоры для изучаемой проблемы. Следующий рисунок, где переменные X и Y представлены на двух осях, иллюстрирует тот факт, что точка (75, 7) является точкой высокого рычага.



Изучая этот рисунок, мы замечаем, что линия регрессии находится под сильным влиянием этой точки рычага и не отражает правильно набор изучаемых данных.

На следующем рисунке мы удалили эту точку из набора данных. Мы видим, что новая линия регрессии дает гораздо лучшее представление набора данных.



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Матрица шляпы
- Матрица
- Множественная линейная регрессия
- Регрессивный анализ
- Простая линейная регрессия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Хубер, П.Дж: Надежная статистика. Уайли, НьюЙорк (1981)

Мостеллер, Ф., Тьюки, Дж.В.: Анализ данных и регрессия: второй курс статистики. Аддисон-Уэсли, Рединг, Массачусетс (1977)

Л

Тест отношения правдоподобия

Проверка отношения правдоподобия является проверкой гипотезы. Он позволяет проверять общие гипотезы, касающиеся интересующих параметров параметрического семейства, а также тестировать две разные модели, построенные на одних и тех же данных. Гл

Идея этого теста заключается в следующем: вычислить вероятность наблюдения данных при нулевой гипотезе H0 и при альтернативной гипотезе с использованием функции правдоподобия.

ИСТОРИЯ

Это Нейман, Ежи и Пирсон, Эгон Шарп (1928), который предложил использовать статистику отношения правдоподобия для проверки

Machine Translated by Google

гипотезы. Вальд, Абрахам (1941) обобщил тест отношения правдоподобия на более

пликативные гипотезы. Асимптотические результаты на распределение этой статистики (с учетом определенные условия регулярности) были впервые представлены Wilks, Samuel Stanley (1962).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В тесте отношения правдоподобия нулевая гипотеза отклоняется, если правдоподобие при альтернативная гипотеза значительно больше правдоподобия при нулевой гипотезе.
Следовательно, проблема не в том, чтобы найти наиболее адекватный параметр, но зная, если при альтернативной гипотезе мы получаем значительно лучшую оценку.
Теорема Неймана – Пирсона утверждает, что для всех значимых уровней α вероятность тест отношения (для проверки простой гипотезы) имеет большую силу, чем любой другой тест.
Пусть x1, x2,...,xn —n независимых наблюдений (n > 1) случайной величины
X после распределения вероятностей с параметр θ (или вектор параметров θ). Мы обозначим через L (θ;x1, x2,...,xn) вероятность функция, которая при фиксированном xi зависит только от θ, и мы обозначаем через пространство всех возможных значений θ и 0, где 1 являются подпространством из .

Мы различаем:

Простая нулевая гипотеза против простой альтернативной гипотезы

$$H0 : \theta = \theta0$$

по сравнению с $H1 : \theta$ знак равно $\theta1$.

В этом случае отношение правдоподобия равно:

$$\lambda = \lambda (x1, x2,...,xn)$$

$$"=" \frac{L (\theta0;x1, x2,...,xn)}{L (\theta1;x1, x2,...,xn)} .$$

Отношение правдоподобия для уровня значимости α определяется решающим правилом:

Отклонить H0, если $\lambda > \lambda0$,

где $\lambda0$ —критическое значение, определяемое уровень значимости (α) и распределение λ при нулевой гипотезе:

$$\alpha = P (\lambda (x1, x2,...,xn) \quad \lambda0 | H0) .$$

Проблема здесь в том, чтобы найти критическую региона , так как общепринятого процедура для этого. Идея состоит в том, чтобы искать статистика, распределение которой известно.

Простая нулевая гипотеза по сравнению с общей альтернативой

$$H0 : \theta = \theta0$$

по сравнению с H1 : θ 1 (например, θ<θ0) .

Мы позируем:

$$\lambda = \frac{\sup\theta \quad 1 \, L (\theta0;x1, x2,...,xn)}{L (\theta1;x1, x2,...,xn)} ,$$

где supθ 1 L(θ; x1, x2,...,xn) представляет отправляет значение функции правдоподобия для параметра, который максимизирует эту подобную функцию надежности в ограниченном параметре пространства (θ 1), и мы снова ищем распределение λ при H0.
Если альтернативная гипотеза H1 есть θ = θ0, мы можем записать λ как:

$$\frac{L \hat{\theta};x1, x2,...,xn}{L (\theta0;x1, x2,...,xn)} ,$$

где $\hat{\theta}$ – оценка максимального правдоподобия для θ.

Общая гипотеза 0 : θ 0 против альтернативной гипотезы

ч АС1 : θ \0 = {θ ; θ / 0}

Рассмотрим следующую статистику :

$$\lambda = \frac{\sup\theta \quad 0 \, L (\theta;x1, x2,...,xn)}{\sup\theta \quad L (\theta;x1, x2,...,xn)}$$

Machine Translated by Google

и поиск области отторжения, как и в предыдущих случаях.

$\sup_{\theta \in \Theta_0} L(\theta; x_1, x_2, \ldots, x_n)$ представляет значение функции правдоподобия для параметр, максимизирующий функцию правдоподобия (оценку максимального правдоподобия) в ограниченном пространстве параметров. $\sup_{\theta \in \Theta} L(\theta; x_1, x_2, \ldots, x_n)$ представляет то же самое, но для всего пространства параметров. Чем больше разница между этими двумя значениями, тем более неподходящей будет ограниченная модель (и H0 будет отклонена). При H0 распределение $-2 \log(\lambda)$ асимптотически следует распределению хи-квадрат с u степенями свободы, где $u = \dim \Theta - \dim \Theta_0$ (то есть число параметров, определяемое H0). Другими словами:

$$\lim_{n \to \infty} \log \frac{\sup_{\theta \in \Theta} L(\theta; x_1, x_2, \ldots, x_n)}{\sup_{\theta \in \Theta_0} L(\theta; x_1, x_2, \ldots, x_n)} = \chi^2(u).$$

Для частных случаев, рассмотренных ранее, результат будет следующим:

$-2 \log \lambda \sim \chi^2$ на 1 степени свободы если $n \to \infty$.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Если распределение данных нормальное, критерий отношения равноправия эквивалентен критерию Фишера. Преимущество последнего (когда его можно построить в явном виде) состоит в том, что он гарантирует наилучшую производительность и мощность [согласно теореме Неймана-Пирсона] для проверки простой гипотезы.

ПРИМЕРЫ

Пример 1

Пусть X — нормальная случайная величина со средним μ и дисперсии σ^2. Мы хотим протестировать

Гипотеза мынания:

$$H_0 : \mu = 0 \quad , \quad \sigma^2 > 0$$

$$H_1 : \mu = 0 \quad , \quad \sigma^2 > 0 .$$

Рассматривая пространство параметров $\Theta = \{(\mu, \sigma^2); -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$ и подпространство $\Theta_1 = \{(\mu, \sigma^2) : \mu = 0\}$, гипотеза теперь может быть записана как:

$$H_0 : \theta = \mu, \quad \sigma^2 \in \Theta_0$$

$$H_1 : \theta = \mu, \quad \sigma^2 \in \Theta \setminus \Theta_0 .$$

Рассмотрим теперь $x_1, \ldots, x_n$ n наблюдений переменной X ($n > 1$). Функция правдоподобия (см. пример для максимально похожей вероятности):

$$L(\theta; x_1, \ldots, x_n) = L(\mu, \sigma^2; x_1, \ldots, x_n)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \cdot \exp\left(\frac{\sum_{i=1}^{n} 2(x_i - \mu)}{2\sigma^2}\right) .$$

Мы оцениваем выборочное среднее и дисперсию, используя оценки максимального правдоподобия:

$$\mu = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \bar{x},$$

$$\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 .$$

Следовательно, функции правдоподобия для нулевой гипотезы:

$$L(\Theta_0, \sigma^2; x_1, \ldots, x_n)$$

$$= \left(\frac{n}{2\pi \sum_{i=1}^{n} x_i^2}\right)^{n/2} \cdot \exp\left(\frac{n \sum_{i=1}^{n} x_i^2}{2 \sum_{i=1}^{n} x_i^2}\right)$$

$$= \left(\frac{n}{2\pi \sum_{i=1}^{n} x_i^2}\right)^{n/2} \cdot \exp\left(\frac{n}{2}\right)$$

Machine Translated by Google

$$= \frac{\text{п exp}(\quad 1)}{2\pi^{\text{н}}_{\text{я=1}} x^2_{\text{я}}}^{\text{н/2}},$$

и для альтернативной гипотезы:

$$L\,\mu,\sigma 2;\, x1,\ldots,x\text{п}$$

$$= \frac{\text{н}}{2\pi^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}^{\text{н/2}}$$

$$\cdot \text{опыт} \frac{\text{н}^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}{2^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}$$

$$= \frac{1}{2\pi^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}^{\text{н/2}} \cdot \text{опыт} \frac{\text{n}}{2}$$

$$= \frac{\text{п exp}(\quad 1)}{2\pi^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}^{\text{н/2}},$$

поэтому отношение правдоподобия равно:

$$\lambda = \lambda\,(x1,\, x2,\ldots,xn)$$

$$= \frac{L\,0,\sigma 2;\, x1,\, x2,\ldots,xn\ L}{\mu,\sigma 2;\, x1,\, x2,\ldots,x\text{п}}$$

$$= \frac{\text{п exp}(\quad 1)}{2\pi^{\text{н}}_{\text{я=1}} x^2_{\text{я}}}^{\text{н/2}}$$

$$\frac{\text{п exp}(\quad 1)}{2\pi^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}^{\text{н/2}}$$

$$= \frac{^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}{\text{п}^{\text{н}}_{\text{я=1}} x^2_{\text{я}}}^{\text{н/2}}.$$

В этой точке необходимо найти распределение $^{\text{н}}_{\text{я=1}} x^2_{\text{я}}$. Для этого напомним, что $^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2} + nx2$, и получаем:

$$\lambda = \frac{^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}{^{\text{н}}_{\text{я = 1}} (x - x)^{-\ 2} + nx2}^{\text{н/2}}$$

$$= 1 \quad 1 + \frac{nx2}{^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}^{\text{н/2}}.$$

Мы ищем $\lambda 0$, для которого нулевая гипотеза не отвергается, если $\lambda \quad \lambda 0$, что мы можем записать как:

$$\lambda \quad \lambda 0$$

$$\text{или } 1 \quad 1 + \frac{nx2}{^{\text{н}}_{\text{я=1}} 2\,(x - x)}^{\text{н/2}} \quad \lambda 0$$

$$\text{или } 1 + \frac{nx2}{^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}}^{\text{н/2}} \quad \lambda\ {}^{1}_{0}$$

$$\text{или } \frac{nx2}{^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2}} \quad \lambda\ {}^{2/n}_{0} \quad 1.$$

Умножая каждый член неравенства на $(n \quad 1)$, получаем:

$$\lambda \quad \lambda 0 \text{ или}$$

$$\frac{\overline{n\ |x|}}{^{\text{н}}_{\text{я=1}} (x - x)^{-\ 2} \quad (\text{п} - 1)}$$

$$(n \quad 1)\,\lambda \quad {}^{2/\text{п}}_{0} \quad 1.$$

Machine Translated by Google

Так как случайная величина

$$T = \frac{\bar{X} - \text{мю}}{\sigma(n - 1)},$$

где σ2 —выборочная дисперсия, соответствует t - распределению с (n 1) степенями свободы; мы можем эквивалентно написать:

$$T = \frac{\bar{n}\,\bar{X}}{\text{мю}\ n\sigma 2\ (n \quad 1)},$$

что для среднего значения, равного нулю дает:

$$T = \frac{\bar{n}\bar{X}}{n\sigma 2\ (n \quad 1)}$$

$$\text{"="} \frac{\bar{n}\bar{X}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 \quad (n - 1)}}.$$

Теперь нужно найти значение $\lambda 0$, для которого c := (n 1) λ 2/n $_0$ 1 – это значение переменной Стьюдента для выбранного уровня значимости α:

$$\lambda 0 = \frac{c2}{1\ n - 1} + \quad ^{n/2}.$$

Например, для размера выборки n = 6 и для α = 0,05 мы находим z5,0,05/2 = 2,751, поэтому достаточно увидеть, что:

$$\lambda \quad \frac{(2,751)^2}{5} + 1 \quad ^3 = 0,063$$

чтобы нулевая гипотеза была верной.

Пример 2

Две независимые случайные выборки X1,...,Xn и Y1,...,Ym (m + n > 2) подчиняются нормальным распределениям с параметрами μ1, σ2 и μ2, σ2 соответственно. Мы хотим проверить следующую гипотезу:

$$H0 : \mu 1 = \mu 2$$

$$H1 : \mu 1 = \mu 2 .$$

Определим множества:

$$= (\mu 1, \mu 2, \sigma 2); \qquad < \mu 1, \mu 2 < \quad , 0 <$$

$$\sigma 2 < \qquad ,$$

$$1 = (\mu 1, \mu 2, \sigma 2) \quad ; \mu 1 = \mu 2 .$$

Теперь мы можем записать гипотезу следующим образом:

$$H0 : \text{мк1, мк2,} \qquad \sigma 2 \quad 1$$

$$H1 : \text{мк1, мк2,} \qquad \sigma 2 \quad \backslash 1 .$$

Итак, функции правдоподобия:

$$L\ \mu 1, \mu 1, \sigma 2; x1,...,xn, y1,...,ym$$

$$\text{"="} \frac{1}{\sqrt{2\pi\sigma}}^{n+m} \cdot \text{опыт} \frac{\sum_{i=1}^{n}(xi \quad \mu 1)^2 + \sum_{i=1}^{m}(yi \quad \mu 1)^2}{2\sigma 2},$$

$$L\ \mu 1, \mu 2, \sigma 2; x1,...,xn, y1,...,ym$$

$$\text{"="} \frac{1}{\sqrt{2\pi\sigma 2}}^{n+m} \cdot \text{опыт} \frac{\sum_{i=1}^{n}(xi \quad \mu 1)^2 + \sum_{i=1}^{m}(yi - \mu 2)^2}{2\sigma 2}.$$

Используя оценки максимального правдоподобия, которые относятся к нулевой гипотезе:

$$\mu 1 = \frac{\sum_{i=1}^{n} xi + \sum_{i=1}^{m} yi}{n + m} \text{"="} \frac{n\bar{x} + m\bar{y}}{n + m},$$

$$\sigma 2 = \frac{\sum_{i=1}^{n}(xi \quad \mu 1)^2 + \sum_{i=1}^{m}(yi - \mu 1)^2}{n + m}$$

Machine Translated by Google

и для альтернативной гипотезы

$$\text{мк}\,1 = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x},$$

$$\text{м}2 = \frac{\sum_{i=1}^{m} y_i}{m} = \bar{y},$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x - x)^2 + \sum_{i=1}^{m} (y - y)^2}{n + m},$$

получаем отношение правдоподобия:

$$\lambda = \frac{L\,\mu 1, \mu 1, \sigma 2; x1, \ldots, xn, y1, \ldots, ym}{L\,\text{мк}\,1, \text{мк}\,2, \sigma^2; x1, \ldots, xn, y1, \ldots, ym}$$

$$= \frac{2\pi}{n+m}^{(n+m)/2} \text{опыт} \frac{n+m}{2} \cdot \sum_{i=1}^{n} x_i - \frac{n\bar{x} + m\bar{y}}{n+m}^2 + m \sum_{i=1}^{n} y_i - \frac{n\bar{x} + m\bar{y}}{n+m}^2 \cdot \frac{2\pi}{n+m}^{(n+m)/2} \text{расширение} \frac{n+m}{2}$$

$$= \sum_{i=1}^{n} (x - x)^2 + m \sum_{i=1}^{n} (y - y)^2 \cdot \sum_{i=1}^{n} x_i - \frac{n\bar{x} + m\bar{y}}{n+m}^2 + m \sum_{i=1}^{n} y_i - \frac{n\bar{x} + m\bar{y}}{n+m}^2 \cdot \sum_{i=1}^{n} (x - x)^2 + m \sum_{i=1}^{n} (y - y)^2$$

$$= \sum_{i=1}^{n} (x - x)^2 + m \sum_{i=1}^{n} (y - y)^2 \cdot \sum_{i=1}^{n} x_i - \frac{n\bar{x} + m\bar{y}}{n+m}^2 + m \sum_{i=1}^{n} y_i - \frac{n\bar{x} + m\bar{y}}{n+m}^2 \cdot \sum_{i=1}^{n} (x - x)^2 + m \sum_{i=1}^{n} (y - y)^2$$

$$= \sum_{i=1}^{n} x_i - \frac{n\bar{x} + m\bar{y}}{n+m}^2 + m \sum_{i=1}^{n} y_i - \frac{n\bar{x} + m\bar{y}}{n+m}^2 \cdot \sum_{i=1}^{n} (x - x)^2 + m \sum_{i=1}^{n} (y - y)^2$$

Статистика $\lambda^{2/(n+m)}$ определяется как:

$$\frac{\sum_{i=1}^{n} Kси - \bar{X}^2 + \sum_{i=1}^{m} Йи - \bar{Й}^2}{\sum_{i=1}^{n} Си - \frac{n\bar{X}+m\bar{Y}}{n+m}^2 + \sum_{i=1}^{m} Йи - \frac{n\bar{X}+m\bar{Y}}{n+m}^2}.$$

Чтобы найти распределение этой статистики, мы используем следующие аргументы (обратите внимание, что $\sum_{i=1}^{n} Xi - \bar{X} = 0$):

$$\sum_{i=1}^{n} Си - \frac{n\bar{X} + m\bar{Y}}{n+m}^2$$

$$= \sum_{i=1}^{n} Kси - \bar{X} + \bar{X} - \frac{n\bar{X} + m\bar{Y}}{n+m}^2$$

$$= \sum_{i=1}^{n} Kси - \bar{X}^2 + n\,\bar{X} - \frac{n\bar{X} + m\bar{Y}}{n+m}^2$$

$$= \sum_{i=1}^{n} Kси - \bar{X}^2 + \frac{нм2}{(n+m)} \bar{X} - Y^2.$$

То же самое верно и для Y:

$$\sum_{i=1}^{m} Йи - \frac{n\bar{X} + m\bar{Y}}{n+m}^2$$

$$= \sum_{i=1}^{m} Йи - \bar{Й}^2 + \frac{мин2}{2\,(n+m)} \bar{X} - Y^2.$$

Machine Translated by Google

Теперь мы можем написать:

$$\lambda^{2/(n+m)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2 + \frac{nm}{n+m}(\bar{X} - \bar{Y})^2}$$

$$= \frac{1}{1 + \dfrac{\frac{nm}{n+m}(\bar{X} - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}}.$$

Если нулевая гипотеза верна, то случайная величина

$$T = \frac{\sqrt{\frac{nm}{n+m}}(\bar{X} - \bar{Y})}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2}}}$$

следует распределению Стьюдента с $n+m-2$ степенями свободы. Таким образом, мы получаем:

$$\lambda^{2/(n+m)} = \frac{1}{1 + \dfrac{(n+m-2)T^2}{n+m}} = \frac{n+m-2}{n+m-2+T^2}.$$

Ищем λ_0, для которого нулевая гипотеза отвергается при $\lambda < \lambda_0$:

$$\lambda = \frac{n+m-2}{n+m-2+T^2}^{2/(n+m)} < \lambda_0$$

$$|T| > \sqrt{(n+m-2)\left(\lambda_0^{-\frac{2}{n+m}} - 1\right)} := c.$$

Теперь достаточно взять критическое значение c из таблицы случайной величины Стьюдента с фиксированным α (тот, который нам нужен) и n размером выборки. Например, при $n = 10$, $m = 6$, и $\alpha = 0{,}05$, находим:

$$c = z_{1-4{,}0{,}025} = 2{,}145,$$

$$\lambda_0 = \frac{c^2}{n+m-2}^{2/(n+m)} - 1$$

$$= \frac{(2{,}145)^2}{14}^{1/8} - 1 = -1{,}671.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Альтернативная гипотеза
- Тест Фишера
- Гипотеза
- Проверка гипотезы
- Максимальная вероятность
- Нулевая гипотеза

Л

ССЫЛКИ

Кокс, Д.Р., Хинкли, Д.В.: Теоретическая статистика. Чепмен и Холл, Лондон (1973)

Эдвардс, AWF: Вероятность. Отчет о статистической концепции вероятности и ее применении к научным выводам. Издательство Кембриджского университета, Кембридж (1972)

Кендалл, М.Г., Стюард, А.: Расширенная теория статистики, том. 2. Гриффин, Лондон (1967)

Нейман, Дж, Пирсон, Э.С.: Об использовании и интерпретации некоторых критериев тестирования для целей статистического вывода, Части I и II. Биометрика 20А, 175–240, 263–294 (1928).

Уолд, А.: Асимптотически самые мощные проверки статистических гипотез. Анна. Мат. Стат. 12, 1–19 (1941)

Уолд, А.: Некоторые примеры асимптотически наиболее мощных тестов. Анна. Мат. Стат. 12, 396–408 (1941).

Machine Translated by Google

Уилкс, С.С.: Математическая статистика. Уайли, Нью-Йорк (1962)

Линейный график

Линейная диаграмма — это особый тип частотного графика. Он полезен для представления частотного распределения дискретной случайной величины. Значения переменной даны по оси абсцисс, а частоты, соответствующие этим значениям, отображены по оси ординат.

ИСТОРИЯ

См. графическое представление.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

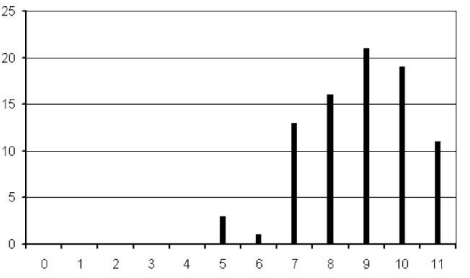
Линейный график имеет две оси. По оси абсцисс отложены значения xi дискретной случайной величины. Над каждым значением xi проводится линия длиной, равной частоте ni (или относительной частоте fi), соответствующей этому значению.

ПРИМЕРЫ

Распределение размеров обуви выборки из 84 мужчин приведено в следующей таблице.

Размер обуви	Частота
5	3
6	1
7	13
8	16
9	21
10	19
11	11
Общий	84

Вот линейная диаграмма, созданная из этого стол:



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Точечный график

Распределение частоты

Графическое представление

Гистограмма

Линейное программирование

Математическое программирование касается оптимального распределения ограниченных ресурсов для достижения желаемых целей. Исследуемая математическая модель линейного программирования выражается с помощью линейных соотношений. Задача линейного программирования появляется в виде набора линейных уравнений (или неравенств), называемых ограничениями, и линейной функции, устанавливающей цель, которая называется целевой функцией.

ИСТОРИЯ

Математическое программирование появилось в области экономики около 1930-х гг.

Замечательные работы по линейному программированию были опубликованы Джоном фон Нейманом (1935–1936). В 1947 году ДжорджБ. Данциг представил симплексный метод решения задач линейного программирования.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Задачу линейного программирования можно определить следующим образом.

Machine Translated by Google

Найдите набор значений для переменных x1, x2,...,xn, линейно максимизирующей (или минимизирующей) целевую функцию

$$z = c1 \cdot x1 + c2 \cdot x2 + ... + cn \cdot xn$$

при линейных ограничениях:

$$a11 \cdot x1 + a12 \cdot x2 + ... + a1n \cdot xn \text{ "=" } б1,$$

$$a21 \cdot x1 + a22 \cdot x2 + ... + a2n \cdot xn \text{ "=" } б2,$$

$$\vdots$$

$$am1 \cdot x1 + am2 \cdot x2 + ... + amn \cdot xn \text{ "=" } бм,$$

и к неотрицательным ограничениям:

$$xj \quad 0, j = 1, 2,...,n,$$

где aij, bi, cj —известные константы.
Другими словами, это означает, что в задаче линейного программирования мы ищем неотрицательное решение, удовлетворяющее поставленным ограничениям и оптимизирующее целевую функцию.
Мы можем написать задачу линейного программирования, используя следующую матричную запись:

Максимизируйте (или минимизируйте) $z = c \cdot x$

при условии $A \cdot x$ "=" б

и $x \quad 0$,

где c представляет собой линейный вектор (1 × n), x представляет собой вектор-столбец (n × 1), A представляет собой матрицу (m × n), b представляет собой вектор-столбец (m × 1), а 0 —нулевой вектор с n компонентами.
Основным инструментом решения задач линейного программирования является симплекс-метод. Он состоит из набора правил, которые должны быть следовал за решением поставленной задачи. Это итерационный метод, дающий точное решение за конечное число итерации.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Линейное программирование ищет оптимальное распределение этих ресурсов, позволяющее производить один или несколько товаров. Это делается путем максимизации (или минимизации) таких функций, как прибыль (или затраты).
Поэтому коэффициенты целевой функции часто называют «цены», связанные с переменными.

ПРИМЕРЫ

Фирма производит столы и стулья, используя два станка A и B. Каждое изделие производится на обоих станках. Для изготовления стула требуется 2 ч станка A и 1 ч станка B. Для изготовления стола требуется 1 ч станка A и 2 ч станка Б. Фирма получает прибыль в размере 3,00 швейцарских франков за каждый стул и 4,00 швейцарских франка за каждый стол. Две машины доступны более 12 часов в день.
Учитывая приведенные выше данные, мы ищем дневной график производства, гарантирующий максимальную прибыль. Другими словами, фирма хочет знать оптимальную производственную программу, то есть количество стульев и столов, которое необходимо производить в день, чтобы получить максимальную прибыль.
Обозначим через x1 количество стульев, а через x2 количество столов, которые необходимо производить в день. Ежедневная прибыль определяется следующей функцией:

г знак равно $3 \cdot x1 + 4 \cdot x2$.

Л

Machine Translated by Google

Эта целевая функция должна быть максимизирована. Мы знаем, что производство x1 требует 2 часа работы машине A, а для продукта x2 требуется 1 ч машине A. Следовательно, общее количество часов в день для машина A это:

$$2 \cdot x1 + 1 \cdot x2$$

так как производятся стулья x1 и столы x2 . Однако общее время машины A не может превышать 12 часов. В математических терминах это означает:

$$2x1 + x2 \quad 12 \quad .$$

Таким же образом для машины B находим:

$$x1 + 2x2 \quad 12 \quad .$$

У нас есть два ограничения. Кроме того, мы не может производить отрицательные значения стульев и таблицы, поэтому два дополнительных ограничения добавлен:

$$x1 \quad 0 \text{ и } x2 \quad 0 .$$

Мы хотим найти значения переменных x1 и x2 , которые удовлетворяют ограничениям и максимизируют прибыль. Наконец, мы можем представить математическую задачу следующим образом:

$$\text{макс } z = 3x1 + 4x2 ,$$

$$\text{ст } 2x1 + x2 \quad 12 \quad ,$$

$$x1 + 2x2 \quad 12 \quad ,$$

$$\text{и } x1, x2 \quad 0$$

(ст означает подлежащий). Максимальный профит дается парой (x1, x2) = (4, 4); 28.00 CHF прибыли от производства 4 стула и 4 стола. Для этой проблемы результат можно найти, вычислив прибыль для каждой возможной пары (их меньше, чем 36 из них.)

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Исследование операций
- Оптимизация

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Артханари, Т.С., Додж, Ю: Математическое программирование в статистике. Уайли, Нью-Йорк (1981)

Данциг, Великобритания: Приложения и расширения линейного программирования. Дун-од, Париж (1966)

Хэдли Г.: Линейное программирование. Эддисон Уэсли, Рединг, Массачусетс (1962)

Фон Нейман, Дж: Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. Эргебн. математика. Коллоку. Вена 8, 73–83 (1937)

Логарифмическое преобразование

Одно из самых полезных и распространенных преобразований – логарифмическое преобразование. Действительно, перед запуском линейной регрессия, возможно, было бы целесообразно заменить зависимая переменная Y ее логарифмом log (Y) или линейной комбинацией из последних a + b · log (Y) (a и b константы, скорректированные методом наименьших квадратов). Такая операция стабилизировала бы дисперсию Y и сделала бы распределение преобразованной переменной ближе к нормальному.

Логарифмическое преобразование также позволяют перейти от мультипликативного к аддитивная модель с более простыми оценками определить и проще интерпретировать. С другой стороны, логарифмическое преобразование гамма - случайная величина Y приводит к новому случайная величина, близкая к нормальной распределение. Последнее может быть легко адаптировано в регрессионном анализе.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Зависимая переменная
- Регрессивный анализ

Логистическая регрессия

При наличии бинарной зависимой переменной логистическая регрессия дает возможность подогнать регрессионную модель. Он позволяет изучать связь между пропорцией и набором объясняющих переменных, как количественных, так и качественных.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Модель логистической регрессии с зависимой переменной Y и независимыми переменными X1, X2,..., _p выражается как:

$$\text{логит}\,\pi(X_{ij})_{i=1,\ldots,p} = \ln\left[\frac{P(Y_j=1)}{P(Y_j=0)}\right] = \beta_0 + \sum_{i=1}^{p}\beta_i X_{ij},$$

$$j = 1,\ldots,n,$$

где Y_j ($j = 1,\ldots,n$) соответствует наблюдениям переменных Y, а X_{ij} наблюдениям независимых переменных, а π —функция вероятности (успеха) логистической регрессии:

$$\pi(X_{ij}) \overset{\text{"="}}{=} \frac{\exp\left(\beta_0 + \sum_{i=1}^{p}\beta_i X_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{p}\beta_i X_{ij}\right)}.$$

Эквивалентно, мы можем написать:

$$P(Y_j = 1) = 1 + \frac{\exp\left(\beta_0 + \sum_{i=1}^{p}\beta_i X_{ij}\right)}{\exp\left(\beta_0 + \sum_{i=1}^{p}\beta_i X_{ij}\right)},$$

что позволяет вычислить вероятность того, что переменная Y примет значение 1 в зависимости от реализации независимых переменных X1, X2,...,Xp.

Интересным свойством является то, что параметры модели ($\beta_0, \beta_1,\ldots, \beta_p$) могут принимать любые значения, так как все значения логит-преобразования $\pi(Y)\ \text{logit}(\pi(Y)) = \log\left[\frac{\pi(Y)}{1-\pi(Y)}\right]$ соответствуют числу от 0 до 1 для P (Y = 1).

Параметры модели можно оценить с помощью метода максимального правдоподобия, тем самым максимизируя следующую функцию

$$L(\beta_0, \beta_1,\ldots, \beta_n) = \prod_{j=1}^{n} \pi(X_{ij})_{i=1,\ldots,p}^{Y_i} \cdot \left[1 - \pi(X_{ij})_{i=1,\ldots,p}\right]^{1-Y_j}$$

или, проще говоря, путем максимизации логарифмической функции правдоподобия:

$$\log L(\beta_0, \beta_1,\ldots, \beta_n) \overset{\text{"="}}{=} \sum_{j=1}^{n} Y_j \log\left[\pi(X_{ij})_{i=1,\ldots,p}\right] + (1 - Y_j)\log\left[1 - \pi(X_{ij})_{i=1,\ldots,p}\right].$$

Поэтому мы можем провести проверку гипотез о значениях параметров (например, $\beta_0 = 0$, $\beta_3 = \beta_2$ и т. д.), используя тест отношения правдоподобия. Интерпретация коэффициентов в регрессии, соответствующих бинарным независимым переменным, описана в следующем абзаце. Когда независимая переменная X непрерывна, коэффициент регрессии отражает относительный риск увеличения X на одну единицу.

Л

независимая переменная X имеет более одного класса, мы должны преобразовать ее в бинарные переменные. Тогда интерпретация коэффициента ясна и понятна.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Логистическая регрессия часто используется в эпидемиологии, поскольку:

- Его сигмоидальная форма (приведенное ранее выражение $P(Y_j = 1)$ очень хорошо соответствует связь, обычно наблюдаемая между дозой X и частотой заболевания $Y: \frac{\exp(x)}{1+\exp(x)}$.
- Это легко интерпретировать: показатель связи между заболеванием и фактором риска M (соответствует двоичной переменной Xi) выражается отношением шансов. Это хорошее приближение относительного риска когда вероятности $P(Y = 1 \mid Xi = 1)$ и $P(Y = 1 \mid Xi = 0)$ малы, и это вычисляется просто как экспонента параметр, связанный с переменной Xi; таким образом:

  шансы—отношение фактора $M = \exp(\beta i)$.

ПРИМЕРЫ

Возьмем пример, данный Хосмером, и Лемешо. Зависимая переменная указывает на наличие или отсутствие заболеваний сердца у 100 человек, а независимая переменная - возраст.

В результате получается модель логистической регрессии:

$$\pi(X) = -5,3 + 0,11 \cdot X$$

со стандартным отклонением для константы и β (коэффициент независимой переменной X) равно 1,1337 и 0,241 соответственно.

Два оцениваемых параметра являются статистически значимыми.

Риск сердечно-сосудистых заболеваний увеличивается на $\exp(0,11) = 1,12$ с каждым годом.

Также можно вычислить достоверность интервал, взяв экспоненту интервал для β:

$$(\exp(1,11 - 1,96 \cdot 0,241), \exp(1,11 + 1,96 \cdot 0,241)) = (1,065, 1,171).$$

В следующей таблице приведены значения для логарифмическое правдоподобие, а также значения параметров, которые оцениваются путем максимизации логарифмической функции правдоподобия в итеративном алгоритме:

	Константа логарифмического правдоподобия		β
1	54,24	4,15	0,872
2	53,68	5,18	1,083
3	53,67	5,31	1.109
4	53,67	5,31	1.109

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Двоичные данные
- Таблица сопряженности
- Тест отношения правдоподобия
- Максимальная вероятность
- Модель
- Шансы и отношение шансов
- Регрессивный анализ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бишоп, Ю.М.М., Финберг, С.Е., Холланд, PW: Дискретный многомерный анализ: теория и практика. Массачусетский технологический институт Пресс, Кембридж, Массачусетс (1975)

Кокс, Д.Р., Снелл, Э.Дж: Анализ бинарных данных. Чепмен и Холл (1989)

Хосмер Д.В., Лемешоу С.: Прикладная логистическая регрессия. Серия Wiley в вероятности и Статистика. Уайли, Нью-Йорк (1989)

Альтман, Д.Г.: Практическая статистика для медицины альные исследования. Чепмен и Холл, Лондон (1991)

Machine Translated by Google

Логнормальное распределение

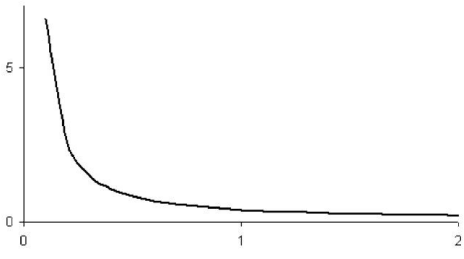
Случайная величина Y следует логнормальному распределению, если значения Y являются функцией значений X согласно уравнению

$$y = \exp(x),$$

где случайная величина X следует за или неправильное распределение.

Функция плотности логнормального распределения определяется как:

$$f(y) = \sigma y \frac{1}{\sqrt{2\pi}} \exp\left(2\sigma^2 \frac{(\log y - \mu)^2}{}\right).$$



Логнормальное распределение, μ = 0, σ = 1

Логнормальное распределение представляет собой непрерывное распределение вероятностей.

ИСТОРИЯ

Логнормальное распределение связано с работами Гальтона Ф. (1879) и Макалистера Д. (1879), которые получили выражения для среднего, медианы, моды, дисперсии и некоторых квантилей результирующего распределения. Гальтон Ф. перешел от n независимых положительных случайных величин и построил произведение и с помощью логарифма перешел от произведения к сумме новых случайных величин.

Kapteyn, JC (1903) пересмотрел конструкцию случайной величины после логнормального распределения с помощью van Uven, MJ (1916). Они разработали графический метод оценки параметров. С тех пор было опубликовано много работ, посвященных логнормальному распределению. Полный список приведен Джонсоном Н.Л. и Коцем С. (1970).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение логнормального распределения определяется выражением:

$$E[Y] = \exp\left(\mu + 2\sigma^2 \frac{1}{}\right).$$

Дисперсия равна:

$$\mathrm{Var}(X) = \exp(2\mu) \cdot \exp 2\sigma^2 \quad \exp\sigma^2 \cdot$$

Если случайная величина Y следует логнормальному распределению, случайная величина

$$X = \text{пер } Y$$

следует нормальному распределению

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В обычной статистической практике широко применяется логарифмически нормальное распределение. Например, критические пропорции в фармацевтике, продолжительность визита к врачу и распределение экономических сил подчиняются логарифмически нормальному распределению. Логнормальное распределение можно использовать как приближение к нормальному распределению. Когда σ становится маленьким, логнормальное распределение стремится к нормальному распределению.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Непрерывное распределение вероятностей
- Нормальное распределение

Л

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Гальтон, Ф.: Среднее геометрическое в естественной и социальной статистике. проц. Рой. соц. Лонд. 29, 365–367 (1879 г.)

Джонсон, Н.Л., Коц, С.: Распределения в статистике: непрерывное одномерное распределение буции, тт. 1 и 2. Уайли, НьюЙорк (1970).

Каптейн, Дж: Кривые перекоса частоты в биологии и статистике. Астрономическая лаборатория Нордхофф, Гронинген (1903 г.)

Каптейн, Дж, ван Увен, М.Дж: Кривые перекоса частоты в биологии и статистике. Братья Хойтсема, Гронинген (1916 г.)

Макалистер, Д.: Закон среднего геометрического. проц. Рой. соц. Лонд. 29, 367–376 (1879 г.)

Продольные данные

Лонгитюдные данные — это данные, собранные по субъектам в течение определенного времени. Это соответствует множественным наблюдениям за субъектом, поскольку одни и те же единицы измеряются более чем один раз. Повторные измерения в биологии и медицине, а также панельные исследования в социальных и экономических науках являются характерными примерами лонгитюдных исследований данных.

ИСТОРИЯ

См. историю данных панели.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Очень общий способ записи линейной модели для лонгитюдного анализа данных следующий:

$$yi = Xi\beta + Wiyi + \varepsilon i,$$

где yi — вектор ответа столбца ni × 1 для испытуемого i, Xi — план ni × b матрица, обычно состоящая из фиктивных закодированных переменных, β — вектор коэффициентов регрессии ni × b, Wi — матрица плана b × 1 для случайных эффектов yi, а εi — субъектная ошибка условия.

Приведенная выше общая модель представляет собой смешанную модель, поскольку она включает как фиксированные (β), так и случайные эффекты (yi).

Оценка параметров часто выполняется с помощью оценки максимального правдоподобия с функция правдоподобия сильно зависит от предположений о ковариационной структуре yi и εi. Несложно построить статистические тесты на предполагаемых параметров, для получения доверительных интервалов и для измерения прогноза. Методы выбора модели также используются для тестирования вложенных моделей.

ПРИМЕР

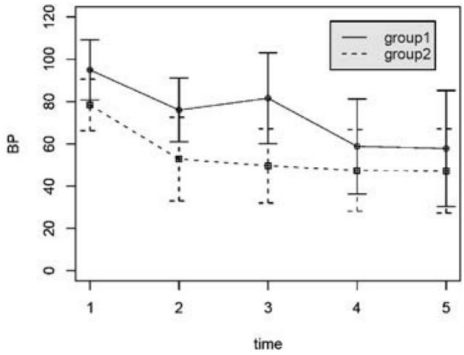
Две группы людей следуют двум разным методам лечения артериального давления в течение 5 месяцев.
Каждая группа состоит из пяти человек. Интересующие меры приведены в таблице ниже:

	Месяц (с января по май)
1-я группа	105 68 95 86 93
	100 92 98 74 35
	102 85 95 62 82
	70 81 72 35 39
	98 34 34 37 40
2 группа	83 25 31 42 38
	69 44 41 30 32
	84 64 77 80 82
	63 54 55 47 44
	93 77 44 38 40

Каждая строка соответствует наблюдениям, сделанным с течением времени для одного и того же человека.

Ниже приведен рисунок, на котором изображены кривые группового среднего ±1 стандартная ошибка.

Machine Translated by Google



Простейшим подходом в этом случае будет игнорировать случайные эффекты ($\gamma_i = 0$) и построить простую модель с фиксированным только эффект, то есть группа, в которой субъект принадлежит. В таком случае модель сводится к

$$y_{ij} = \beta_j + \varepsilon_{ij},$$

где j указывает группу и является простой фиктивной переменной. Если есть значительное разница между людьми, затем случайный эффект γ_i добавлен к модели выше.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Продольные данные в основном используются в эпидемиологии и медицинской статистике, а также в социально-экономические исследования. В бывшем кас.

Продольные данные представляют собой повторные измерения для различных групп лечения в клинические испытания. В последнем отметим панель исследования, которые обычно измеряют социальные и экологические экономические показатели, такие как доход, потребление и удовлетворенность работой, в разное время периоды.

Есть много возможных осложнений в лонгитюдный анализ данных. Во-первых, данные коррелированы, так как одни и те же предметы измеряются во времени. Это надо учитывать в конечном счете, чтобы обеспечить надлежащее Результаты. Более того, часто появляются недостающие данные. в таком анализе данных, так как не все предметы измеряются в течение всего времени период. Последнее становится более проблематичным, например, при анализе выживаемости и последующих исследованиях, когда испытуемые (обычно пациенты) умирают или прекращают испытания и отказываются от лечение.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Панель

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Twisk, JWR: Прикладные продольные данные Анализ для эпидемиологии. Практическое руководство. Издательство Кембриджского университета, Кембридж (2003)

Фицморис, Г., Лэрд, Н., Уэр, Дж: Прикладной Продольный анализ. Уайли, НьюЙорк (2004)

Моленбергс, Г., Герт, В.: Модели для дискретных продольных рядов данных. Спрингер Серия по статистике. Спрингер, Берлин, Хайдельберг, НьюЙорк (2005)

Machine Translated by Google

Расстояние Махаланобиса

Расстояние Махаланобиса – это расстояние от X к величине μ , определяемой как:

$$d2\ M(X,\mu) = (X \quad \mu)^{T} \quad {}^{1} \quad (X - \mu) .$$

Это расстояние основано на корреляции между переменными или дисперсия– ковариационная матрица.

Оно отличается от евклидова расстояния тем, что учитывает корреляцию набор данных и не зависит от масштаба измерения. Расстояние Махаланобиса равно широко используется в кластерном анализе и других методах классификации .

ИСТОРИЯ

Введено расстояние Махаланобиса.

Махаланобиса, ПК, 1936 г.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если μi обозначает E(Xi), то по определению ожидаемое значение вектора X = (X1,...,Xp) :

$$E[X] = м = \begin{matrix} мк1 \\ мк2 \\ . \\ . \\ . \\ мкп \end{matrix} .$$

Матрица дисперсии- ковариации Σ X равна определяется:

$$\Sigma = \begin{matrix} \sigma^2_1 & \operatorname{cov}(X1, X2) & \dots & \operatorname{cov}(X1, Xp) \\ & \sigma^2_2 & & \\ \vdots & & \dots & \vdots \\ & & & \sigma^2_п \end{matrix}$$

$$= E[XX] \quad \mu\mu .$$

Это квадратная симметричная матрица порядка p. Если переменные Xi стандартизированы, то Σ идентична корреляционной матрице :

$$\begin{matrix} 1 & \rho12 & \rho13 & \dots & \rho1p \\ & 1 & \rho23 & \dots & \rho2p \\ & & 1 & \dots & \rho3п \\ & & & \dots & \vdots \\ & & & & 1 \end{matrix} .$$

Тогда Σ 1/2. Y = Σ 1/2(X μ) является dom стандартизированный вектор с некорреляцией ред компоненты. Случайная величина (X μ) Σ 1(X μ) = D2 имеет математическое ожидание p. На самом деле $\sum_{i=1}^{p} Y_i^2$, D2 = где Yi случайных величин со средним значением 0 и 1. D равно называется расстояние м Махаланобиса от X до μ.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Классификация

Кластерный анализ

Анализ корреспонденции

Ковариационный анализ

Расстояние

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Махаланобис, ПК: О тестах и мерах расхождения групп. Часть I. Теоретические формулы. J. Asiatic Soc. Бенгалия 26, 541–588 (1930)

Махаланобис, ПК: Об обобщенном расстоянии в статистике, Proc. физ. Инст. наук. Индия 2, 49–55 (1936)

Махаланобис, Прашанта Чандра

Прашанта Чандра Махаланобис (1893–1972) учился в школе в Калькутте до 1908 года. В 1912 году он получил степень бакалавра наук. по физике в Президентском колледже Калифорнийского университета. В 1915 году он отправился в Англию, чтобы получить степень бакалавра наук по математике и физике в Королевском колледже Кембриджа. Прежде чем начать свои исследования, он вернулся в Калькутту в отпуск, но не вернулся в Англию из-за войны.

Как статистик Махаланобис использовал статистику не как самоцель, а в основном как инструмент для понимания и интерпретации научных данных, а также для принятия решений, касающихся благосостояния общества. Он использовал статистику сначала в своих исследованиях по антропологии (около 1917 г.), а затем наводнения. В 1930 году он опубликовал статью о статистике D2 под названием «Тест и меры групповой дивергенции».

Его величайшим вкладом в статистику, несомненно, является широкомасштабная перепись населения. Он внес большой вклад в эту область, такую как оптимальный выбор выборки с использованием функций дисперсии и стоимости. Он был назначен президентом Подкомиссии ООН по Ста

в 1947 г., занимая эту должность до 1951 г.

Также примечательна его дружба с Рональдом Эйлмером Фишером, которая длилась с 1926 года до смерти Фишера.

Основные работы и статьи Прасанты Чандра Махаланобис:

1930 г. О признаках и мерах расхождения групп. Часть I. Теоретические формулы. Журнал Азиатского общества Бенгалии, 26 стр. 541–588.

1936 г. Об обобщенном расстоянии в статистике, Тр. Нац. Инст. наук. Индия 2, 49–55.

1940 г. Выборочное обследование посевных площадей под джутом в Бенгалии. Санкхья 4, 511–530.

1944 г. О крупномасштабных выборочных обследованиях. Фил. Транс. Рой. соц. Лонд. сер. Б 231329-451.

1963 Подход операционных исследований к планированию в Индии. Издательство Азии, Бомбей.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Расстояние Махаланобиса

Основной эффект

В факторном эксперименте основным эффектом является вклад уровня фактора, характерного для переменной отклика.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Линейная модель, которая принимает во внимание только основные эффекты факторов и игнорирует взаимодействие между факторами, называется линейной моделью

Machine Translated by Google

аддитивная модель. Рассмотрим аддитивную модель конструкции с двойной классификацией:

$$Yij = \mu + \alpha i + \beta j + \varepsilon ij,$$
$$i = 1, 2,\ldots,a \text{ и } j = 1, 2,\ldots,b,$$

где Yij —ответ на лечение ij, μ —среднее значение, общее для всех лечения, αi – основной эффект i-го уровня фактора A, βj – основной эффект j- го уровня уровень фактора B, а ij – экспериментальный ошибка наблюдения Yij .
Если a = b = 2, то основные эффекты факторы A и B, обозначаемые как da и b соответственно, даны:

$$a = \frac{(\alpha 1 \quad \alpha 0)\,(\beta 1 + \beta 0)}{2},$$
$$б = 2\,\frac{(\alpha 1 + \alpha 0)\,(\beta 1 \quad \beta 0)}{}.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Оценка
- Фактор
- Факторный эксперимент
- Проверка гипотезы
- Взаимодействие

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Box, GEP, Hunter, WG, Hunter, JS:
Статистика для экспериментаторов. Введение в проектирование, анализ данных и Построение модели. Уайли, НьюЙорк (1978)

Тест Манна-Уитни

Тест Манна–Уитни непараметрический. тест, целью которого является проверка равенства двух популяций. Он используется, когда у нас есть два образца происходящие из двух популяций.

ИСТОРИЯ

Тест Манна-Уитни был впервые введен для случая m = n Уилкоксон, Франк (1945). Затем критерий Уилкоксона был распространен на случай образцов различных размеров Уайтом К. (1952).
Обратите внимание, что тест эквивалентен тесту Критерий Уилкоксона был независимо разработан и введен Фестингером Л. (1946).
Манн, Х.Б. и Уитни, Д.Р. (1947 г.) первым рассмотрел образцы различных размеры и предоставить таблицы, адаптированные к малым образцам.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть (X1, X2,...,Xn) —выборка размерности n из совокупности 1, и пусть (Y1, Y2,...,Ym) —выборка размерности m происходящие из населения 2.
Мы получаем N = n+m наблюдений, которые классифицируем в порядке возрастания, каждый раз отмечая выборку, к которой относится данное наблюдение. Статистика Манна и Уитни, обозначаемая U определяется как общее количество раз, Xi предшествует Yj в классификации в_ возрастающий порядок NoObservations.Когда размер образца достаточно велик, определение U становится долгим, но мы можно использовать следующее соотношение, которое дает одинаковые результаты:

$$U = мин + \frac{n\,(n + 1)}{2} \quad T,$$

где T – сумма рангов, присвоенных X из теста Уилкоксона.

Гипотезы

Гипотезы, соответствующие Критерий Манна-Уитни можно сформулировать как следует, согласно односторонней или двусторонней случай:

Machine Translated by Google

A: Двусторонний чехол:

$$H0 : P(X < Y) = \frac{1}{2}$$

$$H1 : P(X < Y) = \frac{1}{2}.$$

B: Односторонний чехол:

$$H0 : P(X < Y) \quad \frac{1}{2}$$

$$H1 : P(X < Y) > \frac{1}{2}.$$

C: Односторонний случай:

$$H0 : P(X < Y) \quad \frac{1}{2}$$

$$H1 : P(X < Y) < \frac{1}{2}.$$

В случае A нулевая гипотеза соответствует ситуации, когда нет разницы между популяциями. В случае B нулевая гипотеза означает, что население 1 больше чем популяция 2. В случае C нулевая гипотеза указывает на то, что популяция 1 меньше чем население 2.

Правила принятия решений

Если m, n < 12, мы сравниваем статистику U с значение , найденное в таблице Манна-Уитни, для тест H0. С другой стороны, если m, n 12, то выборочное распределение U очень быстро приближается к нормальному распределению с среднее $\mu = \frac{м \cdot н}{2}$ и стандартное отклонение $\sigma = \sqrt{\frac{мн\,(м + п + 1)}{12}}$.

Таким образом, мы можем определить значение наблюдаемое значение U по:

$$Z = \frac{U - мю}{о},$$

где Z —случайная стандартизованная величина.

Другими словами, мы можем сравнить значение Таким образом Z получается с нормальной таблицей.

Кроме того, правила принятия решений различны. и зависят от выдвигаемых гипотез. В этом таким образом, у нас есть правила A, B и C относительно предыдущие случаи A, B и C.

Правило принятия решения А

Отклоняем нулевую гипотезу H0 на уровне значимости α, если U меньше значения в таблице Манна–Уитни с параметрами n, m и $\frac{\alpha}{2}$ обозначается через $t_{n,m,1-\frac{\alpha}{2}}$, или если U больше, чем значение Манна-Уитни. таблица для n, m и 1 есть, $\frac{\alpha}{2}$, обозначается $t_{n,m,\frac{\alpha}{2}}$, что если

$$t_{n,m,1-\frac{\alpha}{2}} < U < t_{n,m,\frac{\alpha}{2}}.$$

Правило принятия решения Б

Отклоняем нулевую гипотезу H0 на уровне значимости α, если U меньше значения в таблице Манна–Уитни с параметрами n, m и α, обозначаемые через $t_{n,m,\alpha}$, т.е. если

$$U < t_{n,m,\alpha}.$$

Правило принятия решения С

Отклоняем нулевую гипотезу H0 на уровне значимости α, если U больше значения в таблице Манна–Уитни с параметрами n, m и 1 α, обозначаемые через $t_{n,m,1-\alpha}$, т.е. если

$$U > t_{n,m,1-\alpha}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Так как существует линейная зависимость между Тест Уилкоксона и тест Манна-Уитни ($U = mn + n(n+1)/2 - T$), эти два теста эквивалентны.

Таким образом, они имеют два общих свойства, и оба должны соблюдаться:

1. Обе выборки взяты случайным образом. от их соответствующих популяций.
2. Помимо независимости внутри каждой выборки существует взаимная независимость между выборками.
3. Шкала измерения не ниже ordinal.

Machine Translated by Google

ПРИМЕРЫ

В классе у нас девять учеников: пять мальчиков и четыре девушки. Класс сдает тест на умственное расчет, чтобы увидеть, если мальчики, в целом, лучше, чем девушки в этом домене. Данные представлены следующей таблице, наивысший балл соответствует лучшему результату теста.

Мальчики (Си)	Девушки (И)
9.2	11.1
11.3	15,6
19,8	20,3
21.1	23,6
24,3	

Классифицируем баллы в порядке возрастания, отмечая каждый раз выборка, к которой принадлежит данное наблюдение.

Score	Sample	Rank	Количество предшествующих раз	Икс
9,2	х	1	4	
11,1	года	2	–	
11,3	х	3	3	
15,6	г	4	–	
19,8	х	5	2	
20,3	г	6	–	
21,1	х	7	1	
23,6	г	8	–	
24,3	х	9	0	

Находим U = 4 + 3 + 2 + 1 = 10. В

таким же образом, если использовать соотношение $U = mn + \frac{n(n+1)}{2} - T$, то имеем:

$$T = \sum_{i=1}^{n} R(Xi) = 1 + 3 + 5 + 7 + 9 = 25$$

и $U = 5 \cdot 4 + \frac{5 \cdot 6}{2} - 25 = 10.$

Гипотезы, которые необходимо проверить, следующие:

H0: Мальчики обычно не лучше девочек в умственном исчислении.

H1: Мальчики обычно лучше девочек в умственный расчет.

Нулевая гипотеза может быть сформулирована следующим образом: H0: P(X < Y) 1/2 и соответствует к делу С. Решение принимается следующим образом:

Отклонить H0 на уровне значимости α, если

$$U > tn,m,1 \quad \alpha,$$

где tn,m,1 α – значение в Манн–Таблица Уитни с параметрами n, m и 1-а. Напомним, что в нашем случае m = 4 и n = 5. Поскольку у нас есть связь между Уилкокс на тесте и тесте Манна-Уитни, вместо этого использования таблицы U, мы используем критерий Уилкоксона таблица для T. Если мы выберем α = 0,05, то значение таблицы Уилкоксона равно 32. Поскольку T = 25 < 32, мы не отвергаем H0 и сделать вывод, что мальчики в целом не лучше чем девушки в умственном расчете.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Непараметрический тест
- Тест Уилкоксона

ССЫЛКА

Фестингер, Л.: Значение различий между средствами безотносительно к функция распределения частот. Психометрика 11, 97–105 (1946).

Манн, Х.Б., Уитни, Д.Р.: На тесте является ли одна из двух случайных величин стохастически больше другого. Анна. Мат. Стат. 18, 50–60 (1947)

Уайт, К.: Использование рангов в тесте значимости для сравнения двух методов лечения. Биометрия 8, 33–41 (1952).

Уилкоксон, Ф.: Индивидуальные сравнения методы ранжирования. Биометрия 1, 80–83 (1945)

M

Machine Translated by Google

Функция предельной плотности

В случае пары случайных величин (X, Y), когда случайная величина X (или Y) рассматривается сама по себе, ее функция плотности называется предельной функцией плотности.

ИСТОРИЯ

См. вероятность.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В случае пары непрерывных случайных величин (X, Y) с совместной функцией плотности f(x, y) предельная функция плотности X получается путем интегрирования:

$$P(X \quad a, \quad < Y < \quad) \overset{"="}{=} \int_{-\infty}^{a} \int_{-\infty}^{\infty} f(x, y)\,dy\,dx \overset{"="}{=} \int_{-\infty}^{a} fX(x)\,dx ,$$

где $fX(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$ —плотность X. _

Предельная функция плотности Y получается таким же образом:

$$fY(y) = \int_{-\infty}^{\infty} f(x, y)\,dx .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция плотности
- Совместная функция плотности
- Пара случайных величин

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Хогг, Р.В., Крейг, А.Т.: Введение в математическую статистику, 2-е изд. Макмиллан, НьюЙорк (1965)

Росс, С.М.: Введение в вероятностные модели, 8-е изд. Джон Уайли, НьюЙорк (2006 г.)

Предельная функция распределения

В случае пары случайных величин (X, Y), когда случайная величина X (или Y) рассматривается сама по себе, ее функция распределения называется маргинальной функцией распределения.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим пару случайных величин (X, Y) и их совместную функцию распределения:

$$F(a, b) = P(X \quad a, Y \quad b) .$$

Предельная функция распределения X равна:

$$\begin{aligned} FX(a) &= P(X \quad a) \\ &= P(X \quad a, Y < \quad) \\ &= P[\lim_{b} (X \quad a, Y \quad b)] \\ &= \lim_{b} P(X \quad a, Y \quad б) \\ &= \lim_{б} F(a, б) \\ &= F(a, \quad) . \end{aligned}$$

Аналогичным образом, предельная функция распределения Y равна:

$$FY(b) = F(\quad , b) .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция распределения
- Совместная функция распределения
- Пара случайных величин

Марков Андрей Андреевич

Марков Андрей Андреевич родился в 1856 г. в Рязани и умер в 1922 г. в Петербурге.

Этот ученик Чебышева, Пафнутий Львович, внес большой вклад в

Machine Translated by Google

область исчисления вероятностей. Он дал строгое доказательство центральной предельной теоремы. Благодаря его работам о цепях Маркова концепция марковской зависимости превратилась в современную ориюи применение случайных процессов. Глобализм его работ повлиял на развитие вероятности и международной статистики.

Он изучал математику в Санкт-Петербургском университете, где получил степень бакалавра и докторскую степень в 1878 году.
Затем преподавал, продолжая курсы Чебышева, Пафнутий Львович после того, как Чебышев, П.Л. ушел из университета. В 1886 году он был избран членом Математической школы.
математики, основанной Чебышевым П.Л. в Санкт-Петербургской Академии наук. Он стал действительным членом Академии в 1886 г., а в 1905 г. вышел из университета, где продолжал преподавать.
Марков и Ляпунов были известными учениками Чебышева П.Л. и всегда стремились сделать вероятность точной и практической математической наукой.

Первое появление марковских цепей в работах Маркова произошло в «Известиях» Физико-математического общества Казанского университета. Марков завершил доказательство центральной предельной теоремы, которое Чебышев начал, но не закончил. Он подошел к ней методом моментов, и доказательство было опубликовано в третьем издании « Исчисления вероятностей».

Основные работы Маркова Андрея Андреевича:

1899 г. Продолжается применение функций в расчете вероятностей, Касан. Бык. 9(2), 29–34.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Центральная предельная теорема

Теорема Гаусса – Маркова

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Марков, А.А.: Применение функций продолжается в расчете вероятностей.
Касан. Бык. 9 (2), 29–34 (1899 г.)

Матрица

Прямоугольная таблица с m строками и n столбцами называется матрицей (m×n).
Элементы, фигурирующие в таблице, называются коэффициентами или компонентами матрицы.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Матрица может быть представлена следующим образом:

$$A = \begin{matrix} a11 & a12 & \dots & a1n \\ a21 & a22 & \dots & a2n \\ \vdots & \vdots & \dots & \vdots \\ am1 & am2 & \dots & amn \end{matrix} .$$

Элемент характеризуется своим значением и положением. Следовательно, aij является элементом матрицы A, где i указывает номер строки и j номер столбца. Матрица с m строками и n столбцами имеет порядок (m×n). Существуют различные типы матриц: • Квадратная матрица порядка n
Матрица называется квадратной матрицей порядка n, если в ней столбцов столько же строк
(обозначается n). • Нулевая матрица
Матрица, на которой все элементы
null называется нулевой матрицей. Это будет обозначается
O. • Единичная
матрица. Для любой квадратной матрицы порядка n элементы, имеющие одинаковый индекс для прямой

M

Machine Translated by Google

что касается столбца [имеется в виду элементы в позиции (i, i)] образуют главную диагональ aij матрицы A на скаляр α:

Единичная матрица — это квадратная матрица, у которой «1» на главной диагонали и «0» на всех остальных. Его обозначают I или иногда In, где n — порядок единичной матрицы. • Симметричная матрица. Квадратная матрица A является симметричной, если A = , A , где A — транспонированная матрица A. • Антисимметричная матрица. Квадратная матрица A является A , антисимметричной, если A = где A — транспонированная матрица A; это условие приводит к тому, что элементы на главной диагонали aij для i = j равны нулю. • Скалярная матрица Квадратная матрица называется скалярной, если

$$a_{ij} = \beta \text{ для } i = j,\ a_{ij} = 0 \text{ для } i = j,$$

где β — скаляр.

- Диагональная матрица Квадратная матрица A называется диагональной матрицей, если aij = 0 для i = j и где aij — любое число. • Треугольная матрица Квадратная матрица A называется верхнетреугольной, если aij = 0 при i > j, или нижнетреугольной, если aij = 0 при i < j.

Операции над матрицами

- Добавление матриц

Если A = (aij) и B = (bij) являются двумя матрицами порядка (m × n), то их сумма A + B определяется матрицей C = (cij) порядка (m × n) , каждый элемент которой является суммой соответствующие элементы A и B:

$$c_{ij} = a_{ij} + b_{ij}.$$

- Умножение матрицы на скаляр

Если α — скаляр, а A — матрица, произведение αA — это матрица того же порядка, что и A получается умножением каждого элемента

$$\alpha A = \alpha \cdot a_{ij}.$$

- Умножение матриц

Пусть A и B — две матрицы; произведение AB определено тогда и только тогда, когда количество столбцов A равно количеству строк B. Если A имеет порядок (m × n) и B порядка (n × q), то произведение AB равно определяется матрицей C порядка (m × q), элементы которой получаются вычислением:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} \cdot b_{kj}$$

для i = 1,...,m и j = 1,...,q.

Это означает, что элементы i-й линии A умножаются на соответствующие элементы j - го столбца B, и результаты складываются.

- Целая степень квадратных матриц

Пусть A — квадратная матрица; если p - положительное целое число, p -я степень A определяется как:

$$A^p = A \cdot A \cdot \ldots \cdot A \text{ (p раз)}.$$

Обычно при p = 0 полагают A0 = I, где I — единичная матрица того же порядка, что и A.

- Квадратный корень диагональной матрицы

Пусть A — диагональная матрица неотрицательных коэффициентов. Обозначим через $\sqrt{A}$ квадратный корень из A, определяемый диагональной матрицей того же порядка, что и A, в которой мы заменяем каждый элемент его квадратным корнем:

$$\sqrt{A} = \begin{pmatrix} \sqrt{a_{11}} & 0 & \ldots & 0 \\ 0 & \sqrt{a_{22}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{a_{nn}} \end{pmatrix}.$$

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Использование матриц или вообще матричного исчисления широко применяется в регрессионном анализе, дисперсионном анализе и во многих других областях статистики. позволяет описать в наиболее лаконичным образом представляет собой систему уравнений и значительно облегчает обозначения. По тем же причинам матрицы появляются при анализе данных, где мы обычно сталкиваемся с большими таблицами данных.

ПРИМЕРЫ

Пусть A и B —две квадратные матрицы порядка 2:

$$A = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \quad \text{и} \quad B = \begin{matrix} 6 & 2 \\ 2 & 1 \end{matrix}.$$

Сумма двух матриц определяется как:

$$A + B = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} + \begin{matrix} 6 & 2 \\ 2 & 1 \end{matrix} \overset{"="}{} \begin{matrix} 1+6 & 2+2 \\ 1+2 & 3+1 \end{matrix} \overset{"="}{} \begin{matrix} 7 & 4 \\ 1 & 4 \end{matrix}.$$

Умножение матрицы A на 3 дает:

$$3 \cdot A = 3 \cdot \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \overset{"="}{} \begin{matrix} 3 \cdot 1 & 3 \cdot 2 \\ 3 \cdot 1 & 3 \cdot 3 \end{matrix} \overset{"="}{} \begin{matrix} 3 & 6 \\ 3 & 9 \end{matrix}.$$

Матричный продукт A и B равен:

$$A \cdot B = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \cdot \begin{matrix} 6 & 2 \\ 2 & 1 \end{matrix} \overset{"="}{} \begin{matrix} 1 \cdot 6 + 2 \cdot (-2) & 1 \cdot 2 + 2 \cdot 1 \\ 1 \cdot 6 + 3 \cdot (-2) & 1 \cdot 2 + 3 \cdot 1 \end{matrix}$$

$$A \cdot B = \begin{matrix} 2 & 4 \\ 0 & 5 \end{matrix}.$$

Замечание: $A \cdot B$ и $B \cdot A$, вообще говоря, не равны, что иллюстрирует некоммутативность матричного умножения. Квадрат матрицы A получается умножением A на A. Таким образом:

$$A2 = A \cdot A = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \cdot \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \overset{"="}{} \begin{matrix} 1 \cdot 2 + 2 \cdot 1 & 1 \cdot 2 + 2 \cdot 3 \\ 1 \cdot 2 + 3 \cdot 1 & 1 \cdot 2 + 3 \cdot 3 \end{matrix}$$

$$A2 = \begin{matrix} 3 & 8 \\ 4 & 11 \end{matrix}.$$

Пусть D будет следующей диагональной матрицей порядка 2:

$$D = \begin{matrix} 9 & 0 \\ 0 & 2 \end{matrix}.$$

Определим квадратный корень из квадратной матрицы D, обозначаемый $\sqrt{D}$, с помощью

$$\sqrt{D} = \begin{matrix} \sqrt{9} & 0 \\ 0 & \sqrt{2} \end{matrix} \overset{"="}{} \begin{matrix} 3 & 0 \\ 0 & \sqrt{2}. \end{matrix}$$

M

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Анализ корреспонденции
- Анализ данных
- определитель
- Матрица шляпы
- Матрица инерции
- Обратная матрица
- Множественная линейная регрессия
- Регрессивный анализ
- След
- Транспонировать

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Балестра, П.: Матричный расчет для экономистов. Editions Castella, Альбёв, Швейцария (1972)

Machine Translated by Google

Грейбилл, Ф.А.: Теория и приложения
Линейная модель. Даксбери, Северный
округ, Массачусетс (Уолдсворт и Брукс/Коул,
Пасифик-Гроув, Калифорния) (1976)

Серл, С.Р.: Матричная алгебра полезна для
Статистика. Уайли, НьюЙорк (1982)

Максимальная вероятность

Термин «максимальная вероятность» относится к
метод оценивания параметров совокупности
по случайной выборке. Применяется
когда мы знаем общую форму размещения
населения, но когда одно или
больше параметров этого распределения
неизвестный. Метод заключается в выборе
оценщик неизвестных параметров,
значения максимизируют вероятность
получения наблюдаемой выборки.

ИСТОРИЯ

Обычно метод максимального правдоподобия
приписывается Фишеру, Рональду Эйлмеру.
Тем не менее, метод можно проследить
работам ученых 18 века
Ламберт, Дж Х. и Бернулли, Д. Фишер, Р. А.
ввел метод максимального правдоподобия
Капкшон в своих первых статистических публикациях в
1912.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим выборку из n случайных величин
X1, X2,...,Xn, где каждая случайная величина
распределена согласно распределению
вероятностей, задающему функцию плотности
f (x, θ), где θ, где – пространство параметров.
Совместная функция плотности
X1, X2,...,Xn есть

$$f(x1, \theta) \cdot f(x2, \theta) \cdot \quad \cdots \quad \cdot f(xn, \theta).$$

При фиксированном x эта функция является функцией θ
и называется функцией правдоподобия L:

$$L(\theta; x1, x2,...,xn) = f(x1, \theta) \cdot f(x2, \theta) \cdot \cdots \cdot f(xn, \theta).$$

Оценщик – это максимальная вероятность
оценка θ, если она максимизирует вероятность
функция капкшона. Обратите внимание, что эта функция L может
максимизировать, установив в ноль первый
частная производная по θ и разрешение
полученные таким образом уравнения. Более того, когда
каждая функция L и lnL максимизируется
для каждого значения θ часто можно
решать:

$$\frac{\ln L(\theta; x1, x2,...,xn)}{\theta} = 0.$$

Наконец, оценки максимального
правдоподобия не всегда уникальны, и этот метод
не всегда дает объективную оценку
тор.

ПРИМЕРЫ

Пусть X1, X2,...,Xn — случайная выборка из
нормального распределения с
среднее значение μ и дисперсией σ2. Метод
оценки максимального правдоподобия кон
помогает выбрать в качестве оценок неизвестных
параметры μ и σ2 – значения,
максимизирующие вероятность наличия наблюдаемого
образец.

Таким образом, максимизируя функцию правдоподобия
влечет за собой следующее:

$$L(\mu, \sigma^2; x1 \ldots, xn) = \frac{1}{\sigma \sqrt{2\pi}^{n}} \exp\left[\frac{\sum_{i=1}^{n} (xi - \mu)^2}{2\sigma^2}\right],$$

Machine Translated by Google

или, что то же самое, его натуральный логарифм:

$$\ln L = \ln \frac{1}{\sigma \sqrt{2\pi}^{n}} \cdot \exp\left[-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \ln \sigma \sqrt{2\pi}^{-n} \cdot \exp\left[-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \ln(2\pi)^{-\frac{n}{2}} + \ln(\sigma)^{-n} - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}$$

$$= -\frac{n}{2} \cdot \ln(2\pi) - n \cdot \ln(\sigma) - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}.$$

Приравняем к нулю частные производные $\frac{\ln L}{\mu}$ и $\frac{\ln L}{\sigma}$:

$$\frac{\ln L}{\mu} = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu) = 0, \quad (1)$$

$$\frac{\ln L}{\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma \cdot \sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 = 0. \quad (2)$$

Из (1) получаем:

$$\frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu) = 0,$$

$$\sum_{i=1}^{n} (x_i - \mu) = 0,$$

$$\sum_{i=1}^{n} x_i - n \cdot \mu = 0,$$

$$\hat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \bar{x}.$$

Таким образом, для нормального распределения среднее арифметическое есть оценка максимального правдоподобия.

матор.

Из (2) получаем:

$$\frac{n}{\sigma} + \frac{1}{\sigma \cdot \sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 = 0,$$

$$\frac{1}{\sigma \cdot \sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{n}{\sigma},$$

$$\hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu)^2.$$

Находим в этом случае как максимальную вероятность оценка эмпирической дисперсии выборки.

Обратите внимание, что оценка максимального правдоподобия беспристрастной и _ _ дисперсия смещена (делится на n вместо из n 1). Мы находим, что для нормального распределения среднее значение и дисперсия выборки являются оценкой максимального правдоподобия. среднего и дисперсии населения. Это не обязательно верно для других распределений: рассмотрим, например, случай равномерного распределения на интервале [a, b], где a и b неизвестны и где оцениваемым параметром является среднее значение распределение μ = (a + b)/2. Мы можем проверить, что оценка максимально похожей вероятности μ равна:

$$\hat{\mu} = \frac{x_1 + x_n}{2},$$

где x1 и xn – наименьшие и самые большие наблюдения выборки (x1, x2,...,xn) соответственно. Этот оценщик является средней точкой интервала с пределы x1 и xn. Мы видим, что в этом случае

Machine Translated by Google

оценка максимального правдоподобия отличается от среднего значения выборки:

$$\mu^{*} = \bar{x}.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Оценка
- Оценщик
- Тест отношения правдоподобия
- Надежная оценка

ССЫЛКИ

Фишер, Р.А.: Об абсолютном критерии подгонки частотных кривых. Беспорядок. Мат. 41, 155–160 (1912)

Иметь в виду

Среднее значение случайной величины является центральной мерой тенденции этой переменной. Его также называют ожидаемым значением. На практике термин среднее часто используется в значении среднего арифметического.

ИСТОРИЯ

См. среднее арифметическое.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X — непрерывная случайная величина с функцией плотности $f(x)$. Среднее значение μ для X, если оно существует, определяется по формуле:

$$\mu = \int x \cdot f(x)\, dx.$$

Среднее значение дискретной случайной величины X с функцией вероятности $P(X = x)$, если она существует, определяется как:

$$\mu = x \cdot P(X = x).$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Эмпирически определение среднего арифметического набора наблюдений дает нам меру центрального значения этого набора.

Эта мера дает нам оценку среднего μ случайной величины X, из которой даны наблюдения.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Среднее арифметическое
- Ожидаемое значение
- Мера центральной тенденции
- Мера эксцесса
- Мера асимметрии
- Среднее взвешенное арифметическое

Среднее абсолютное отклонение

Среднее отклонение набора количественных наблюдений является мерой дисперсии, показывающей среднее значение абсолютных значений отклонений каждого наблюдения с относительно меры положения распределения.

ИСТОРИЯ

См. оценку L1.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим $x_1, x_2, \ldots, x_n$ набор из n величин или n наблюдений, связанных с количественной переменной X. Согласно определению, среднее отклонение вычисляется следующим образом:

$$\mathrm{MD} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n},$$

где $\bar{x}$ — среднее арифметическое праздники. Когда наблюдения упорядочены

Machine Translated by Google

в частотном распределении среднее отклонение рассчитывается следующим образом:

$$\text{Э М} = \frac{\sum_{\text{я}=1}^{k} f_i \cdot |x_i - \bar{x}|}{\sum_{\text{я}=1}^{k} \text{фи}},$$

где

x_i представляет различные значения переменная,

f_i — частоты, связанные с этими ценности и

k — количество различных значений.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Среднее абсолютное отклонение является мерой дисперсии более надежной, чем стандартная отклонение.

ПРИМЕРЫ

Пять студентов последовательно сдали два экзамена для чего они получили следующие оценки:

Экзамен 1:

$$3,5 \quad 44,5 \quad 3,5 \quad 4,5 \quad \bar{x} = \frac{20}{5} = 4$$

Экзамен 2:

$$2,5 \quad 5,5 \quad 3,5 \quad 4,5 \quad 4 \quad \bar{x} = = 4 \quad \frac{20}{5}$$

Среднее арифметическое $\bar{x}$ оценок одинаково для обоих экзаменов. Тем не менее изменчивость градаций (или дисперсия) не одинаковый.

Для теста 1 среднее отклонение равно:

$$\text{Иметь ввиду отклонение} = \frac{|3,5 - 4|}{5} + \frac{|4 - 4|}{5} + \frac{|4,5 - 4|}{5} + \frac{|3,5 - 4|}{5} + \frac{|4,5 - 4|}{5} = \frac{2}{5} = 0,4.$$

Для экзамена 2 среднее отклонение очень высокое э:

$$\text{Иметь ввиду отклонение} = \frac{|2,5 - 4|}{5} + \frac{|5,5 - 4|}{5} + \frac{5\,|3,5 - 4|}{5} + \frac{|4,5 - 4|}{5} + \frac{5\,|4 - 4|}{5} = \frac{4}{5} = 0,8.$$

Таким образом, оценки второго экзамена более рассеяны вокруг среднего арифметического чем оценки первого экзамена.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- оценка L1
- Мера рассеивания
- Среднеквадратичное отклонение
- Дисперсия

Среднеквадратическая ошибка

Пусть $\hat{\theta}$ — оценка параметра θ.

Квадрат ошибки оценки θ по $\hat{\theta}$ равен $(\theta - \hat{\theta})^2$.

Среднеквадратическая ошибка (MSE) представляет собой математическое ожидание этого значения:

$$\text{СКО}(\hat{\theta}) = E(\theta - \hat{\theta})^2.$$

Мы можем доказать, что среднеквадратическая ошибка равно сумме дисперсии оценки и квадрата смещения:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + E(\hat{\theta} - \theta)^2.$$

Таким образом, когда рассматриваемая оценка с смещения, среднеквадратическая ошибка соответствует точно к дисперсии оценщика.

Machine Translated by Google

ИСТОРИЯ

В начале 19 века де Лаплас, Пьер Симон (1820) и Гаусс, Карл Фридрих (1821) сравнивали оценку неизвестной величины, основанную на наблюдениях со случайными ошибками, с игрой случайностей. Они провели параллель между ошибкой оценочной стоимости и потерями из-за такая игра. Гаусс, Карл Фридрих предложил меру неточности, квадрат ошибки. Он признавал, что его выбор был случайным, но оправдывал его математической простотой функции «сделать квадрат».

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X1, X2,...,Xn — выборка размерности n , а $\hat{\theta}$ — оценка параметра θ; отклонение (квадраты) оценки задается выражением $(\theta - \hat{\theta})^2$.

Среднеквадратическая ошибка определяется по формуле:

$$СКО\,\hat{\theta} = E\,(\theta - \hat{\theta})^2 .$$

Мы можем доказать, что:

$$MSE\,\hat{\theta} = Var\,\hat{\theta} + E\,\hat{\theta} - \theta^2 ,$$

то есть среднеквадратическая ошибка равна сумме дисперсии оценки и квадрату клона .

Фактически:

$$СКО(\theta)\hat{} = E\,(\theta - \hat{\theta})^2$$

$$= E\,\theta - E\,\hat{\theta} + E\,\hat{\theta} - \hat{\theta}^2$$

$$= E[\theta - E\,\hat{\theta}^2 + E\,\hat{\theta} - \hat{\theta}^2 + 2\cdot\theta - E\,\hat{\theta}\cdot E\,\hat{\theta} - \hat{\theta}]$$

$$= E\,\theta - E\,\hat{\theta}^2 + E\,E\,\hat{\theta} - \hat{\theta}^2 + 2\cdot\theta - E\,\hat{\theta}\cdot(E\,\hat{\theta} - E\,\hat{\theta})$$

$$= E\,E\,\hat{\theta} - \hat{\theta}^2 + \theta - E\,\hat{\theta}^2$$

$$= Var\,\hat{\theta} + (смещение)^2 ,$$

где смещение равно $E[\hat{\theta}] - \theta$.

Таким образом, когда оценка $\hat{\theta}$ не имеет смещения, то есть $E[\hat{\theta}] = \theta$, мы получаем:

$$MSE\,\hat{\theta} = Var\,\hat{\theta} .$$

Эффективность оценки обратно пропорциональна ее среднеквадратической ошибке .

Мы говорим, что оценка $\hat{\theta}$ непротиворечива, если

$$\lim_{n} СКО(\theta)\hat{} = 0 .$$

В этом случае дисперсия и смещение оценка стремится к 0, когда количество наблюдателей n стремится к бесконечности .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Некоторые авторы предпочитают использовать среднее отклонение по абсолютной величине, определяемое формулой $E[|\theta - \hat{\theta}|]$, где $\hat{\theta}$ – оценка параметра θ. Недостаток последнего определения находится на уровне математических свойств, так как функция абсолютной величины не может быть различна. везде дифференцированные .

ПРИМЕРЫ

Мы рассматриваем множественную линейную регрессию в матричной форме:

$$Y = X\cdot\beta + \varepsilon ,$$

где

Y - вектор порядка n (n наблюдений независимой переменной),

X — заданная матрица n × p,

β — вектор порядка p коэффициентов быть определено, и

Machine Translated by Google

ε —вектор ошибок, которые мы предполагаем не зависимыми, и каждое последующее аномальное распределение со средним значением 0 и дисперсией σ2.

Вектор β^ = (X · X) 1 · X · Y дает оценку без смещения β.

Затем мы получаем:

MSE β^ = Var β^ .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Предвзятость

Оценщик

Ожидаемое значение

Дисперсия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Гаусс, К. Ф.: В: Работа Гаусса (1803–1826) по теории наименьших квадратов.

Транс. Trotter, HF Statistical Techniques Research Group, Технический отчет № 5.

Издательство Принстонского университета, Принстон, Нью-Джерси (1821 г.)

Лаплас, PS de: Аналитическая теория вероятностей, 3-е изд. Курсье, Париж(1820 г.)

Мера центральной тенденции

Мерой центральной тенденции является статистика, которая обобщает набор данных относительно к количественной переменной. Точнее, он позволяет определить фиксированное значение, называемое центральным значением, вокруг которого набор данных имеет тенденцию группироваться. Основными мерами центральных тенденций являются:

- Среднее арифметическое
- Медиана
- Режим

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Среднее значение μ, медиана Md и мода Mo являются мерами центра распределения.

В симметричном распределении среднее, медиана и мода совпадают по общему значению.

Это значение дает центр симметрии распределения. Для симметричного унимодального распределения имеем:

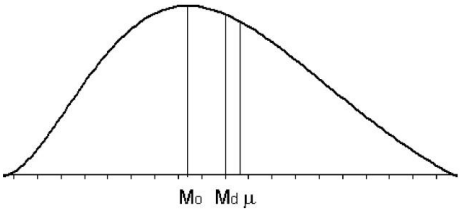
μ знак равно Md знак равно Mo .

В этом случае оценка центра распределения может быть дана по среднему, медиане или моде.

В асимметричном распределении разные меры центральной тенденции имеют разные значения. (Их сравнение можно использовать как меру асимметрии.)

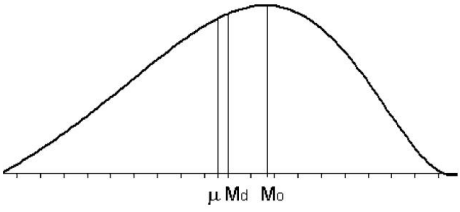
Для унимодального распределения, растянутого до правильно, у нас обычно есть:

μ Md Mo .



Это соотношение инвертируется, когда распределение функция унимодальна и вытянута влево:

μ Md Mo .



Для одномодального распределения, умеренно асимметричного, эмпирически доказано, что

M

Machine Translated by Google

мода Mo, медиана Md и среднее значение μ часто приближенно удовлетворяют следующему соотношению:

$$\text{Mo} \approx \mu = 3 \cdot (\text{Md} - \mu).$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

На практике выбор меры, позволяющей наилучшим образом охарактеризовать центр множества наблюдений, зависит от специфики статистических исследований и информацию, которую мы хотим получить.

Сравним среднее арифметическое, моду и медиану в зависимости от разных критериев:

1. Среднее арифметическое:

- Зависит от значения всех наблюдений.
- Простота интерпретации.
- Является наиболее знакомым и наиболее используемым мера.
- Часто используется в качестве оценки среднего значения генеральной совокупности.
- Имеет значение, которое может быть искажено выбросами.
- Кроме того, сумма квадратов отклонений каждого наблюдения x_i набора данных и значения α минимальна, когда α равно среднему арифметическому:

$$\min_{\alpha} \sum_{i=1}^{n} (x_i - \alpha)^2$$

α = арифметика. иметь ввиду

2. Медиана:

- Легко определить, поскольку требуется только одна классификация данных.
- Легко понять, но используется реже, чем среднее арифметическое.
- Не подвержен влиянию выбросов, что дает ему преимущество перед арифметикой.

среднее значение, если в сериале действительно есть выбросы.

- Используется в качестве оценки центрального значения Характеристики распределения, особенно когда оно асимметрично или имеет выбросы.
- Сумма квадратов отклонений по абсолютной величине между каждым наблюдением x_i набора данных и значением α минимальна когда α равно медиане:

$$\min_{\alpha} \sum_{i=1}^{n} |x_i - \alpha| \quad \alpha = \text{Md}.$$

3. Режим:

- Имеет практический интерес, поскольку является наиболее представленным значением множества.
- В любом случае является редко используемой мерой.
- Имеет значение, на которое мало влияют выбросы.
- Имеет значение, на которое сильно влияют колебания выборки. Он может сильно варьироваться от одного образца к другому.

Кроме того, в наборе данных может быть много (или не быть) мод.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Среднее арифметическое

Ожидаемое значение

Иметь ввиду

Мера асимметрии

медиана

Режим

Средневзвешенное арифметическое

ССЫЛКИ

Стюарт, А., Орд, Дж К.: Кендалл разработал теорию статистики. Том. I. Теория распределения. Уайли, Нью-Йорк (1943)

Додж, Ю: Premiers pas en statistique. Спрингер, Париж (1999)

Machine Translated by Google

Мера дисперсии

Мера дисперсии позволяет описать набор данных, касающихся конкретной переменной, дающий указание на изменчивость значений внутри набора данных. Мера уверенв дисперсии завершает описание определяется мерой центральной тенденции распределения.

Если мы наблюдаем различные распределения, мы можем сказать, что для некоторых из них все данные сгруппированы более или менее близком расстоянии от центральное значение; для других расстояние намного лучше.

Мы можем классифицировать меры дисперсии на две группы:

1. Меры, определяемые расстоянием между двумя репрезентативных значения распределения:
   - Диапазон, также называемый интервалом изменения ация
   - Межквартильный диапазон
2. Показатели, рассчитываемые взависимости от отклонений каждого исходного значения от центрального значения:
   - Геометрическое отклонение
   - Среднее отклонение
   - Среднее абсолютное отклонение
   - Стандартное отклонение

Среди мер дисперсии наиболее важным и наиболее часто используемым является стандартное отклонение.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

За исключением описательного аспекта, позволяющего охарактеризовать распределение или Если сравнивать множество распределений, то знание дисперсии или изменчивости распределения может иметь значительный практический интерес. Что случилось бы если бы наши решения основывались на среднем один.

- Дороги будут построены только для средний трафик, а пробки впраздничные дни были бы неизмеримы.
- Дома будут построены в таких способ противостоять средней силе ветра с последствиями, которые могут возникнуть на случай сильного шторма.

Эти примеры показывают, что в некоторых случаях наличие меры центральной тенденции недостаточно для принятия надежного решения. Более того, мера

дисперсия является показателем характера центральной меры тенденции.

чем меньше изменчивость данного множества, тем больше значение меры центрального тенденция будет репрезентативной для данных набор.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коэффициент вариации

Межквартильный диапазон

Среднее абсолютное отклонение

Мера центральной тенденции

Диапазон

Среднеквадратичное отклонение

Дисперсия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Додж, Ю: Premiers pas en statistique.
Спрингер, Париж(1999)

Мера не похожести

Мерой не похожести является расстояние мера, приписывающая паре объектов реальное положительное (или нулевое) число.

Если предположить, например, что каждый объект характеризуется n переменными, мера не похожести между двумя объектами будет заключаются в указании количества различных точки, представляющие два рассматриваемых объекта.

Machine Translated by Google

Это целое число от 0

и n и равно n , когда два объекта не имеют общих точек.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Позвольте быть населением, на котором мы хотели бы определить меру несходства s . Назовем мерой несходства s любое отображение × в действительные положительные (или нулевые) числа, удовлетворяющее следующим двум
свойствам: • Для каждого w,w в ,s (w, w) = s (w , w);

то есть s симметричен.

• Для каждого w в s (w, w) = 0, т. е. мера несходства между объектом и самим собой равна нулю

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В качестве меры несходства s на конечной совокупности имеет ограниченное применение, то есть:

макс у в, ш ; w, w = m < .

Мы можем определить меру подобия населения следующим образом:

шв, ш знак равно м - шв, ш .

Эти два понятия, мера сходства и мера несходства, можно считать взаимодополняющими.

Для решения задач классификации мы используем меру несходства, а не расстояние , которое является ее частным случаем и обладает так называемым свойством треугольного неравенства.

ПРИМЕРЫ

Рассмотрим популяцию, состоящую из трех цветов

= {ш1, ш2, ш3}

на котором были измерены следующие три переменные:

• v1: цвет цветка • v2:
количество лепестков •
v3: количество листьев, на
которых

v1 резюмируется здесь как дихотомическая переменная ,
потому что мы рассматриваем только красный цвет.
(R) и желтые (Y) цветы и

v2 и v3 —дискретные случайные величины,
принимающие свои значения в количественных числах.

Мы наблюдали:

Переменные

	B_1	B_2	B_3
$Ж_1$	p	5	2
$Ж_2$	г	4	2
$Ж_3$		p	5 1

Определяем для каждой пары цветов количество баллов, отличающих их. Мы видим, что у нас есть мера несходства для набора из трех цветов:

s (w1, w1) = s (w2, w2) = s (w3, w3) = 0

потому что цветок не отличен от самого себя;

s (w1, w2) = s (w2, w1) = 2

потому что цвет и количество лепестков на w1 и w2 разные ;

s (w1, w3) = s (w3, w1) = 1

так как только число листьев позволяет отличить w1 от w3;

s (w2, w3) = s (w3, w2) = 3

asw2 и w3 различны по каждой из трех исследуемых характеристик.

С помощью этих возможных значений мы можем сделать вывод, что s удовлетворяет двум свойствам меры несходства.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Классификация
- Расстояние
- Мера подобия
- Население

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Селё, Г., Дидай, Э., Говерт, Г., Лешеваль Лиер, Ю, Рал амбондрани, Х.: Автоматическая классификация понятий—аспекты статистика и информатика. Дуно, Париж(1989)

Хэнд, Д.Дж: Дискриминация и классификация. Уайли, НьюЙорк (1981)

Мера эксцесса

Меры эксцесса являются частью меры формы и характеризуют аспект вида заданного распределения. Точнее, они характеризуют степень эксцесса распределения в сторону нормального распределения. бутион. Некоторые дистрибутивы близки к нормальное распределение, не будучи полностью идентичный. Очень полезно иметь возможность проверить, форма распределения представляет собой отклонение от эксцесса нормального распределения бутион.

Мы говорим о платикуртическом распределении, если кривая более сглажена, чем обычно кривой и лептокуртикального распределения, если она более острая, чем нормальная кривая. Чтобы проверить эксцесс кривой, мы используем коэффициент эксцесса.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Дана случайная величина X с вероятностью распределения fx коэффициент эксцесса равен определяется как:

$$\gamma 2 = \frac{\mu 4}{\sigma 4} - 3,$$

где μ4 обозначает четвертый момент относительно среднее значение, σ4 – квадрат дисперсии. Эксцесс выборки получается с использованием аналогов выборки, то есть:

$$\gamma \hat{} \text{ знак равно } \frac{\mu \hat{} 4}{\hat{} \sigma 4} - 3.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент эксцесса
- Мера формы
- Мера асимметрии

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Джоанс, Д.Н. и Гилл, Калифорния: Сравнение показателей асимметрии пробы и эксцесса, JRSS.D 47(1), 183–189 (1998)

Мера местоположения

Мерой местоположения (или положения) является мера, которая предлагает синтезировать набор статистические данные только на одно фиксированное значение, выделяя конкретную точку из участвующей переменной. Наиболее часто используемые меры позиции:

- Показатели центральной тенденции (среднее, мода, медиана), которые, как правило, определяют центр набора данных.
- Квантиль (квартиль, децильь, центильь), которые стремятся определить не центр, а конкретная позиция подмножества данных.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Меры местоположения, а также меры дисперсии и формы, используются особенно для сравнения многих наборов данных или многих частотных распределений.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Мера центральной тенденции
- квантиль

Machine Translated by Google

Мера формы

Можно попытаться охарактеризовать форму частотного распределения с помощью соответствующих коэффициентов. Определенные частотные распределения близки к нормальному распределению («кривая колокола»). Эти распределения могут также представлять собой асимметриюили эксцесс от нормальной кривой. Две меры формы являются:

- Мера асимметрии • Мера эксцесса.

ИСТОРИЯ

Пирсон, Карл (1894, 1895) был первым, кто проверил различия между определенными дистрибутивами. и нормальное распределение. Мы показали, что отклонения от нормы

Кривая может характеризоваться моментами 3-го и 4-го порядка распределения.

До 1890 г. Грам, Йорген Педерсен и Тиле, Торвальд Николаи из Дании разработали теориюсимметрии частотных кривых. После того, как Карл Пирсон опубликовал своюсложную и трезвыйайно интересную систему (1894, 1895), на эту тему было опубликовано много статей.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Мера рассеивания
- Мера эксцесса
- Мера местоположения
- Мера асимметрии

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Пирсон, К.: Вклад в математическую теорию эволюции. I. В: Ранние статистические статьи Карла Пирсона. Издательство Кэмбриджского университета, Кембридж, стр. 1–40 (1948). Впервые опубликовано в 1894 г.

Пирсон, К.: О рассечении асимметричных частотных кривых. Филос. Транс. Рой. соц. Лонд. сер. А 185, 71–110

Пирсон, К.: Вклад в математическую теорию эволюции. II: Косая вариация в однородном материале. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, Кембридж, стр. 41–112 (1948). Впервые опубликовано в 1895 году в Philos. Транс. Рой. соц. Лонд. сер. А 186, 343–414

Мера подобия

Мерой сходства является мера расстояния, присваивающая вещественному положительному (или нулю) числу соответствующие пары объектов из учаемой совокупности. Если мы предположим, например, что каждый объект характеризуется n переменными, то мера сходства между двумя объектами будет заключаться в присвоении числа общих точек двум рассматриваемым объектам. Это целое число от 0 до n и равно n, когда два объекта идентичны.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Позвольте быть населением, на котором мы хотели бы определить меру сходства s. Назовем мерой подобия s любое отображение из x в действительные положительные (или нулевые) числа, удовлетворяющее следующим условиям: • , ж ,

Для каждого w, w в ,sw, w = sw , т.е. симметрично; • с ждя каждого w, w в

с(ш, ш) = ш , ш > ш, ш ,

то есть объект более «похож» на себя, чем на другой объект.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Мера сходства s в популяции имеет ограниченное применение, то есть:

Machine Translated by Google

макс у в, ш ; ш, ш е =

s(w, w) < ,

где w любой объект ; если мы положим m =
s(w, w), то мы можем определить меру
не похожести совокупности по формуле:

шв, ш знак равно м - шв, ш .

Эти два понятия, мера подобия и
Мерой не похожести можно считать
дополняющий.
Мы используем меры подобия, а также
как меры не сходства, для задач классификации .

ПРИМЕРЫ

Пусть население будет формами трех
цветы

= {ш1, ш2, ш3}

на котором мы измеряем три переменные:

- v1: цвет цветка
- v2: количество лепестков
- v3: количество листье в

где

v1 резюмируется здесь как дихотомический

переменная , потому что мы рассматриваем только красный

(R) и желтые (Y) цветы и

v2 и v3 —дискретные случайные величины,

принимающие свои значения в количественных числах.

Мы наблюдали:

Переменные

	в 1 2	в	в 3
ж 1 р 5			2
ж 2	Д 4		2
ж 3 Р 5			1

Определяем для каждой пары цветков
количество общих для них точек.
Мы видим, что у нас есть мера подобия
из набора из трех цветов:

s(w1, w1) = s(w2, w2) = s(w3, w3) = 3 ,

что соответствует числу переменных;

s(w1, w2) = s(w2, w1) = 1

потому что цвет и количество лепестков
разные при w1 и w2;

s(w1, w3) = s(w3, w1) = 2

так как только число листье в позволяет
отличить w1 от w3;

s(w2, w3) = s(w3, w2) = 0

так как w2 и w3 не имеют общих точек.
С помощью этих возможных значений мы можем
вывести, что s удовлетворяет двум свойствам
мера сходства.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Расстояние
- Мера не похожести
- Население

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

См. меру не похожести.

M

Мера асимметрии

В асимметричном распределении медиана ,
среднее арифметическое и мода совпадают
центральная точка. Это не верно, когда
распределение асимметрично. В этом случае режим
отдельно от среднего арифметического, а
медиана находится между двумя из них. В конце концов
Следовательно, необходимо разработать меры
асимметрии для изучения степени отклонения
формы распределения от
симметричное распределение . Основной перекос
Меры безопасности:

- Коэффициент Юла и Кендалла
- Коэффициент асимметрии

Machine Translated by Google

МАТЕМАТИЧЕСКИЙ АСПЕКТ

Дана случайная величина X с вероятностью распределения fx коэффициент асимметрии равен определяется как:

$$\gamma 1 = \frac{\mu 3}{\sigma 3}$$

где μ3 – третий момент относительно среднее значение, σ3 – отклонение стенда.

ИСТОРИЯ

См. коэффициент асимметрии.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент асимметрии
- Мера эксцесса
- Мера формы
- Коэффициент Юла и Кендалла

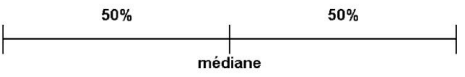
медиана

Медианамера центрального сгиба

су определяется как значение, которое находится вцентре набора упорядоченных наблюдений, когда он находится в порядке возрастания или убывания.
Для случайной величины X медиана, обозначаемая по Md, обладает тем свойством, что:

$$P(X \leq Md) \geq \frac{1}{2} \text{ и } P(X \geq Md) \geq \frac{1}{2}.$$

Затем мы находим 50% наблюдений на каждом сторона медианы:



Если у нас есть нечетное количество наблюдений, медиана соответствует значениям среднее наблюдение. Если у нас есть четное число Среди наблюдений нет единственного наблюдения в середине. Медиана будет даваться средним арифметическим значений два наблюдения середины заказанного наблюдения.

ИСТОРИЯ

В 1748 году Эйлер, Леонард и Майер, Иоганн Тобиас независимо предложил метод заключается вразделении наблюдений набор данных на две равные части.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Когда наблюдения даны индивидуально, процесс вычисления медианы простой:

1. Расположить n наблюдений в порядке возрастания или в порядке убывания.
2. Если количество наблюдений нечетное:
   - Найдите наблюдение середины $\frac{n+1}{2}$.
   - Медиана равна значению среднее наблюдение.
3. Если число наблюдений четное:
   - Найдите два наблюдения середины дель $\frac{n}{2}$ и $\frac{n}{2}+1$.
   - Медиана равна арифметической среднее значение этих двух наблюдателей праздники.

Когда у нас нет индивидуального обсерва ции, но они сгруппированы в классы, медиана определяется следующим образом:

1. Определить срединный класс (тот, который содержит $\frac{n}{2}$-е наблюдение).
2. Рассчитайте медиану, используя следующую формула:

$$Md = L + \frac{\frac{n}{2} - \text{плавник}}{fMD} \cdot c,$$

где

L — нижний предел медианы сорт,

n общее количество наблюдений ции,

Machine Translated by Google

finf - сумма частот
ниже среднего класса,

fMd – частота медианы
класс, и

c это длина интервала
средний класс.

Эта формула основана на гипотезе
согласно которым наблюдения
равномерно распределены внутри каждого класса. Это
также предполагает, что нижняя и верхняя
границы медианного класса определены и
известен.

Характеристики

Сумма отклонений по абсолютным значениям
между каждым наблюдением xi набора данных и
значение α является минимальным, когда α равно медиану:

$$\text{минимизировать}_{\alpha} \sum_{i=1}^{n} |x_i - \alpha| \quad \alpha = \text{Md}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

При расчете среднего арифметического
мы учитываем ценность каждого наблюдателя
рация Таким образом, выброс сильно влияет
среднее арифметическое. С другой стороны, в
расчет медианы, мы принимаем в
учитывать только ранг наблюдений.
Таким образом, выбросы не влияют на медиану.

Когда мы имеем сильно перекошенное
частотное распределение, нас интересует
выбор медианы в качестве меры центральной
тенденции, а не среднего арифметического, чтобы
нейтрализовать действие экстремального вала
уэс.

ПРИМЕРЫ

Медиана набора из пяти чисел 42, 25,
75, 40 и 20 — это 40, потому что 40 — это середина.
номер, когда номера упорядочены:

20 25 40 42 75 .
медиана

Если к этому множеству добавить число 62,
число наблюдений будет четным, а медиана
равно среднему арифметическому двух
средних наблюдений:

20 25 40 42 62 75 .
$\frac{40+42}{2}$ =41=медиана

Рассмотрим теперь медиану данных, сгруппированных
в классы. Возьмем, к примеру, следующую
таблицу частот, которая представляет
дневная выручка 50 продуктовых магазинов:

Сорт (доход в евро)	Частота f_i (количество продуктовые магазины)	Накопленный частота
500–550	3	3
550–600	12	15
600–650	17	32
650–700	8	40
700–750	6	46
750–800	4	50
Общий	50	

Средний класс - это класс 600–650, т.к.
он содержит среднее наблюдение (25-е
наблюдение).
Предположим, что наблюдения равномерно
распределены в каждом классе, медиана
равно:

$$\text{Md} = 600 + \frac{\frac{50}{2} - (650 - 600)}{17} \cdot 15$$

$$= 629{,}41 .$$

Это результат означает, что 50 % продуктовых
магазинов имеют ежедневную выручку, превышающую
629,41 евро, а остальные 50% имеют обр.
enues на общую сумму менее 629,41 евро.

M

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Иметь в виду

Мера центральной тенденции
Режим

квантиль

ССЫЛКИ

Боскович, Р. Дж: De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex instanceria etiam sensorum impress sa. Bononiensi Scientiarium et Artium

Instituto Atque Academia Commentarii, vol. IV, стр. 353–396 (1757 г.)

Эйлер, Л.: Исследования по вопросу о наименьших Галитов движения Сатурна и Юпитера, часть, о которой сообщается о премии 1748 года, парижская Королевская академия наук. Republié en 1960, в исполнении Леонарди Эйлери, Opera Omnia, 2ème série. Турич и, Базе, 25, стр. 47–157 (1749 г.)

Mayer, T.: Abhandlung über dieUmwälzung desMonds um seineAchse und die schein

голый Bewegung der Mondfleken. Kos mographische Nachrichten und Samm lungen auf das Jahr 17481, 52–183 (1750)

Среднее абсолютное отклонение

Среднее абсолютное отклонение набора количественные наблюдения являются мерой дисперсии. Он соответствует среднему значению абсолютных значений отклонения каждого наблюдения относительно медианы.

ИСТОРИЯ

См. оценку L1.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x1, x2,...,xn —набор из n наблюдений относительно количественной переменной X.

Среднее абсолютное отклонение, обозначаемое EMAD рассчитывается следующим образом:

$$\text{Э МАД} = \frac{\sum_{i=1}^{n} |x_i - \text{Md}|}{n},$$

где Md —медиана наблюдений.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Среднее абсолютное отклонение является фундаментом ментальное понятие оценки L1. Оценка L1 параметра мерыцентральной тенденции θ на самом деле является методом оценки, основанным на минимальном абсолютном

отклонения наблюдений xi от параметра θ:

$$\sum_{i=1}^{n} |x_i - \theta|.$$

Значение, которое минимизирует это выражение, является медианой выборки. В случае данных, содержащих выбросы, этот метод более эффективен, чем метод наименьших квадратов.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

оценка L1

Мера рассеивания

Среднеквадратичное отклонение

Дисперсия

Метод моментов

Для оценки используется метод моментов.

сопоставление параметров распределения из выборки. Идея заключается в следующем. Если количество оцениваемых параметров равно k, положим k первых моментов распределения (выражение которого зависит от неизвестных параметров) равными соответствующим эмпирическим моментам, т. е.

Machine Translated by Google

оценки моментов порядка k cal
на образце. Нам нужно решить систему из k
уравнений с k неизвестными переменными.

ИСТОРИЯ

В конце 19 века Пирсон, Карл использовал метод моментов для оценки

параметры нормальной смешанной модели (с параметрами p1, p2, μ1, μ2, σ2 σ2), задан2ные формулой

$$P(x) = \sum_{i=1}^{2} \pi_i \cdot g_i(x),$$

где gi нормально распределены со средним значением μi и дисперсией σ2 . Дальнейшее развитие и изучение метода получили Ч уров А.А. (1874–1926), Тиле, Торвальд Николаи, Фишер, Рональд Эйлмер, Пирсон, Карл и другие.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть θ1,..., θk —k параметров, которые необходимо оценить для случайной величины X с заданным распределением.

Обозначим через μj((θ1,..., θk)) = E Xj момент порядка j функции X , выраженный как функция оцениваемых параметров.

Метод моментов заключается в решении следующей системы уравнений:

$$\mu_j(\theta 1,\ldots,\theta k) = mj \text{ для } j = 1,\ldots,k.$$

Чтобы различать реальные параметры и оценки, обозначим последние через θ^1,..., θ^ к.

ПРИМЕРЫ

1. Пусть X —случайная величина, имеющая экспоненциальное распределение с неизвестным параметром θ. У нас есть n наблюдений x1,...,xn для X, и мы хотим оценить θ.

   Поскольку момент порядка 1 является средним значением распределения, для оценки θ следует уравнять теоретическое и эмпирическое среднее. Мы знаем, что (см. экспоненциальное распределение) среднее (теоретическое) равно θ. Искомая оценка $\hat{\theta}$ удовлетворяет

   $$\frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

2. Согласно методу моментов среднее μ и дисперсия σ2 случайной величины X , подчиняющейся нормальному распределению, оцениваются по эмпирическому среднему $\bar{x}$ и дисперсии

   $$\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

   выборки {x1, x2,...,xn} из X.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Во-первых, мы должны наблюдать, имеет ли система уравнений, данная ранее, решения.
Он не может иметь решений, если, например, оцениваемые параметры имеют одно или несколько конкретных ограничений. В любом случае метод не гарантирует особо эффективных оценок.
Например, найденная оценка дисперсии нормального распределения
это необъективно.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Расстояние хи-квадрат
- Оценщик
- Наименьших квадратов
- Максимальная вероятность
- Момент
- Параметр
- Образец

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Моменты и моменты произведения выборочных распределений. проц. Лонд. Мат. соц. 30, 199–238 (1929)

M

Machine Translated by Google

Фишер, Р.А.: Моменты дистрикта
Для нормальных выборок мера отклонения от
нормальности. проц. Рой. соц.
Лонд. сер. А 130, 16–28 (1930)

Кендалл, М.Г., Стюард, А.: Расширенная теория
статистики. Том. 1. Теория распределения, 4-е
изд. Гриффин, Лондон (1977)

Пирсон, К.: Вклад в математическую теорию
эволюции. I. В: Ранние статистические статьи
Карла Пирсона. Издательство Кэмбриджского
университета, Кембридж, стр. 1–40 (1948). Впервые
опубликовано в 1894 г. под названием:
«Рассечение асимметричных частотных
кривых». Фил ос. Транс. Рой. соц.
Лонд. сер. А 185, 71–110

Тиле, Теннесси: Теория наблюдений. К. и Э.
Лейтон, Лондон; Анна. Мат. Стат. 2, 165–308 (1903)

Недостающие данные

Отсутствующие данные возникают в статистике, когда определенный результат отсутствует. Это происходит, когда данные недоступны (данные были и у них тоже ны ошибкой сохранения, животное погибло во время эксперимента и т. д.).

ИСТОРИЯ

Аллан Ф.Е. и Уишарт Дж (1930), по-видимому, были первыми, кто рассмотрел проблему недостающих данных при анализе плана эксперимента. За их работами последовали работы Йейтса, Франка (1933).

Аллан Ф.Е. и Уишарт Дж, а также Йейтс Ф. попытались найти оценку недостающих данных методом итераций. В ситуации, когда многие наблюдения отсутствуют, предложенный метод не дает удовлетворительного решения. Бартлетт, MS (1937) использовал анализ ковариации для решения той же проблемы.

С тех пор было предложено много других методов, среди них метод Рубина Д.Б. (1972). Бирке с, Д. и соавт. (1976) представили точный метод решения проблемы отсутствия данных с использованием матрицы инциденто сти. Обобщенный матричный подход был предложен Dodge, Y. и Majum dar, D. (1979). Более подробную информацию об исторических аспектах отсутствия данных с 1933 по 1985 год можно найти у Dodge, Y. (1985).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Мы можем решить проблему отсутствия данных, повторив эксперимент в тех же условиях и таким образом получив значения для недостающие данные. Такое решение, несмотря на то, что оно идеально, не всегда экономически и материально возможно.

Существуют и другие подходы к решению проблемы отсутствия данных при планировании эксперимента. Упомянем подход по обобщенной обратной или по матрице инциденто сти, которая указывает, какие наблюдения отсутствуют в плане.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Данные
- План экспериментов
- Эксперимент
- Наблюдение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Аллан, Ф.Е., Уишарт, Дж: Метод оценки урожайности
отсутствующего участка в полевых
экспериментальных работах. Дж Сельское
хозяйство. наук. 20, 399–406 (1930)

Machine Translated by Google

Бартлетт, М.С.: Некоторые примерыстатистических методовисследования вселькомхозяйстве и прикладнойбиологии. Дж Рой. Стат. соц. (Приложение) 4, 137–183 (1937).

Биркес, Д., Додж, Ю, Сили, Дж: Охватывающие наборы для оцениваемых контрастов в классах фикационные модели. Анна. Стат. 4, 86–107 (1976)

Додж, Ю: Анализ экспериментов с отсутствующими данными. Уайли, НьюЙорк (1985)

Додж, Ю, Маджумдар, Д.: Алгоритм поиска обобщенных обратных наименьших квадратов для моделей классификации с произвольными шаблонами. Дж Стат. вычисл. Симул. 9, 1–17 (1979)

Рубин, Д.Б.: Неитерационный алгоритм для оценки отсутствующих значений по методу наименьших квадратов в любом дисперсионном анализе. заявл. Стат. 21, 136–141 (1972)

Йейтс, Ф.: Анализ повторяющихся экспериментов, когда полевые результаты не получены полный. Империя J. Опыт. Сельскохозяйственный. 1, 129–142 (1933)

Отсутствующий анализ данных

См. отсутствующие данные.

Режим

Мода является мерой центральной тенденции су. Мода набора наблюдений – это значение наблюдения, имеющего наибольшую частоту. Согласно этому определению, распределение может иметь уникальный режим (называемый унимодальным распределением). В некоторых ситуациях распределение может иметь много моду сов (так называемое бимодальное, тримодальное, мультимодальное и т. д. распределение).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим случайную величину X, ден функция города есть f(x). Мо является модой распределения, если f(x) имеет локальный максимум в точке Мо.

Эмпирически мода набора наблюдений определяется следующим образом:

1. Установить частотное распределение набор наблюдений.
2. Режим или режимы – это значения, частота больше или равна частоте остальных наблюдений.

Для распространения наблюдений сгруппированы в классы того же диапазона (= максимум - минимум), пределы которых известны, мода определяется следующим образом: 1. Определить модальные классы - классы с частотой большей или равной частотам других классов. (Если много занятий с одинаковой частотой, мы группируем эти занятия таким образом, чтобы вычислить только одну моду этого множества занятий).
2. Рассчитать моду Мо с учетом совместных занятий:

$$Мо = Л + \frac{д1}{д1 + д2} \cdot с,$$

где

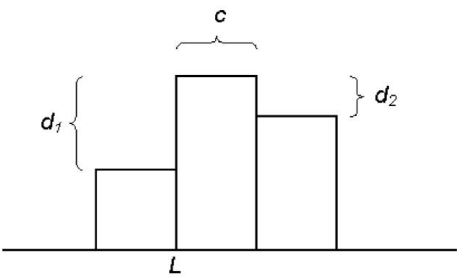
L —меньший предел модального класса,

d1 —разница между частотой модального класса и частотой предыдущего класса,

d2 —разница между частотой модального класса и частотой следующего класса, а

с —длина модального класса, общая для всех классов.

Machine Translated by Google



Если продолжительность занятий не одинакова, мода должна быть рассчитана путем изменения разделение классов для получения, если возможно, классы равной длины.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Для дискретной переменной мода имеет Преимущество в том, что его легко определить и интерпретировать. Он используется особенно тогда, когда распределение несимметрично.

Для непрерывной переменной определение режима не всегда точен, потому что он зависит от группировки по классам. В целом режим является хорошим индикатором центр данных только в том случае, если в наборе данных есть одно доминирующее значение . Если есть много доминирующих значений, распределение называется плюримодальным. В этом случае режимы не являются мерой центральной тенденции. Бимодальное распределение обычно указывает что рассматриваемая популяция в действительности неоднородна и состоит из двух субпопуляции, имеющие разные центральные значения у эс.

ПРИМЕРЫ

Пусть в наборе будут следующие числа:

1223334445555667789 .

Таблица частот , полученная из этого набора является:

Ценности	1 2 3 4 5 6 7 8 9	
Частота	1 2 3 3 4 2 2 1	1

Мода этого набора чисел равна 5, потому что он появляется с наибольшей частотой (4). Рассмотрим теперь следующую таблицу частот, что представляет собой ежедневный доход 50 магазины, сгруппированные по классам выручки:

Класс (доход в евро)	Частота фи (количество магазинов)
500–550	3
550–600	12
600–650	17
650–700	8
700–750	6
750–800	4
Общий	50

Сразу видно, что модальный класс составляет 600–650, потому что у него наибольшая частота качество Режим рассчитывается следующим образом:

$$Mo = L1 + \frac{d1}{д1 + д2} \cdot c$$

$$= 600 + \frac{5}{5 + 9} \cdot 50 = 617,86 .$$

Возьмем еще раз тот же пример, но группировка чисел в разные классы:

Класс (доход в евро)	Частота фи (количество магазинов)
500–600	15
600–650	17
650–700	8
700–800	10
Общий	50

Классы не имеют одинаковой продолжительности. Мы сделает новую группировку для получения классов такой же длины.

Machine Translated by Google

Дело 1

Возьмем, например, в качестве ссылки класс 600–650, длина которого равна 50, и предположим, что доходы равномерно распределены в классе.

Таблица, производящая эту новую группировку, является следующий:

Класс (доход в евро)	Частота f_i (количество магазинов)
500–550	7,5
550–600	7,5
600–650	17
650–700	8
700–750	5
750–800	5
Общий	50

Модальный класс — это класс 600–650, а значения режима:

$$Mo = 600 + \frac{9,5}{9,5 + 9} \cdot 50 = 625,67 .$$

Случай 2

Мы также можем основывать наши расчеты на длина первого класса; это дает следующую таблицу:

Класс (доход в евро)	Частота f_i (количество магазинов)
500–600	15
600–700	25
700–800	10
Общий	50

Модальный класс — это класс 600–700, а значения режима:

$$Mo = 600 + \frac{10}{10 + 15} \cdot 100 = 640 .$$

Эти результаты показывают, что мы должны быть осторожны к использованию режима, потому что полученный результаты различаются в зависимости от того, как данные группируются в классы.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Иметь ввиду

Мера центральной тенденции
Мера асимметрии
медиана

Модель

Модель – это теоретическое представление реальная ситуация. Он состоит из математических символов. математическая модель, на основе определенного количества наблюдений и гипотезы, пытается дать как можно лучшее описание изучаемого явления.
Математическая модель содержит существенные два типа элементов: непосредственно или косвенно наблюдаемые переменные, которые касаются изучаемое явление и параметры, которые являются фиксированными величинами, обычно неизвестными, которые связывают переменные между их внутри модели.

M

ИСТОРИЯ

Понятие «модель» впервые появилось в 18 века, в трех основных проблемах:

1. Понимание наблюдаемого неравенства между движениями Юпитера и Сатурн.
2. Определение и математическое описание движений Луны.
3. Определение формы земли.

В этих трех задачах использовались наблюдения из астрономии и некоторых теоретических знаний о гравитации. Они заняли многие ученые, особенно Эйлер, Леонгард (1749), который первым обратился к этой проблеме, и

Machine Translated by Google

Майер, Тобиас (1750 г.), который интересовался в движениях луны. Третий проблему лечил Боскович , Роджер Джозефи Мэр, Кристофер (1755 г.).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Существует множество типов моделей. Среди их, детерминистская модель — это модель, где отношения считаются точными. На с другой стороны, стохастическая модель предполагает, что соотношения не точны, а устанавливается случайным процессом. Стохастическая модель в основном определяется природой независимых переменных. (экзогенные или объяснительные). Мы говорим о непрерывные модели, когда независимые переменные Xj являются количественными непрерывными переменные . Мы пытаемся выразить зависимое переменная Y (также называемая эндогенной или ответ) функцией вида:

$$Y = f X1,...,Xp|\beta0, \beta1,..., \beta p \quad 1 + \varepsilon \quad ,$$

где βj — оцениваемые параметры (или коэффициенты), а ε — ненаблюдаемая случайная ошибка. Из этого типа модели мы различать линейные модели и нелинейные модели; линейность относится к коэффициентам. Линейную модель можно записать на вид полинома. Например:

$$Y = \beta0 + \beta1X + \beta2X2 + \varepsilon$$

является линейной моделью, называемой «второго порядка» (приказ исходит от величайшей силы переменная X). Модель

$$\text{лог } V = \beta0 + \frac{\beta1}{\text{Вт}}$$

также является линейной моделью для параметров β0 и β1. Эта модель также становится линейной при постановке переменных:

$$Y = \text{журнал } V \text{ и } X = \frac{1}{\text{Вт}} \cdot$$

Это дает:

$$Y = \beta0 + \beta1 \cdot X \quad .$$

С другой стороны, модель

$$Y = + \varepsilon \frac{\beta0}{X + \beta1}$$

является нелинейной моделью для параметра β1. Существует множество методов оценки параметры модели. Среди этих методов можно назвать метод наименьших квадратов и метод наименьших квадратов. Методы оценки L1 .

За исключением непрерывных моделей, мы говорим о дискретных моделях, когда изучаемую популяцию можно разделить на подпопуляции на основе качественных категориальных переменных (или факторов), каждая из которых может иметь определенное число уровней. Тип модели зависит от установленного экспериментального дизайн. Цель такой конструкции состоит в том, чтобы определить, как условия эксперимента влияют на наблюдаемые значения отклика переменная. Мы можем различать разные типы дискретных моделей:

- Модель с фиксированными эффектами (модель I) линейная модель, для которой факторы фиксированный, то есть взять конечное число уровней все они представлены в установленном эксперименталь ном плане . В этом случае, модель можно записать:

$$\text{наблюдаемое значение} = \text{линейная функция неизвестных параметры} + \text{ошибка};$$

ошибки являются независимыми случайными величинами, средние равны нулю

- Модель с переменными эффектами (модель II) является линейной моделью, для которой значения каждого фактора были выбраны случайным образом

Machine Translated by Google

из конечного или не конечного числа возможных уровней. В этом случае мы можем написать:

наблюдаемое значение = константа + линейная функция случайных величин + ошибка;

случайные величины не зависимы со средним значением, равным нулю.

Модель со смешанными эффектами (модель III) — это линейная модель, содержащая одновременно фиксированные во времени и случайные факторы. Таким образом, мы можно написать:

наблюдаемое значение = линейная функция неизвестных параметры + линейная функция случайного переменные + ошибка.

Когда план эксперимента содержит один фактор, обозначаемый α, модель можно записать как:

$$Yij = \mu + \alpha i + \varepsilon ij,$$
$$i = 1,...,I \text{ и } j = 1,...,ni,$$

где I — количество уровней фактора α, ni — количество наблюдений на каждом уровне, μ — среднее значение генеральной совокупности, αi — эффекты, связанные с каждым уровнем фактора α и εij — независимые случайные ошибки, нормально распределенные со средним значением до нуля и однородной дисперсией σ2.

Такая модель может быть типа I или типа II, в зависимости от природы фактора α: это тип I, если уровни фактора α фиксированы и если они все присутствуют на всем плане. Он относится к типу II, если α — случайная величина, не зависящая от распределения по какому-либо распределению с нулевым математическим ожиданием и переменным σ2 (если уровни α выбираются случайным образом из конечного или бесконечного числа возможных уровней).

Когда план эксперимента содержит много факторов, например:

$$Yijk = \mu + \alpha i + \beta j + (\alpha\beta)ij + \varepsilon ijk,$$
$$i = 1,...,I,$$
$$j = 1,...,J \text{ и } k = 1,...,nij,$$

модель может быть I типа (уровень всех факторов фиксирован), II (все факторы являются случайными величинами с теми же ограничениями, что и в предыдущей модели), или III (некоторые факторы имеют фиксированные уровни, а другие — переменные). Термин $(\alpha\beta)ij$ представляет здесь эффект взаимодействия между уровнями I фактора α и J уровней фактора β.

Модель типа I можно анализировать с помощью дисперсионного анализа.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Зависимая переменная
- План экспериментов
- Независимая переменная
- Параметр
- Случайная переменная
- Регрессивный анализ
- Переменная

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Андерсон В.Л., Макклин Р.А. Планирование экспериментов: реалистичный подход. Марсель Деккер, НьюЙорк (1974)

Boscovich, RJ et Maire, C.: De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus. Палладис, Рим (1755 г.)

M

Machine Translated by Google

Эйлер, Л.: Исследования по вопросу о desiné galités du Mouvement de Saturne et de Юпитер, получивший премию l'Année 1748, par l'Académie royale des наук Парижа. Republié en 1960, данс Леонарди Эйлери, Opera Omnia, 2ème серия. Турич и, Базе, 25, стр. 47–157 (1749 г.)

Mayer, T.: Abhandlung über dieUmwälzung desMonds um seineAchse und die schein голый Bewegung der Mondfleken. Kos mographische Nachrichten und Samm lungen auf das Jahr 17481, 52–183 (1750)

Де Муавр, Абрахам

Абрахам де Муавр(1667–1754) сначала учился в Седане и Сомюре (оба во Франции). прежде чем продолжить изучение математики и физика в Сорбонне. Приходящий из протестантской семье, он был вынужден покинуть Франция в 1685 году в связи с отменой Нантским эдиктом и был сослан в Лондон.

После тяжелого старта на новой родине он мог заниматься математикой. Он был избран членом Лондонского королевского общества. в 1697 году.

Наиболее важными его работами были «Доктрина шансов», вышедшая в 1718 г. и был посвящен вопросам, которые продолжают играть фундаментальную роль в современной теории вероятностей и Miscellanea Analytica de Seriebus et Quadraturis, где для впервые появилась функция плотности нормального распределения в кон тексте расчетов вероятностей для игр опасности. Карл Фридрих Гаусс взял ту же функциюплотности и применил ее к задачам измерения в астрономии. Он также внес вклад в актуарные науки и математический анализ.

Основные произведения Абрахама де Муавра:

1718 г. Доктрина шансов: или, Метод расчета вероятности событий в игре. Пирсон, Лондон.

1725 год. Аннуитеты на жизнь. Пирсон, Лон Дон.

1730 г. Miscellanea Analytica de Seriebus и др. Квадратурис. Тонсон и Уоттс, Лон Дон.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Теорема де Муавра – Лапласа
Нормальное распределение

Теорема де Муавра – Лапласа.

Эта теорема установила соотношение между биномиальным распределением с параметрыn и р и нормальное распределение для n, стремящееся к бесконечности. Его полезность заключается прежде всего в оценке функция распределения F(k) бинома распределения по функции распределения нормальное распределение, которое оказывается будь проще.

ИСТОРИЯ

Эта теорема восходит к 1870 г., когда Пьер Симон де Лаплас учился вместе с другими (среди них де Муавр, Абрахам), проблемы, связанные с аппроксимацией биномиальное распределение и к теории ошибки.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X1, X2,...,Xn,... —последовательность независимых и одинаково распределенных случайных величин, следующих биномиальному распределениюс параметрами n и р, где 0 < р < 1.

Machine Translated by Google

Рассмотрим соответствующие скорректированные переменная Zn, n = 1, 2,... например:

$$Zn = \frac{Xn - np}{\sqrt{npq}},$$

где np представляет собой среднее значение биномиального распределения, а $\sqrt{npq}$ — его дисперсию. переменная Zn соответствует нормировке Xn).

Теорема де Муавра–Лапласа выражается следующим образом:

$$P(Zn \quad x) \quad n \quad (x),$$

где – функция распределения стандартное нормальное распределение.

Аппроксимацию также можно улучшить с использованием поправочного коэффициента «половина». Отсюда следующее приближение:

$$P(Xn \quad x) = 1 \quad \frac{x - 2\frac{1}{} \quad np}{\sqrt{npq}}$$

лучше, чем полученные без учетом поправочного коэффициента $\frac{1}{2}$.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Эта теорема является частным случаем центральной предельной теоремы. Биномиальное распределение действие (такое, как было определено ранее) может быть записывается в виде суммы n случайных величин dis по бернуллиевскому распределению параметра p . Среднее значение случайной переменной Бернулли равно p и его дисперсия до p (1 p). де Муавр – Теорема Лапласа легко выводится из Центральная предельная теорема.

ПРИМЕРЫ

Мы хотим рассчитать вероятность более чем 27 успехов в биномиальном эксперименте из 100 испытаний, каждое испытание имеет вероятность успеха p = 0,2. Искомый бином вероятность может быть выражена суммой:

$$P(Xn > 27) \quad "=" \quad \sum_{k=27}^{100} \binom{100}{k} \cdot (0,2)^{k} (0,8)^{100-k}.$$

Прямой расчет этой вероятности требует оценки 74 терминов, каждый формы

$$\binom{100}{k} (0,2)^{k} (0,8)^{100-k}.$$

Используя теорему де Муавра–Лапласа, это вероятность можно приблизительно оценить с учетом соотношения между биномиальное и нормальное распределения. Таким образом:

$$P(Xn > 27) = 1 \quad \frac{27 - np}{\sqrt{npq}}$$

$$= 1 - \frac{27 \quad 100 \cdot 0,2}{\sqrt{100 \cdot 0,2 \cdot 0,8}}$$

$$= 1 - (1,75).$$

Обращаясь к обычной таблице, получаем (1,75) = 0,9599, и поэтому:

$$P(Xn > 27) \quad = 0,0401.$$

Точное значение до 4 знаков после запятой вероятность: P (Xn > 27) = 0,0558. Сравнение двух значений 0,0401 и 0,0558 показывает, что аппроксимация нормальное распределение дает результат, который очень близко к точному значению.

Получаем следующее приближение используя поправочный коэффициент:

$$P(Xn > x) \quad = 1 - (1,625) = 0,0521.$$

Результат ближе к точному значению 0,0558 чем полученные без учета поправочного коэффициента (0,0401).

M

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Биномиальное распределение

Центральная предельная теорема

Нормальное распределение

Нормализация

ССЫЛКА

Лаплас, П.С.: Память о приближениях формул, которые выполняют функции трех великих имен и о возможных приложениях. Mémoires de l'Académie Royale des Sciences de Paris, 10. Воспроизведено в: Œuvres de Laplace 12, 301–347 (1810).

Moivre, A. de: Approximatio ad summam terminorum binomii (a + b) n, in seriem expansi. Приложение II к Miscellanae Analytica, стр. 1–7 (1733 г.). Фотографически перепечатано вредкой брошюре о Муавре и некоторых его открытиях. Паб

опубликовано Арчибальдом, RC Isis 8, 671–683 (1926)

Шейнин, О.Б.:Работы П.С.Лапласа о вероятности, Архив истории точных наук, 16, 137–187 (1976).

Шейнин О.Б.: Теория ошибок П.С. Лапласа, Архив истории точных наук, 17, 1–61 (1977).

Момент

Назовем момент порядка k случайной переменной X относительно любого значения x0 средняя разница между случайной величиной и x0 в степени k. Если x0 = 0, мы имеем в виду момент (или начальный момент) порядка k.

Если x0 = μ, где μ —математическое ожидание случайной величины X, то мы имеем в виду центральный момент порядка k.

ИСТОРИЯ

Понятие «момент» в статистике происходит от понятия момента в физике. который происходит от открытия Архимедом (287–212 до н.э.) действия рычага, где мы говорим о моменте сил. Исправлены формулы для расчета моменты сгруппированных данных были адаптированы Sheppard, WF (1886).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Начальный момент порядка k дискретного случайная величина X выражается как:

$$E\ Xk = \sum_{i=1}^{n} x_i^k P(x_i),$$

если он существует, т. е. если $E\ Xk < \infty$, где P (x) – функция вероятности дискретной случайной величины X, принимающей n значений x1,...,xn.

В случае непрерывной случайной величины, если момент существует, он определяется:

$$E\ Xk = \int x^k \phi(x)\, dx,$$

где f (x) —функция плотности непрерывной случайной величины.

При k = 1 начальный момент совпадает с ожидаемым значением случайной величины X (если это E(X) существует):

$$\text{(X дискретный) } E[X] = \sum_{i=1}^{n} x_i P(x_i),$$

$$\text{(X непрерывный) } [X] = \int x\phi(x)\, dx.$$

Центральный момент порядка k дискретного случайная величина X, если она существует, выражается как:

$$E\ (X - E\ [X])^k = \sum_{i=1}^{n} (x_i - E[X])^k P(x_i).$$

Machine Translated by Google

Точно так же в случае непрерывной случайной величины имеем:

$$E\left[(X - E[X])^k\right] = \int_{-\infty}^{+\infty} (x - E[X])^k f(x)\,dx$$

если интеграл корректно определен.

Очевидно, что для любой случайной величины X, имеющей конечное математическое ожидание, центральный момент порядка 1 равен нулю

$$E[X - \mu] = \sum_{i=1}^{n} (x_i - \mu) P(x_i) = \sum_{i=1}^{n} x_i P(x_i) - \mu \sum_{i=1}^{n} P(x_i) = 0$$

знак равно $\mu - \mu$ знак равно 0 .

С другой стороны, если он существует, то центральный момент порядка 2 совпадает с переменной случайной величины:

$$(X \text{ дискретный}) \; E(X - \mu)^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(x_i) = \mathrm{Var}(X),$$

$$(X \text{ непрерывный}) \; E(X - \mu)^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\,dx = \mathrm{Var}(X).$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Оценка момента порядка k кал для выборки x1, x2,...,xn обозначается через mk. Это равно:

$$m_k = \frac{\sum_{i=1}^{n} (x - x_0)^k}{n},$$

где n —общее количество наблюдений.

Если x0 = 0, то имеем момент (или начальный момент) порядка k.

Если $x_0 = \bar{x}$, где $\bar{x}$ —среднее арифметическое, мы имеем центральный момент порядка k.

Для случая, когда случайная величина X принимает свои значения xi с частотами fi, i = 1, 2,...,h, момент порядка k определяется выражением:

$$m_k = \frac{\sum_{i=1}^{h} f_i (x_i - x_0)^k}{\sum_{i=1}^{h} f_i}.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Среднее арифметическое
- Функция плотности
- Ожидаемое значение
- Функция вероятности
- Случайная переменная
- Дисперсия
- Среднее взвешенное арифметическое

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Шеппард, В.Ф.: О вычислении наиболее вероятных значений констант частоты для данных, расположенных в соответствии с равноудаленными делениями шкалы. Лондонская математика. соц. проц. 29, 353–380 (1898 г.)

M

Метод Монте-Карло

Любой численный метод решения математических задач, использующий случайные или псевдослучайные числа, называется методом Монте-Карло. Этот метод. Эти математические или статистические задачи не имеют аналитических решений.

ИСТОРИЯ

Название Монте-Карло происходит от одноименного города в Монако, известного своими казино. Рулетка —один из самых простых механизмов генерации случайных чисел.

Согласно Соболю И.М. (1975), метод Монте-Карло обязан своим существованием американские математики фон Нейман,

Machine Translated by Google

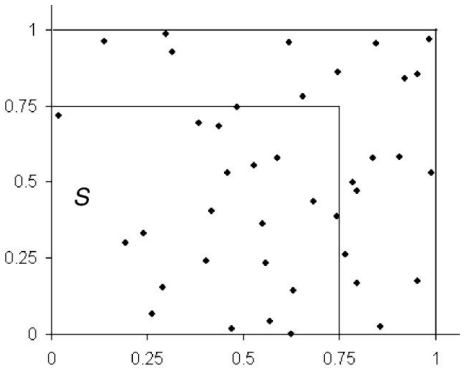
Джон (1951), Улам, Станислав Марцин и Метрополис, Николай (1949). Он был разработан примерно в 1949 году, но стал практичным с изобретением компьютеров.

ПРИМЕРЫ

Предположим, что нам нравится вычислять поверхность фигуры S, расположенной внутри квадрата; его сторона равна 1. Мы сгенерируем N случайных точки внутри квадрата. Для этого отметим квадрат в системе перпендикуляров оси, началом которых является нижний левый угол квадрат. Это означает, что все точки, которые мы будем генерировать внутри квадрата будет иметь координаты между 0 и 1. Достаточно взять равномерные случайные величины, чтобы получить точка. Первое случайное число будет абсцисса, а вторая ордината.

В следующем примере выберем 40 случайные точки. Нам понадобятся два образца 40 случайных чисел. Поверхность S, которую мы хотел бы оценить квадрат 0,75 стороны, другими словами точное значение равняется 0,5625.

После создания 40 случайных точек в на единичный квадрат, достаточно посчитать количество точек внутри фигуры S (21 в этом случай). Оценка искомой поверхности получается по соотношению $\frac{21}{40}$= 0,525 .



Можно улучшить эту оценку, проводя эксперимент многократно и взяв среднее значение полученных поверхностей.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Генерация случайных чисел
- Метод складного ножа
- Случайное число
- Моделирование

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Метрополис, Н., Улам, С.: Монте-Карло метод. Варенье. Стат. доц. 44, 335–341 (1949)

Соболь И.М. Метод Монте-Карло. Мир Издательство, Москва (1975)

Фон Нейман, Дж: Использовались различные методы в связи со случайными цифрами. Симпозиум Национального бюро стандартов, NBS, серия «Прикладная математика» 12, стр. 36–38, Национальное бюро стандартов, Вашингтон, округ Колумбия (1951).

Скользящее среднее

Мы называем скользящее среднее порядка k среднее арифметическое, рассчитанное на k последовательных значения временного ряда. Остальные эти арифметические средние дают ряд скользящих средних порядка k. Процесс, который состоит в замена исходного ряда называется сглаживанием временного ряда.

Что касается временных рядов, состоящих из годовых или месячных данных, то средние значения носят, соответственно, название скользящих средних за k лет или k месяцев. Когда у нас есть месячные данные, например, расчет скользящей средней за 12 месяцев,

Machine Translated by Google

получ енные результаты соответствуют среднему рассматриваемого периода, в конце шестом месяце (или в первый день седьмого месяца), а не в середина месяца, как и в исходных данных. Мы исправим это, сделав центрированное движение в среднем за 12 мес. Как правило, вычисляется центрированное скользящее среднее порядка k. построение скользящей средней 2-го порядка скользящее среднее порядка k нач ального времени ряд.

ИСТОРИЯ

Смотрите временной ряд.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Имеем временные ряды Y1, Y2, Y3, ..., YN, и мы определяем серию скользящих средних порядка k рядом средних арифметических:

$$\frac{Y1 + Y2 + \cdots + Yk}{k},$$

$$\frac{Y2 + Y3 + \cdots + Yk+1}{k}, \ldots,$$

$$\frac{YN\quad k+1 + YN\quad k+2 + \cdots + YN}{k}.$$

Сумма числителей называется скользящей суммы порядка k.

Мы можем использовать взвешенные средние арифметические с заранее указанными весами; остальная часть то, что получается таким образом, называется рядом взвешенных скользящих средних порядка k.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Скользящие средние используются при изучении временных рядов для оценки вековой тенденции, сезонных изменений и циклические колебания.

Применение операции «скользящее среднее» к временному ряду позволяет:

- Определить сезонные колебания пределы, где они являются строго периодическими,
- Сглаживание неравномерных вариаций и
- Примерно сохранить дополнительный сезон все движение.

Отметим недостатки метода скользящей средней:

- Данные начала и окончания серии отсутствуют.
- Скользящие средние могут вывать циклы или другие движения, отсутствующие в оригинале. данные.
- Скользящие средние находятся под сильным влиянием по выбросами и случайным значениям.
- Когда график исследуемого временного ряда демонстрирует экспоненциальную восходящую тенденцию или, в более общем смысле, сильная кривая, метод скользящей средней не дает очень точные результаты для оценки вне сезонное движение.

ПРИМЕРЫ

С данными временного ряда 7, 2, 3, 4, 5, 0, 1, получаем скользящие средние третьего порядка:

$$\frac{7 + 2 + 3}{3}, \quad \frac{2 + 3 + 4}{3}, \quad \frac{3 + 4 + 5}{3},$$

$$\frac{3\,4 + 5 + 0}{3}, \quad \frac{3\,5 + 0 + 1}{3},$$

то есть 4, 3, 4, 3, 2.

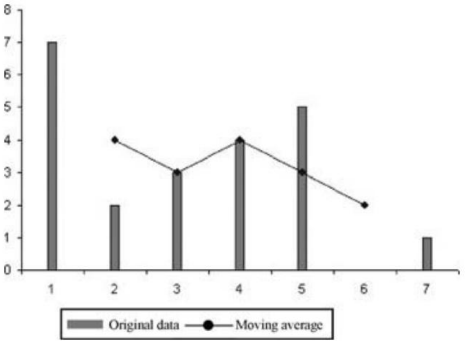
В ряду скользящих средних обычно локализовать каждое число на его относительной позиции к исходным данным. В этом примере мы писать:

Исходные данные	7, 2, 3, 4, 5, 0, 1
Скользящее среднее порядка 3	4, 3, 4, 3, 2

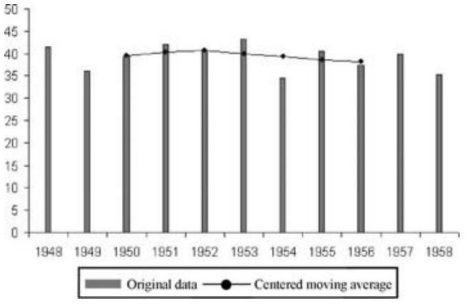
M

Machine Translated by Google

На приведенном ниже графике каждое число ряда скользящих средних представляет собой среднее арифметическое три числа непосредственно под ним.



Мы предлагаем рассчитать центрированное скользящее среднее за 4 года среднемесячного добычи нефти в стране. На первом шаге рассчитываем скользящие суммы за 4 года, добавление данных за 4 последовательных года. Затем, разделив каждую скользящую сумму на 4, получим скользящие средние за 4 года, последняя из которых соответствует 1 января третьего года, учитываемого в разных суммах. Наконец, мы получаем скользящие средние за 4 года, расчет среднее арифметическое двух скользящих средних в течение 4 лет подряд.



Среднемесячная добыча нефти в стране в миллионы тонн

Перемещение	дата	сумма за 4 года	движущийся средний более 4	по центру движущийся средний более 4
	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		

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Среднее арифметическое

Временная последовательность

Среднее взвешенное арифметическое

Мультиколлинеарность

Мультиколлинеарность — это термин, используемый для описания ситуация, когда переменная почти равна линейной комбинации других переменных.

Мультиколлинеарность иногда присутствует в проблемы регрессии, когда есть сильный корреляция между объясняющими переменными.

Machine Translated by Google

Существование линейного или почти линейного отношения

Между независимыми переменными возникает большая неточность в оценке коэффициентов регрессии . Если отношения строго линейны, коэффициенты будут смещены.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

См. регрессию и гребня.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коэффициент корреляции

Множественная линейная регрессия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Белсли, Д., Кух, Э ., Уэлш, Р.Э .: Диагностика регрессии Идентификация влиятельных данных и источников коллинеарности. Уайли, Нью-Йорк (1976)

Мультимодальная дистрибуция

См. распределение частот.

Полиномиальное распределение

Рассмотрим случайный эксперимент с n независимыми испытаниями. Предположим, что каждое испытание может дать только одно из событий Ei с вероятностью pi.
Случайная величина Xi представляет собой количество появлений события Ei.
Полиномиальное распределение с параметрами n и pi определяется следующей совместной функцией плотности:

$$P(X1 = x1, X2 = x2, \ldots, Xs = xs) = \frac{n!}{x1!\,x2!\cdots xs!}\,p1^{x1}\cdot p2^{x2}\cdots ps^{xs},$$

с целыми числами xi 0 такими, что $\sum_{i=1}^{s} xi = n$.
Действительно, вероятность получения последовательность

$$\underbrace{E1,\ldots,E1}_{x1};\ \underbrace{E2,\ldots,E2}_{x2};\ldots;\underbrace{Es,\ldots,Es}_{xs}$$

равно $\prod_{i=1}^{s} pi^{xi}$.
Количество существующих возможностей — это количество перестановок n объектов, из которых x1 — событие E1, x2 — событие E2. . . , xs — событие Es, означающее, что:

$$\frac{n!}{\prod_{i=1}^{s} xi!}.$$

Полиномиальное распределение является дискретным распределением вероятностей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Для полиномиального распределения

$$\sum_{i=1}^{s} pi = 1 \text{ и } \sum_{i=1}^{s} xi = n,$$

с $0 < pi < 1$.
Ожидаемое значение рассчитывается для каждой случайной величины Xi.
Таким образом, мы имеем:

$$E[Xi] = n \cdot pi .$$

Действительно, каждая случайная величина может быть sideredindividually как после абиномиального распределения.
То же самое можно сделать и для переменной анс. Тогда он равен:

$$Var(Xi) = n \cdot pi \cdot qi .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Полиномиальное распределение имеет множество применений в анализе статистических данных.

Machine Translated by Google

данные. Он используется особенно в ситуациях, когда данные должны быть классифицированы по нескольким категориям. истории событий (например, при анализе категорийных данных).

ПРИМЕРЫ

Рассмотрим следующий случайный эксперимент:

В урне девять шаров: два белых, три красных и четыре черных. Мяч выбирается случайным образом, отмечают его цвет и кладут обратно в урну. Это повторяется 5 раз.

Мы можем описать этот случайный эксперимент следующим образом:

У нас есть три возможных события:

E1: выбор белого шара

E2: выбор красного шара

E3: выбор черного шара

и три случайные величины:

X1: количество выбранных белых шаров

X2: количество выбранных красных шаров

X3: количество выбранных черных шаров

Вероятности, связанные с этими тремя события:

$$p1 = 9/3$$

$$p2 = 9/4$$

$$p3 = 9/2.$$

Подсчитаем вероятность того, что из пяти выбранных шаров (n = 5) один окажется белым, два — красными и два — черными.

Количество возможностей равно:

$$5!/(1!2!2!).$$

Вероятность получения последовательности (w, r, r, b, b), где w, r и b — соответственно белые, красные и черные шары, равна:

$$\prod_{i=1}^{3} p_i^{x_i} = \left(\frac{2}{9}\right)^1 \cdot \left(\frac{3}{9}\right)^2 \cdot \left(\frac{4}{9}\right)^2 = 0{,}0049.$$

Следовательно,

$$P(X1 = x1, X2 = x2, X3 = x3) = \frac{5!}{x_1!\,x_2!\,x_3!} \cdot \left(\frac{2}{9}\right)^{x1} \cdot \left(\frac{3}{9}\right)^{x2} \cdot \left(\frac{4}{9}\right)^{x3},$$

где

$$P(X1 = 1, X2 = 2, X3 = 2) = 30 \cdot 0{,}0049 = 0{,}147.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Биномиальное распределение
- Дискретное распределение вероятностей

Множественный коэффициент корреляции

См. коэффициент корреляции.

Множественная линейная регрессия

Регрессионный анализ, в котором независимая переменная Y линейно зависит от многих независимых переменных X1, X2,...,Xk, называется множественной линейной регрессией.

Уравнение множественной линейной регрессии имеет вид:

$$Y = f(X1, X2,...,Xk),$$

где f (X1, X2,...,Xk) — линейная функция от X1, X2,...,Xk.

ИСТОРИЯ

См. регрессионный анализ.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Общая модель множественной линейной регрессии, содержащая k = (p 1) независимых переменных (и n наблюдений), записывается как:

$$Yi = \beta 0 + \sum_{j=1}^{p-1} Xji\beta j + \varepsilon i, \qquad i = 1,...,n,$$

Machine Translated by Google

где Yi — зависимая переменная, Xji, j = 1,...,p 1, — независимые переменные, βj, j = 0,...,p 1, — оцениваемые параметры, εi — член случайной наблюдаемой ошибки.

В матричной форме эта модель записывается так:

$$Y = X\beta + \varepsilon \quad ,$$

где Y — вектор (n×1) наблюдений, относящихся к зависимой переменной (n наблюдений), β — вектор (p×1) оцениваемых параметров, ε — вектор (n×1) ошибок,

$$X = \begin{matrix} 1 & X11 & \dots & X1(p\ \ 1) \\ \vdots & \vdots & & \vdots \\ 1 & Xn1 & \dots & Xn(p\ \ 1) \end{matrix}$$

это (n×p) матрица независимых переменных.

Оценка вектора β

Начиная с модели

$$Y = X\beta + \varepsilon \quad ,$$

получаем оценку $\hat{\beta}$ вектора β методом наименьших квадратов:

$$\hat{\beta} = (X\ X)\quad 1X\ Y \quad ,$$

и мы вычисляем оценочное значение Y для Y:

$$Y \text{ знак равно } X \cdot \hat{\beta} .$$

На этом шаге мы можем вычислить невязки, обозначаемые вектором e, которые мы находим следующим образом:

$$e = Y - Y \quad .$$

Мера достоверности оценки Y

Чтобы знать, какой мере доверять чо

Для линейной модели полезно провести дисперсионный анализ и проверить гипотезы о векторе β регрессионной модели. Для проведения этих тестов мы должны сделать следующие предположения:

- Для каждого значения Xji и для всех i = 1,...,n и j = 1 ...,p 1 существует случайная переменная Y, распределенная согласно нормальному распределению
- Дисперсия Y одинакова для всех Xji ; оно равно σ2 (не известно).
- Различные наблюдения над Y не зависят друг от друга, но обусловлены значениями Xji.

Дисперсионный анализ в матричной форме

В матричной форме таблица дисперсионного анализа для регрессии выглядит следующим образом:

Дисперсионный анализ

Источник из вариация	Степени свобода	Сумма квадраты	Среднее значение квадраты
Регрессия	p - 1	$\hat{\beta}\ XY \quad ny^{-}2$	$\frac{\hat{\beta}\ XY \quad nY\hat{}\ 2}{p \quad 1}$
Остаток	n - p	$YY \quad \hat{\beta}XY\ _$	$\frac{YY \quad \hat{\beta}XY\ _}{n - p}$
Общий	n - 1	$Y\ Y \quad ny^{-}2$	

Если модель правильная, то

$$S2 = \frac{YY \quad \hat{\beta}XY\ _}{n - p}$$

является несмещенной оценкой σ2 .

Дисперсионный анализ позволяет проверить нулевую гипотезу:

$$H0 : \beta j = 0 \text{ для } j = 1,...,p \quad 1$$

против альтернативной гипотезы:

H1: хотя бы один из параметров βj, j = 0, отличен от нуля

расчет статистики:

$$Ф = \frac{Э\ МСЭ}{СКО} \text{ "=" } \frac{Э\ МСЭ}{S2} ,$$

Machine Translated by Google

где EMSE — среднее квадратов регрессии, RMSE — среднее квадратов остатков, а TMSE — общее среднее квадраты.

Эту F-статистику необходимо сравнить со значением Fα,p 1,n p таблицы Фишера, где α — уровень значимости теста:

если F Fα,p 1,n p, принимаем H0;
если F > Fα,p 1,n p, мы отказываемся от H0 в пользу H1.

Коэффициент детерминации R2 рассчитывается следующим образом:

$$P2 = \frac{E\,CC}{TCC} \text{ "=" } \frac{\hat{\beta}\,XY \quad ny^{-}2}{YY \quad ny^{-}2},$$

где ESS — сумма квадратов регрессии, а TSS — общая сумма квадратов.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

См. анализ регрессии.

ПРИМЕРЫ

Следующие данные касаются десяти компаний химической промышленности. Мы пытаемся установить связь между производством, рабочим временем и капиталом.

Производственная работа		Капитал
(100 тонн)	(час)	(машино-часы)
60	1100	300
120	1200	400
190	1430	420
250	1500	400
300	1520	510
360	1620	590
380	1800	600
430	1820 г.	630
440	1800	610
490	1750	630

Матричная модель записывается как:

$$Y = X\beta + \varepsilon \quad ,$$

где

$$Y = \begin{matrix} 60 \\ 120 \\ 190 \\ 250 \\ 300 \\ 360 \\ 380 \\ 430 \\ 440 \\ 490 \end{matrix}, \quad X = \begin{matrix} 1\ 1100\ 300 \\ 1\ 1200\ 400 \\ 1\ 1430\ 420 \\ 1\ 1500\ 400 \\ 1\ 1520\ 510 \\ 1\ 1620\ 590 \\ 1\ 1800\ 600 \\ 1\ 1820\ 630 \\ 1\ 1800\ 610 \\ 1\ 1750\ 630 \end{matrix},$$

$$e = \begin{matrix} \varepsilon 1 \\ \varepsilon 2 \\ \varepsilon 3 \\ \varepsilon 4 \\ \varepsilon 5 \\ \varepsilon 6 \\ \varepsilon 7 \\ \varepsilon 8 \\ \varepsilon 9 \\ \varepsilon 10 \end{matrix}, \quad \beta = \begin{matrix} \beta 0 \\ \beta 1 \\ \beta 2 \end{matrix}.$$

Уравнения, формирующие модель, следующие:

$$60 = \beta 0 + 1100\beta 1 + 300\beta 2 + \varepsilon 1,$$
$$120 = \beta 0 + 1200\beta 1 + 400\beta 2 + \varepsilon 2,$$
$$\ldots$$
$$490 = \beta 0 + 1750\beta 1 + 630\beta 2 + \varepsilon 10.$$

Мы вычисляем оценки $\hat{\beta}$, используя результат:

$$\hat{\beta} = X\,X \quad {}^{1}\,XY \quad ,$$

с

$$(X\ X) = \begin{matrix} 10 & 15540 & 5090 \\ 15540 & 24734600 & 8168700 \\ 5090 & 8168700 & 2720500 \end{matrix},$$

$$(X\ X)^{1} \text{ "=" } \begin{matrix} 6{,}304288 & 0{,}007817 & 0{,}011677 \\ 0{,}007817 & 0{,}000015 & 0{,}000029 \\ 0{,}011677 & 0{,}000029 & 0{,}000066 \end{matrix},$$

Machine Translated by Google

и

$$XY = \begin{matrix} 3020 \\ 5012000 \\ 1687200 \end{matrix} .$$

Это дает:

$$\hat{\beta} = \begin{matrix} 439,269 \\ 0,283 \\ 0,591 \end{matrix} \overset{\hat{\beta}_0}{=} \begin{matrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{matrix} .$$

Таким образом, получаем уравнение для расчетной линии:

$$\hat{Y}_i = 439,269 + 0,283X_{i1} + 0,591X_{i2} .$$

Дисперсионный анализ в матричной форме

Расчет различных элементов таблица дает нам 1. Расчет степеней свободы:

$$dlreg = p - 1 = 3 - 1 = 2 ,$$
$$dlres = n - p = 10 - 3 = 7 ,$$
$$dltot = n - 1 = 10 - 1 = 9 ,$$

где dlreg — степени свободы регрессии, dlres — степени свободы остатков, dltot — полные степени свободы.

2. Расчет суммы квадратов:

$$ESS = \hat{\beta}' XY - n\bar{Y}^2 = [439,269 \; 0,283 \; 0,591] \cdot \begin{matrix} 3020 \\ 5012000 \\ 1687200 \end{matrix} - 912040$$
$$= 179063,39 .$$

$$TSS = YY - n\bar{Y}^2 = [60 \ldots 490] \cdot \begin{matrix} 60 \\ \vdots \\ 490 \end{matrix} - 912040$$
$$= 187160 .$$

$$RSS = YY - \hat{\beta}' XY$$
$$= TCC - ESS$$
$$= 187160,00 - 179063,39$$
$$= 8096,61 .$$

Получаем следующую таблицу:

Дисперсионный анализ

Источник вариации	Степень свободы дом	Сумма квадраты	Среднее значение квадраты
Регрессия	2	179063,39	89531,70
Остаток	7	8096,61	1156,66 = S2
Общий	9	187160.00	

Это позволяет нам проверить нулевую гипотезу:

$$H0 : \beta 1 = \beta 2 = 0 .$$

Расчетное значение F: _

$$\Phi = \frac{ESS}{CKO} \overset{=}{} \frac{ESS}{S2} \overset{=}{} \frac{89531,70}{1156,66}$$
$$= 77,41 .$$

Значение F в таблице Фишера с уровнем значимости α = 0,05 равно

$$F0,05,2,7 = 4,74 .$$

В результате нулевая гипотеза H0 отвергается. Альтернативная гипотеза может быть одной из трех следующих возможностей:

$$H1 : \beta 1 = 0 \text{ и } \beta 2 = 0$$
$$H2 : \beta 1 = 0 \text{ и } \beta 2 = 0$$
$$H3 : \beta 1 = 0 \text{ и } \beta 2 = 0$$

M

Machine Translated by Google

Можно рассчитать коэффициент детерминации R2:

$$R2 = \frac{ESS}{TSS} = \frac{179063,39}{187160.00} = 0,96 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Анализ остатков
- Дисперсионный анализ
- Коэффициент детерминации
- Коллинеарность
- Коэффициент корреляции
- Матрица шляпы
- Наименьших квадратов
- Точка рычага
- Нормальные уравнения
- Регрессивный анализ
- Остаток
- Простая линейная регрессия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Вайсберг, С.: Прикладная линейная регрессия. Уайли, НьюЙорк (2006)

Чаттерджи, П., Хади, А.С.: Регрессионный анализ на примере. Wiley, НьюЙорк (2006)

Machine Translated by Google

Отрицательное биномиальное распределение

Случайная величина X следует за отрицательным биномиальное распределение с параметрами sk и r, если его функция вероятности имеет вид:

$$P(X = k) = C_{k+r-1}^{k} \cdot p^{r} \cdot q^{k},$$

где p — вероятность успеха, q — вероятность отказа и C_{n}^{k} это число комбинаций k объектов среди n.
В ситуации, когда «успех-неудача» случайный эксперимент повторяется независимым образом, вероятность успеха равна обозначается p, а вероятность отказа q = 1 p. Эксперимент повторяется столько раз, сколько требуется для получить r успехов. Количество полученных количество неудач до достижения этой цели является случайной величиной, подчиняющейся отрицательному биномиальному распределению, описанному выше. Отрицательный биномиальное распределение является дискретным биномиальным распределение.

ИСТОРИЯ

Первыми, кто рассмотрел отрицательное биномиальное распределение, были Паскаль, Блез (1679). Тогда де Montmort, PR (1714 г.) применил отрицательный биномиальное распределение, представляющее количество раз, которое нужно подбросить монету получить определенное количество голов. Стьюдент (1907) использовали отрицательное биномиальное распределение как альтернатива распределениюПуассона. бутион.
Гринвуд, М. и Юл, Г.У. (1920) и Эггенбергер Ф. и Поля Г. (1923) обнаружили приложения отрицательного биномиального распределения. С тех пор число применений этого распределения растет, и статистические методы, основанные на на этом дистрибутиве были разработаны в параллельный путь.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если X1, X2,...,Xn —n независимых случайных переменных, следующих геометрическому распределению, то случайная величина

$$X = X1 + X2 + \ldots + Xr$$

следует отрицательному биномиальному распределению.
Для вычисления ожидаемого значения X используется следующее свойство, где Y и Z случайные переменные:

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

Таким образом, мы имеем:

$$E[X] = E[X1 + X2 + \ldots + Xr]$$

$$= E[X1] + E[X2] + \ldots + E[Xr]$$

$$= \frac{r}{p} + \frac{r}{p} + \ldots + \frac{r}{p}$$

$$= r \cdot \frac{1}{p}.$$

Machine Translated by Google

Для вычисления дисперсии X используется следующее свойство, где Y и Z — независимые переменные:

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

Таким образом, мы имеем:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(X1 + X2 + \ldots + Xr)\\ &= \mathrm{Bap}(X1) + \mathrm{Bap}(X2) + \ldots\\ &\quad + \mathrm{Bap}(Xp)\\ &= \frac{Д}{} + \frac{Д}{} + \ldots + \frac{Д}{} \\ &= p \cdot \frac{Д}{p2} .\end{aligned}$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Среди конкретных областей, для которых было применено отрицательное биномиальное распределение статистика несчастных случаев, биологические науки, экология, маркетинговые исследования, медицинские исследования, и психология.

ПРИМЕРЫ

Монета подбрасывается несколько раз. Нас интересует вероятность выпадения орла в третий раз на шестом броске.

Тогда мы имеем:

Количество успехов: r = 3

Количество отказов: k = 6 - p = 3

Вероятность одного успеха: $p = \frac{1}{2}$ (получение головы)

Вероятность одного отказа: $q = \frac{1}{2}$ (получение хвостов)

Вероятность выпадения k решек до третья голова дается:

$$P(X = k) = C_{3+k-1}^{k} \cdot \left(\frac{1}{2}\right)^{3} \cdot \left(\frac{1}{2}\right)^{k} .$$

Вероятность выпадения третьей головы на шестой бросок, означающий вероятность получение решки три раза до получения третья головка, следовательно, равна:

$$\begin{aligned}P(X = 3) &= C_{3+3-1}^{3} \cdot \frac{1}{23} \cdot \frac{1}{23}\\ &= C_{5}^{3} \cdot \frac{1}{23} \cdot \frac{1}{23}\\ &= \frac{5!}{3!(5-3)!} \cdot \frac{1}{23} \cdot \frac{1}{23}\\ &= 0{,}1563 .\end{aligned}$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение Бернулли
- Биномиальное распределение
- Дискретное распределение вероятностей
- распределение Пуассона

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Эггенбергер, Ф. и Поля, Г.: Уберу мереть Statistik verketteter Vorgänge. Zeitschrift für angewandte Mathematis che Mechanik 3, 279–289 (1923)

Гринвуд, М., Юл, Г.У.: Расследование характерчастотных распределений, представляющих множественные события с особая ссылка на возникновение множественные приступы болезней или повторные несчастные случаи. Дж. Рой. Стат. соц. сер. А 83, 255–279 (1920)

Montmort, PR de: Essai d'analyse sur les jeux de hasard, 2-е изд. Кийо, Париж (1713)

Паскаль, Б.: Varia Opera Mathematica. Д. Петри де Ферма. Толосы (1679 г.)

Госсет, С.В. («Студент»): Об ошибке подсчета гемоцитометром. Биометрика 5, 351–360 (1907).

Machine Translated by Google

Нелдер, Джон А.

Джон Нелдер родился в 1924 году в Далвере.
тонн (Сомерсет, Англия). После получения его
диплом по математической статистике в Кэмбридже,
он был назначен президентом секции статистики
Национальной овощной
Исследовательская станция в Уэллсборне, позиция
он занимал с 1951 по 1968 год. Он заработал
звание доктора наук в Бирмингеме
затем был избран президентом статистического
отдела Ротамстедской экспериментальной лаборатории.
Станция в Харпендене с 1968 по 1984 год. Он
ныне приглашенный профессор Императорского
колледжа науки, техники и медицины в
Лондон. Он был избран членом Королевского
общества в 1984 году и был президентом двух
престижные общества: Международное биометрическое
общество и Королевское статистическое общество.

Нелдер является автором статистических вычислительных
систем Genstat и GLIM, используемых в настоящее время.
в более чем 70 странах. Автор более
более 140 статей, опубликованных в статистических и
биологических обзоров, он также является автором
книги «Компьютеры в биологии» и соавтором с Питером
МакКаллах в книге, посвященной обобщенному
линейная модель. Нельдер также развил идею
обычно сбалансированных конструкций.

Избранные произведения и статьи Джона А.
Нелдер:

1965 (совместно с Мидом Р.) Симплексный метод.
для минимизации функции. Вычислительный
журнал, 7, 303–333.
1971 Компьютеры в биологии. Уайкхэм,
Лондон и Винчестер, стр. 149.
1972 г. (совместно с Wedderburn, RWM) Общие линейные
модели. Дж Рой. Стат. соц.
сер. А 135, 370–384.
1974 Genstat: статистическая система. В:
COMPSTAT: Труды в Com
путационная статистика. Физика, Вена
нет
1983 г. (совместно с Маккаллахом, П.) Обобщенный
Линейные модели. Чепмен и Холл,
Лондон, стр. 261.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Статистическое программное обеспечение

Ньюкомб, Саймон

Ньюкомб, Саймон родился в 1835 году в Нове.
Шотландии (Канада) и умер в 1909 г. Этот астроном
канадского происхождения внес свой вклад в
обработки выбросов в статистике, применения теории
вероятностей для интерпретации данных и разработки
того, что
сегодня мы называем робастными методами в статистике.
До 16 лет он учился в основном по
сверяясь с книгами, которые его отец нашел для
ему. После этого он начал изучать медицину
с планами стать врачом. В 1857 году он
занимался американскими эфемеридами и
Морской альманах в Кембридже в штате
Массачусетса. В то же время он учился в Гарварде,
который окончил в 1858 году.
1861 г., он стал профессором математики.
в Военно-морской обсерватории.
Сборник заметок по теории
Вероятности (Mathematics Monthly, 1859–1859 гг.).
61) представляет собой произведение, которое и
сегодня считается современным. Самое замечание Ньюкомба
способный вклад в статистику - его подход
к обработке выбросов в астрономии
данные; утверждая, что нормальное распределение
не подходит, он изобрел нормальное загрязненное
распределение.

Избранные произведения и статьи Ньюкомба,
Саймон:

1859–61 Заметки по теории вероятностей.
Ежемесячник по математике, 1, 136–

Machine Translated by Google

139, 233–235, 331–355, 349–350; 2,
134–140, 272–275; 3, 68, 119–125,
341–349.

Нейман, Е жи

Нейман, Е жи (1894–1981) был одним из основоположники современной статистики. Один из его установление выдающихся вкладов, с Пирсоном, Э гоном Шарпом, основа теория проверки гипотез. Родился в России, Нейман, Е жи изучал физику и математику в Харьковском университете. В 1921 году он уехал в Польшу, страну свое го происхождения, где некоторое время работал статистиком в Национальном институте сельского хозяйства в Быдгоще. В 1924 году он провел некоторое время в Лондоне, где он смог учиться в Университетском колледже под руководством Пирсона, Карла. В э то время он встретил Пирсона, Э гона Шарп Госсет, Уильям Сили и Фишер, Рональд Э йлмер. В 1937 году, в конце после поездки в Соединенные Штаты, где он выступал с докладами на многих конференциях, он принял должность профессора Калифорнийского университета в Беркли. Он создал кафедрой статистики и завершил свою блестящую карьеру в э том университете.

Избранные статьи Неймана, Е жи:

1928 г. (совместно с Пирсоном, Э .С.) Об использовании и интерпретация некоторых тестовых критериев для целей статистического вывода, I, II. Биометрика 20A, 175–240, стр. 263–295.

1933 (совместно с Пирсоном, Э .С.) Проверка статистических гипотез по отношению к априорным вероятностям. проц. Камб. Фил ос. соц. 29, 492–510.

1934 г. О двух различных аспектах репрезентативный метод: метод стратифицированная выборка и метод целенаправленный отбор. Дж Рой. Стат. соц. 97, 558–606, обсуждение, стр. 607–625.

1938 г. Вклад в теорию выборки человеческих популяций. Варенье. Стат. доц. 33, 101–116.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы

Нелинейная регрессия

Анализ регрессии, в котором зависимая переменная Y зависит от одного или многих независимых переменных X1,...,Xk называется нелинейная регрессия, если уравнение регрессия Y = f X1,...,Xk; β0,..., βр есть нелинейна по параметрам β0,..., βр.

ИСТОРИЯ

Нелинейная регрессия восходит к 1920-м годам. Фишеру, Рональду Э йлмеру и Макензи, В. А. Однако использование и более подробное исследование э тих моделей пришлось отложить за достижения в области автоматических расчетов в 1970-е годы

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть Y1, Y2,...,Yr и X11,...,X1r, X21,...,X2r,..., Xk1,...,Xkr —r наблюдений (соответственно) зависимой переменная Y и независимые переменные X1,...,Xк.

Цель состоит в том, чтобы оценить параметры β0,..., βр модели:

$$Yi = f X1i,...,Xki; \beta 0,..., \beta p + \varepsilon i, \quad я = 1,...,r.$$

Machine Translated by Google

Методом наименьших квадратов и при соответствующей гипотезе об ошибках (см. анализ невязок) оценивает параметры β0,..., βp значениями β^ 0,..., β^, минимизирующими сумму квадратов ошибки, обозначаемые S β0,..., βp =

$$\sum_{i=1}^{p} \left(Y_i - f(X1_i,\ldots,Xk_i;\ \beta_0,\ldots,\beta_p) \right)^2 :$$

$$\min_{\beta_0,\ldots,\beta_p} \varepsilon_i^2 = \min_{\beta_0,\ldots,\beta_p} S(\beta_0,\ldots,\beta_p) . \quad \sum_{i=1}^{p}$$

Iff является нелинейным, разрешение нормальных уравнений (ни одно из которых не является линейным):

$$\sum_{i=1}^{p} \left(Y_i - f(X1_i,\ldots,Xk_i;\ \beta_0,\ldots,\beta_p) \right) \cdot \frac{\partial f(X1_i,\ldots,Xk_i;\ \beta_0,\ldots,\beta_p)}{\partial \beta_j} = 0$$

(j = 0,...,p) может быть затруднено.

Проблема в том, что данные уравнения не обязательно имеют решение или имеют более одного решения.

Далее мы обсудим различные подходы к решению проблемы.

1. Гаусс–Ньютон.

   Процедура, имеющая итерационный характер, состоит в линеаризации f с помощью разложения Тейлора при последовательном оценивании параметров методом линейной регрессии.

   Мы разработаем этот метод более подробно на следующем примере.

2. Самый крутой спуск

   Идея этого метода состоит в том, чтобы приближенно зафиксировать параметры βj для оценок, а затем добавить аппроксимацию

$$\frac{\partial S(\beta_0,\ldots,\beta_p)}{\partial \beta_j}$$

   что соответствует максимуму спуска с функции S.

   Итеративно определяем параметры, минимизирующие S.

3. Метод Маркара–Левенбурга. Этот метод, не развиваемый здесь, объединяет преимущества двух упомянутых процедур, ввывающих и ускоряющих сходимость аппроксимаций оцениваемых параметров.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В множественной линейной регрессии есть функции, которые не являются линейными по параметрам, но становятся линейными после преобразования. Если это не так, мы называем эти функции нелинейными и трактуем их как нелинейные. методы.

Пример нелинейной функции:

$$Y = \frac{\beta_1}{\beta_1 - \beta_2} \exp(-\beta_2 X) - \exp(-\beta_1 X) + \varepsilon .$$

ПРИМЕРЫ

Пусть P1 = (x1, y1), P2 = (x2, y2) и P3 = (x3, y3) —три точки на графике, а d1, d2 и d3 —аппроксимативные расстояния («приближения» распределены согласно распределение Гаусса той же дисперсии) между точками данных и точкой P = (x, y) неизвестных и искомых координат.

Возможная регрессионная модель для оценки координат P выглядит следующим образом:

$$d_i = f(x_i, y_i; x, y) + \varepsilon_i , \qquad i = 1, 2, 3 ,$$

где функция f представляет собой расстояние между точкой Pi и точкой P, то есть:

$$f(x_i, y_i; x, y) = \sqrt{(x - x_i)^2 + (y - y_i)^2} .$$

Machine Translated by Google

Функция расстояния явно нелинейна и не может быть преобразована в линейную форму, поэтому соответствующая регрессия является нелинейной. Применим метод Гаусса–Ньютона. Рассматривая f как функцию оцениваемых параметров x и y, тейлоровское развитие f до тех пор, пока оно не станет линейным членом в (x0, y0), определяется выражением:

$$f(x_i, y_i; x, y) \approx f(x_i, y_i; x_0, y_0) + \left(\frac{\partial f(x_i, y_i; x, y)}{\partial x}\right)_{\substack{x=x_0 \\ y=y_0}} \cdot (x - x_0) + \left(\frac{\partial f(x_i, y_i; x, y)}{\partial x}\right)_{\substack{x=x_0 \\ y=y_0}} \cdot (y - y_0),$$

где

$$\left(\frac{\partial f(x_i, y_i; x, y)}{\partial x}\right)_{\substack{x=x_0 \\ y=y_0}} = \frac{x_i - x_0}{f(x_i, y_i; x_0, y_0)}$$

и

$$\left(\frac{\partial f(x_i, y_i; x, y)}{\partial x}\right)_{\substack{x=x_0 \\ y=y_0}} = \frac{y_i - y_0}{f(x_i, y_i; x_0, y_0)}.$$

Solet(x0, y0) — первая оценка P, найденная, например, геометрически.

Таким образом, модель линеаризованной регрессии выражается следующим образом:

$$d_i = f(x_i, y_i; x_0, y_0) + \frac{x_i - x_0}{f(x_i, y_i; x_0, y_0)} \cdot (x - x_0) + \frac{y_0}{f(x_i, y_i; x_0, y_0)} \cdot (y - y_0) + \varepsilon_i, \quad i = 1, 2, 3.$$

С помощью множественной линейной регрессии (без констант) можно оценить параметры $(\Delta x, \Delta y) = (x - x_0, y - y_0)$ следующей модели («наблюдения» модели указаны в скобках):

$$(d_i - f(x_i, y_i; x_0, y_0)) = \Delta x \cdot \frac{x_i - x_0}{f(x_i, y_i; x_0, y_0)} + \Delta y \cdot \frac{y_i - y_0}{f(x_i, y_i; x_0, y_0)} + \varepsilon_i, \quad i = 1, 2, 3.$$

Позволять $\hat{x}$, $\hat{y}$ — найденная оценка (x, y). Таким образом, мы получаем лучшую оценку координат точки P по формуле $x_0 + \Delta x, y_0 + \Delta y$.

Используя эту новую точку в качестве отправной точки, мы можем снова использовать ту же процедуру. Остальные координаты находятся сходящимся путем к искомой точке, т. е. к точке, которая лучше всего подходит.

Возьмем теперь конкретный пример: P1 = (0, 5); P2 = (5, 0); P3 = (10, 10); д1 = 6; д2 = 4; д3 = 6.

Если в качестве начальной точки процедуры Гаусса–Ньютона выбрать следующую точку: $P = \left(\frac{0+5+10}{3}, \frac{5+0+10}{3}\right) = (5, 5)$, то получим 3 набор следующих координат:

(5, 5)

(6,2536, 4,7537)

(6,1704, 4,7711)

(6,1829, 4,7621)

(6,1806, 4,7639)

сходится к точке (6.1806,4.7639). Расстояния между найденной точкой и точками P1, P2 и P3 равны соответственно 6,1856, 4,9078 и 6,4811.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Наименьших квадратов

Machine Translated by Google

Модель

Множественная линейная регрессия

Нормальные уравнения

Регрессивный анализ

Простая линейная регрессия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бейтс Д.М., Уоттс Д.Г.: Нелинейный регрессионный анализ и его приложения. Уайли, НьюЙорк (1988)

Дрейпер, Н.Р., Смит, Х.: Прикладной регрессионный анализ, 3-е изд. Уайли, НьюЙорк (1998)

Фишер, Р.А.: О математических основах теоретической статистики. Фил ос. Транс. Рой. соц. Лонд. сер. А 222, 309–368 (1922)

Фишер, Р.А., Маккензи, В.А.: Реакция разных сортов картофеля на мануриал. Дж Сельское хозяйство. наук. 13, 311–320 (1923)

Флетчер, Р.: Практические методыоптимизации. Уайли, НьюЙорк (1997)

Себер А.Ф, Уилт С.К.: Нелинейная регрессия. Уайли, НьюЙорк (2003)

Не параметрическая статистика

Статистические процедуры, которые позволяют нам обрабатывать данные из небольших выборок, по переменным о котором нич его не известно относительно их распространение. В частности, непараметрические статистические методы были разработаны для используется в тех случаях, когда исследователь знает нич его о параметрах интересующей нас переменной (отсюда название непараметрический). Непараметрический методы не полагаются на оценку параметры (такие как среднее значение или стандартное отклонение), описывающие распределение интересующей нас переменной.

Непараметрические модели отличаются от параметрических моделей тем, что структура модели не определено априори, но вместо этого определено из данных. Поэтому эти методы также иногда называемые методами без параметров или безраспространительные методы.

ИСТОРИЯ

Термин «непараметрический» впервые был использован Вулфовиц, 1942 год.

См. также непараметрический тест.

ДОМЕН И ОГРАНИЧЕНИЯ

Непараметрический метод отличается от анализ модели односторонней классификации для сравнения лечения с регрессией и проблемыс подгонкой кривой.

Наиболее часто используемые непараметрические тесты: тест Андерсона-Дарлинга, критерий хи-квадрат, тау Кендалла, критерий Колмогорова-Смирнова, критерий Краскела-Уоллиса, критерий суммы рангов Уил-Коксона, ранг Спирмена. коэффициент корреляции и знак Вилкоксона ранговый тест.

Непараметрические критерии менее эффективны, чем соответствующие параметрические тесты, но более надежным, когда предположения, лежащие воснове параметрический тест не удовлетворяет.

ПРИМЕРЫ

Гистограмма представляет собой простую непараметрическую оценку распределения вероятностей. Обобщение гистограммы—ядерное сглаживание. метод, с помощьюкоторого из заданного набора данных очень гладкая функция плотности вероятности можно построить.

См. также Непараметрические тесты.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Не параметрический тест

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Парзен, Э .: Об оценке функции и режима плотности вероятности. Анна. Мат. Стат. 33, 1065–1076 (1962)

Сэвидж, Р.: Библиография по непараметрической статистике и смежным темам: Введение. Варенье. Стат. доц. 48, 844–849 (1953)

Гиббонс, Дж.Д., Чакраборти, С.: Непараметрический статистический вывод, 4-е изд. КПР (2003)

Не параметрический тест

Непараметрический тест — это вид проверки гипотез, при котором не требуется указывать форму распределения исследуемой совокупности. В любом случае у нас должны быть независимые наблюдения, т. е. выбор особей, образующих включенную в выборку, не должен влиять на выбор других особей.

ИСТОРИЯ

Первый непараметрический критерий появился в работах Арбатнота Дж (1710 г.), который ввел критерий знаков. Но большинство непараметрических тестов было разработано между 1940 и 1955 годами.

Особо отметим статьи Колмогорова Андрея Николаевича (1933), Смирнова Владимира Ивановича (1939), Уилкоксона Ф. (1945, 1947), Манна Х.Б. и Уитни Д.Р. (1947), Мудд А.М. (1950).), и Крускал, В. Х. и Уоллис, В. А. (1952).

Позже к этому списку добавилось много других статей. Savage, IR (1962) опубликовал библиографию

рафия около 3000 статей, написанных до 1962 г. и посвященных непараметрическим тестам.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Быстрое развитие непараметрических тестов можно объяснить следующими моментами: •

Непараметрические методы требуют небольшого количества предположений относительно изучаемой совокупности, таких как предположение о нормальности.

- Непараметрические методы часто легче понять и применять, чем эквивалентные параметрические тесты.
- Непараметрические методы применимы в ситуациях, когда нельзя использовать параметрические критерии, например, когда переменные измеряются только в порядковых шкалах. • Несмотря на то, что, на первый взгляд, непараметрические методы приносят в жертву существенную часть информации, содержащейся в выборках, теоретические исследования показали, что непараметрические тесты лишь немного уступают своим параметрическим аналогам, когда распределение изучаемой совокупности указано, например, нормальное распределение.

Непараметрические тесты, с другой стороны, превосходят параметрические тесты, когда распределение генеральной совокупности далеко от заданного (нормального) распределения.

ПРИМЕРЫ

См. критерий Колмогорова-Смирнова, критерий Крускала-Уоллиса, критерий Манна-Уитни, критерий Уилкокса по знаковому критерию, знаковый критерий.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Тест на пригодность

Проверка гипотезы

Тест Колмогорова – Смирнова

Тест Крускала-Уоллиса

Machine Translated by Google

Тест Манна – Уитни
Подписать тест
Коэффициент ранговой корреляции Спирмена
Тест на независимость
Подписанный тест Уилкоксона
Тест Уилкоксона

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Арбетнотт, Дж: Аргумент в пользу Божественного провидения, основанный на постоянной регулярности, наблюдаемой в рождении обоих полов. Фил. Транс. 27, 186–190 (1710) (3.4).

Колмогоров, А.Н.: Эмпирическое определение пути распределения.
Giornale dell'Instituto Italiano degli Attuari 4, 83–91 (6.1) (1933)

Крускал, В. Х., Уоллис, В. А.: Использование рангов в однокритериальном дисперсионном анализе. Варенье. Стат. доц. 47, 583–621 и опечатки, там же. 48, 907–911 (1952)

Манн, Х.Б., Уитни, Д.Р.: На тесте, является ли одна из двух случайных величин стохастически больше другого. Анна. Мат. Стат. 18, 50–60 (1947)

Настроение, AM: Введение в теорию статистики. Макгроу-Хилл, НьюЙорк (1950) (глава 16).

Savage, IR: Библиография непараметрической статистики. Издательство Гарвардского университета, Кембридж, Массачусетс (1962)

Смирнов Н.В.: Оценка отклонения между эмпирическими функциями распределения в двух независимых выборках. (Русский). Бык. Московский ун-т. 2 (2), 3–16 (6.1, 6.2) (1939)

Уилкоксон, Ф.: Индивидуальные сравнения методами ранжирования. Биометрия 1, 80–83 (1945).

Уилкоксон, Ф.: Некоторые быстрые приближенные статистические процедуры. американский цианамид, Стэмфордские исследовательские лаборатории, Стэмфорд, Коннектикут (1957 г.)

Норма вектора

Норма вектора указывает длину этого вектора, определяемую скалярным произведением вектора на самого себя. Норма вектора получается путем извлечения квадратного корня из этого скалярного произведения.
Вектор нормы 1 называется единичным вектором.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Для заданного вектора x

$$x = \begin{matrix} x1 \\ x2 \\ \vdots \\ xn \end{matrix},$$

мы определяем скалярное произведение x на себя:

$$x \cdot x = \sum_{i=1}^{n} x_i^2.$$

Таким образом, норма вектора x равна:

$$x = \sqrt{x \cdot x} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Поскольку норма вектора вычисляется с помощью

С помощью скалярного произведения свойства нормы получаются из свойств скаляра.

продукт. Таким образом:
- Норма вектора x строго положительна, если вектор не равен нулю, и равна нулю, если он:

  $x > 0$, если $x = 0$ и

  $x = 0$ тогда и только тогда, когда $x = 0$.

Machine Translated by Google

- Норма результата умножения вектора x на скаляр равна произведению нормы x на абсолютное значение скаляра:

$$\|k \cdot x\| = |k| \cdot \|x\|.$$

- Для двух векторов x и y имеем:

$$\|x + y\| \leq \|x\| + \|y\|,$$

то есть норма суммы двух векторов меньше или равна сумме норм этих векторов.

- Для двух векторов x и y абсолютное значение скалярного произведения x и y меньше или равно произведению норм:

$$|x \cdot y| \leq \|x\| \cdot \|y\|.$$

Этот результат называется неравенством Коши–Шварца.

Норма вектора также используется для получения единичный вектор, имеющий то же направление, что и данный вектор. В этом случае достаточно разделить исходный вектор на его норму:

$$u = \frac{x}{\|x\|}.$$

ПРИМЕРЫ

Рассмотрим следующий вектор, определенный в Евклидово пространство трех измерений:

$$x = \begin{matrix} 12 \\ 15 \\ 16 \end{matrix}.$$

Скалярное произведение этого вектора само по себе равно:

$$\begin{aligned} x \cdot x &= 12^2 + 15^2 + 16^2 \\ &= 144 + 225 + 256 = 625 \end{aligned}$$

откуда мы получаем норму x:

$$\|x\| = \sqrt{x \cdot x} = \sqrt{625} = 25.$$

Единичный вектор, имеющий направление x, получается путем деления каждого компонента значение x на 25:

$$y = \begin{matrix} 0{,}48 \\ 0{,}6 \\ 0{,}64 \end{matrix},$$

и мы хотим убедиться, что $\|y\| = 1$. Отсюда:

$$\begin{aligned} \|y\|^2 &= 0{,}48^2 + 0{,}6^2 + 0{,}64^2 \\ &= 0{,}2304 + 0{,}36 + 0{,}4096 = 1. \end{aligned}$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Наименьших квадратов

- оценка L1
- Вектор

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Додж, Ю: Математика базы для экономистов. Спрингер, Берлин Гейдельберг, НьюЙорк (2002)

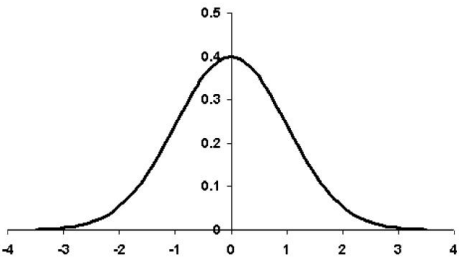
Додж, Ю, Руссон, В.: Многомерное L1-среднее. Метрика 49, 127–134 (1999)

Нормальное распределение

Случайная величина X распределена согласно к нормальному распределению, если оно имеет функциюплотности вида:

$$f(x) = \exp\frac{1}{\sigma\sqrt{2\pi}} \quad \frac{(x-\mu)^2}{2\sigma^2},$$

$$(\sigma > 0).$$



Нормальное распределение, $\mu = 0$, $\sigma = 1$

Machine Translated by Google

Будем говорить, что X следует нормальному распределению среднего μ и дисперсии σ2. Нормальное распределение представляет собой непрерывное распределение вероятностей.

ИСТОРИЯ

Нормальное распределение часто связывают с Лаплас, П.С. и Гаусс, К.Ф., чье имя он носит. Однако его происхождение восходит к работам Бернулли, Дж., который в своем труде Ars Conjectandi (1713) дал первые основные элементы закона больших чисел.

В 1733 г. А. де Муавр впервые получил нормальное распределение как аппроксимацию биномиального распределения. Эта работа была написана на латыни и опубликована в английской версии в 1738 г. А. Де Муавр назвал то, что он нашел, «кривой»; он обнаружил эту кривую, вычисляя вероятности выигрыша для различных азартных игр.

Лаплас, П.С., вслед за де Муавром, А., изучил это распределение и получил более формальный и общий результат приближения де Муавра. В 1774 г. он получил нормальное распределение как аппроксимацию гипергеометрического распределения.

Гаусс К.Ф. изучил это распределение с помощью задач измерения в астрономии. Его работы 1809 и 1816 годов заложили основы техники методы, основанные на нормальном распределении, которые стали стандартными методами в 19 веке.

Заметим, что даже если первое приближение этого распределения обусловлено де Муавра, Галилей уже обнаружил, что ошибки наблюдения распределяются в симметричным образом и имели тенденцию перегруппировываться вокруг их истинного значения.

В литературе можно найти множество наименований нормального распределения.

Кетле, А. (1796–1874) говорил о «кривой возможностей» или «распределении возможностей».

Отметим также, что Гальтон Ф. говорил о «частоте распределения ошибок» или о «распределении отклонений от среднего».

Stigler, S. (1980) представляет более подробное обсуждение различных названий этого изгиб.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение нормального распределения определяется выражением:

$$E[X] = \mu .$$

Дисперсия равна:

$$Var(X) = \sigma^2 .$$

Если среднее значение μ равно 0, а дисперсия σ2 равна 1, то мы получаем центрированное и редуцированное нормальное распределение (или стандартное нормальное распределение), функция плотности которого определяется выражением:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) .$$

Если случайная величина X следует нормальному распределения среднего μ и дисперсии σ2, то случайная величина

$$Z = \frac{X - \mu}{\sigma}$$

после центрированное и уменьшенное нормальное распределение (среднего 0 и дисперсии 1).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Нормальное распределение является наиболее известным непрерывное распределение вероятностей. Она играет центральную роль в теории вероятностей и ее статистических приложениях. Многие измерения, такие как размер или вес людей, диаметр машины, результаты теста IQ и т. д.

Machine Translated by Google

примерно соответствует нормальному распределению. Нормальное распределение часто используется в качестве аппроксимации, либо когда нормальность приписывается распределениюв конструкции модели или когда известное распределение заменяется нормальным распределением с тем же ожидаемым значением и дисперсией. Он используется для аппроксимации распределения хи-квадрат, распределения Стьюдента и дискретных вероятностных распределений, таких как биномиальное распределение и распределение Пуассона.

Нормальное распределение также является фундаментом важный элемент теории выборки, где его роль важна при изучении корреляции, регрессионного анализа, дисперсионного анализа и ковариационного анализа.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Непрерывное распределение вероятностей

Логнормальное распределение

Обычный стол

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бернулли, Дж: ArsConjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurni siorum, Fratrum, Базель (1713 г.)

Гаусс, CF: Theoria Motus Corporum Coelestium. Верке, 7 (1809 г.)

Gauss, CF: Bestimmung der Genauigkeit der Beobachtungen, vol. 4, стр. 109–117 (1816 г.). В: Gauss, CF Werke (опубликовано в 1880 г.). Dieterichsche Universitäts Druckerei, Геттинген

Лаплас, PS de: Mémoire sur la probabi lite des причины par les événements. Мем. акад. Рой. наук. (представлено разными учеными) 6, 621–656 (1774) (или Лаплас, PS de (1891) Œuvres complètes, том 8. Gauthier-Villars, Париж, стр. 27–65).

Moivre, A. de: Approximatio ad summam terminorum binomii (a + b) n, in seriem expansi. Приложение II к Miscellanae Analytica, стр. 1–7 (1733 г.). Фотографически перепечатано вредкой брошюре о Муавре и некоторых его открытиях. Паб опубликовано Арчибальдом, RC Isis 8, 671–683 (1926)

Муавр, А. де: Доктрина шансов: или Метод расчета вероятности событий в игре. Пирсон, Лондон (1718 г.)

Стиглер, С.: Закон понимии Стиглера, Труды Нью-Йоркской академии наук, 2-я серия 39, 147–157 (1980).

Нормальные уравнения

Нормальные уравнения — это уравнения, полученные путем приравнивания к нулю частных производных суммы квадратов ошибок (наименьших квадратов); нормальные уравнения позволяют оценить параметры множественной линейной регрес

ИСТОРИЯ

См. анализ регрессии.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим общую модель множественной линейной регрессии:

$$Yi = \beta 0 + \sum_{j=1}^{p-1} \beta j \cdot Xji + \varepsilon i, \qquad я = 1,\ldots,n,$$

где Yi — зависимая переменная, Xji, j = 1,...,p 1, — независимые переменные, εi — член случайной ненаблюдаемой ошибки,

Machine Translated by Google

βj, j = 0,...,p 1, —оцениваемые параметры, n — число наблюдений.

Чтобы применить метод наименьших квадратов, мы должны минимизировать сумму квадратов ошибок:

$$f(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j \cdot X_{ji} \right)^2 .$$

Полагая частных производных равными нулю относительно оцениваемых параметров, мы получаем p нормальных уравнений:

$$\frac{\partial f(\beta_0, \beta_1, \ldots, \beta_{p-1})}{\partial \beta_0} = 0 ,$$

$$\frac{\partial f(\beta_0, \beta_1, \ldots, \beta_{p-1})}{\partial \beta_1} = 0 ,$$

$$\vdots$$

$$\frac{\partial f(\beta_0, \beta_1, \ldots, \beta_{p-1})}{\partial \beta_{p-1}} = 0 .$$

Мы также можем выразить ту же модель в матричной форме:

$$Y = X\beta + \varepsilon ,$$

где Y —вектор (n × 1) наблюдений, относящихся к зависимой переменной (n наблюдений), X — матрица (n × p) плана, относящаяся к независимым переменным, ε —(n × 1) вектор ошибок, а β —(p × 1) вектор оцениваемых параметров.

Методом наименьших квадратов мы должны минимизировать:

$$\varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

$$= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta$$

$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta .$$

Обнуление производных по β (в матричной форме), соответствующих частному производные, но записанные в векторной форме, находим нормальные уравнения:

$$X'X\hat{\beta} = X'Y .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Мы можем обобщить понятие нормальных уравнений в случае нелинейной регрессии (с n наблюдениями и p + 1 оцениваемым параметром).

Они выражаются:

$$\sum_{i=1}^{n} \left(Y_i - f(X1i, \ldots, Xki; \beta_0, \ldots, \beta_p) \right) \cdot \frac{\partial f(X1i, \ldots, Xki; \beta_0, \ldots, \beta_p)}{\partial \beta_j} = 0 ,$$

где j = 0,...,p. Поскольку f нелинейна в оцениваемых параметрах, нормальные уравнения также нелинейны, и поэтому их решение может быть очень сложным. Более того, система уравнений может иметь более одного решения, что соответствует возможности многих минимумов f . Обычно мы пытаемся разработать итерационную процедуру для решения уравнений или обращаемся к методам, объясняемым нелинейной регрессией.

ПРИМЕРЫ

Если модель содержит, например, 3 параметра β0, β1 и β2, нормальные уравнения будут следующими:

$$\frac{\partial f(\beta_0, \beta_1, \beta_2)}{\partial \beta_0} = 0 ,$$

$$\frac{\partial f(\beta_0, \beta_1, \beta_2)}{\partial \beta_1} = 0 ,$$

$$\frac{\partial f(\beta_0, \beta_1, \beta_2)}{\partial \beta_2} = 0 .$$

Machine Translated by Google

То есть:

$$\beta 0 \cdot n + \beta 1 \cdot \sum_{i=1}^{n} X1i + \beta 2 \cdot \sum_{i=1}^{n} X2i = \sum_{i=1}^{n} Yi,$$

$$\beta 0 \cdot \sum_{i=1}^{n} X1i + \beta 1 \cdot \sum_{i=1}^{n} X1_i^2 + \beta 2 \cdot \sum_{i=1}^{n} X1i \cdot X2i = \sum_{i=1}^{n} X1i \cdot Yi,$$

$$\beta 0 \cdot \sum_{i=1}^{n} X2i + \beta 1 \cdot \sum_{i=1}^{n} X2i \cdot X1i + \beta 2 \cdot \sum_{i=1}^{n} X2_i^2 = \sum_{i=1}^{n} X2i \cdot Yi.$$

См. простую линейную регрессию

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Ошибка

Наименьших квадратов

Множественная линейная регрессия

Нелинейная регрессия

Параметр

Простая линейная регрессия

ССЫЛКИ

Лежандр,

А. М.: Новые методы определения орбит комет.

Курсье, Париж(1805 г.)

График нормальной вероятности

График нормальной вероятности позволяет проверить, распределен ли набор данных в соответствии с нормальным распределением. Когда это случае можно оценить среднее значение и стандартное отклонение от графика.

Накопленные частоты записываются по ординате гауссовой шкалы, градуируются путем помещения значения функции распределения F(t) нормального распределения (центрированного и редуцированного) напротив точки, расположенной на расстоянии t от начала координат. По оси абсцисс представлены наблюдения по арифметической шкале. На практике используются нормальные вероятностные бумаги, продаваемые на рынке.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В системе осей выбор начала и размера единичной длины производится в соответствии с наблюдениями, которые должны быть представлены.

Затем абсцисса арифметически градуируется, как миллиметровая бумага.

По оси ординат значения функции распределения F(t) нормального распределения отложены на высоте t фиктивного начала, расположенного на середине высоты листа бумаги.

Поэтому:

ф(t)	t
0,9772	2
0,8412	1
0,5000	0
0,1588	1
0,0	

будет размещен, и будет сохранена только шкала слева, которая будет градуирована от 0 до 1 (или от 0 до 100, если она написана в %).

Рассмотрим набор из n точечных данных, которые предполагается классифицировать в порядке возрастания. Назовем этот ряд x1, x2,...,xn, и для каждой абсциссы xi (i = 1,...,n) ордината вычисляется как

$$100 \cdot \frac{3i - 8}{1n + 4}$$

и представлена на графике нормальной вероятности.

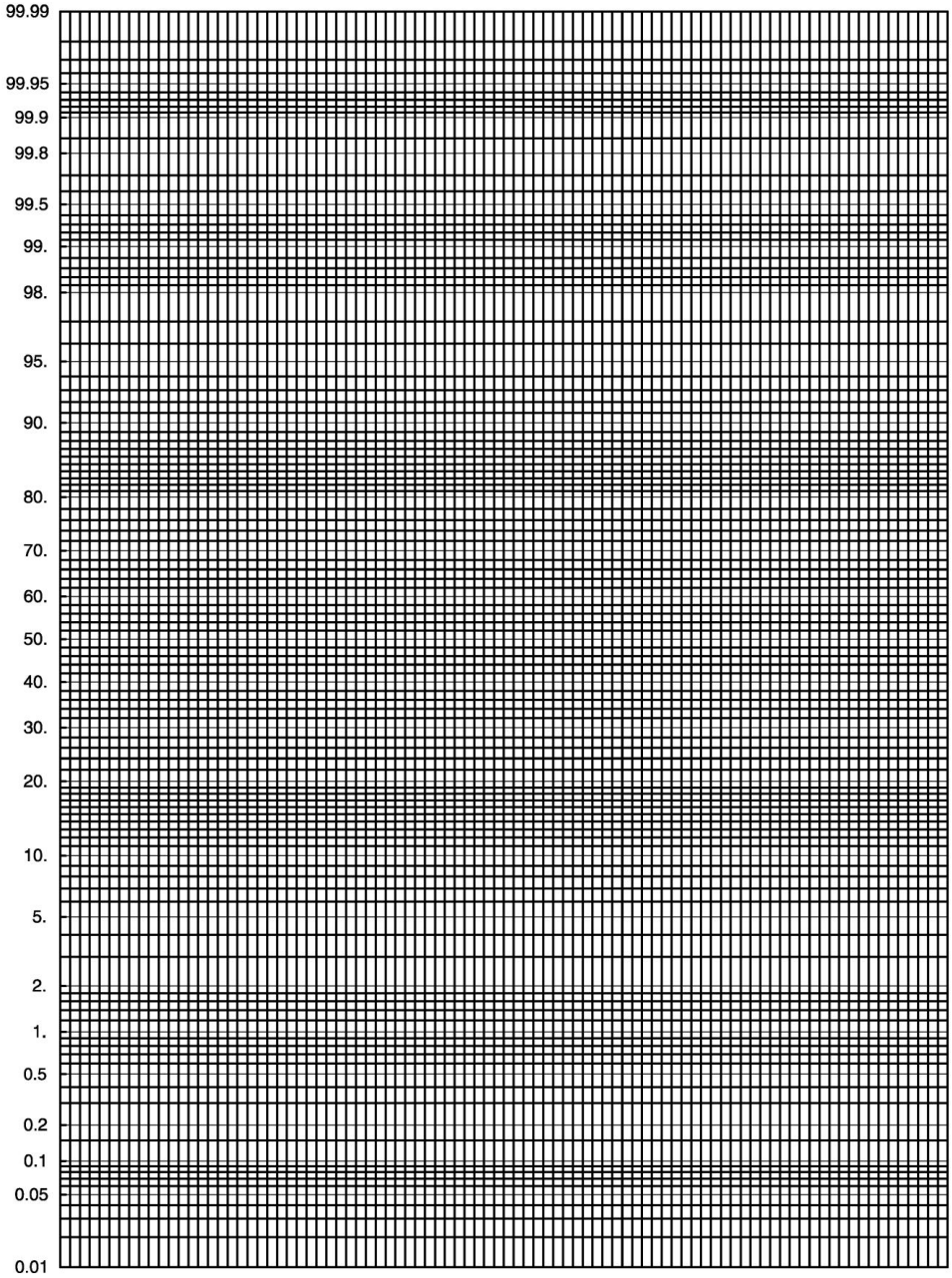
Также возможна транскрипция

$$100 \cdot \frac{1i - 2}{n}$$

как функция xi на графике. Это второе приближение кумулятивной частоты объясняется следующим образом. Если поверхность, расположенная под нормальной кривой (площадью равной

Machine Translated by Google

Papier gausso-arithmetique. Echelle verticale : 0.01 - 99.99 % (made with Mathematica™)



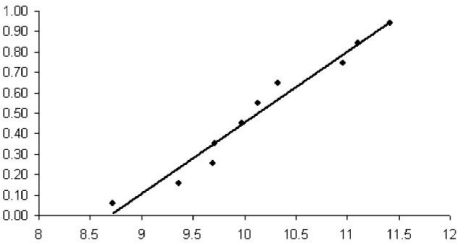
Нормальная вероятностная бумага. Вертикальная шкала: 0,01–99,99 %

Machine Translated by Google

1) делится на n равных частей, можно предположить, что каждое из наших n наблюдений можно найден в одной из этих частей. Поэтому _ наблюдение xi(в порядке возрастания)расположено в середине i-й части, что с точки зрения накопленной частоты соответствует

$$\frac{я - \frac{1}{2}}{н}.$$

Нормальность наблюдаемого распределения Затем можно наблюдать, изучая совмещение точек: если точки совмещены, тогда распределение нормальное.



ОБЛАСТИ И ОГРАНИЧЕНИЯ

Представляя кумулятивные частоты в зависимости от значений статистического переменная на графике нормальной вероятности, это можно проверить, является ли наблюдаемое распределение следует нормальному распределению Действительно, если это так, точки примерно выровнены.

Поэтому графическая корректировка линии к набору баллов позволяет:

- Визуализируйте нормальный характер распределение лучше, чем с его тограммой.
- Сделать графическую оценку среднее значение и стандартное отклонение наблюдаемое распределение.

Линия наименьших квадратов пересекает ординату 0,50 (или 50%) при оценке среднего (что можно прочитать по оси абсцисс).

расчетное стандартное отклонение определяется обратный наклону линии.

Наклон легко определить, просто рассмотрев две точки. Давайте выберем

(μˆ,0,5) и точка на ординате

$$0,9772 = \Phi(2) = \Phi\frac{X0,9772 - \hat{}\mu}{\sigma\hat{}}$$

абсцисса которого X0,9772 читается на линии.

У нас есть:

$$\frac{1}{\sigma\hat{}} "=" \frac{F\text{-}1(0,9772) - F\text{-}1(0,5)}{X - мю}$$

= наклон линии наименьших квадратов,

откуда окончательно получаем:

$$\sigma\hat{} = \frac{X0,9772 - \hat{}\mu}{2}.$$

ПРИМЕРЫ

Рассмотрим десять деталей, изготовленных машиной. X, взятые случайным образом: они имеют следующие диаметры:

9,36 , 9,69 , 11.10 , 8,72 , 10.13 ,
9,98 , 11.42 , 10.33 , 9,71 , 10.96 .

Эти куски, предположительно распределенные согласно нормальному распределению, затем классифицируются в порядке возрастания и для каждого диаметр

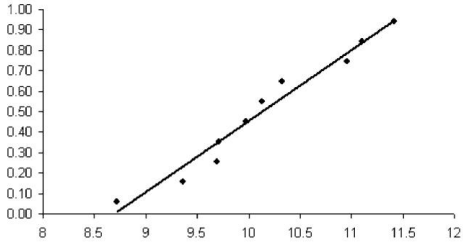
$$фи = \frac{я - 8\frac{3}{}}{10 + \frac{1}{4}},$$

рассчитываются, i - порядковый номер каждое наблюдение:

xi	фи
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	фи
10,13	45/82
10,33	53/82
10,96	61/82
11.10	69/82
11.42	77/82



Графическое представление fi по отношению xi на листе бумаги с нормальными вероятностями дает для нашей случайной выборки: оценочное среднее диаметров: μˆ = 10,2, оценочное стандартное отклонение: σˆ = 0,9.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция распределения
- Частота
- Распределение частоты
- Нормальное распределение

Нормальная таблица

Таблица нормалей, также называемая таблицей Гаусса или таблицей нормального распределения, дает значения функции распределения ранга.

dom, подчиняющаяся стандартному нормальному распределению, т. е. имеющему среднее значение, равное 0, и дисперсию, равную 1.

ИСТОРИЯ

де Лаплас, Пьер Симон (1774 г.) получил нормальное распределение из гипергеометрического распределения, а во второй работе (датированной 1778 г., но опубликованной в 1781 г.) он дал нормальную таблицу. Пирсон, Эгон Шарпи Хартли Х.О. (1948) отредактировали обычную таблицу, основанную на значениях, рассчитанных Шепардом В.Ф. (1903, 1907). Исчерпывающие списки обычных таблиц на рынке были предоставлены Национальным бюро стандартов (до 1952 г.) и Гринвудом, Дж А. и Хартли, Х.О. (до 1958 г.).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

См. центральную предельную теорему.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть случайная величина Z следует нормальному распределение среднего 0 и дисперсии 1.

Таким образом, функция плотности f (t) равна:

$$f(t) = \cdot \exp \frac{1}{2\pi} \quad t \frac{2}{2} \quad .$$

Функция распределения случайных величин возможность Z определяется:

$$F(z) = P(Z \quad z) = \quad \int_{-\infty}^{z} f(t)\,dt.$$

Представленная здесь нормальная таблица дает значения F (z) для различных значений z от 0 до 3,5.

При нормальном распределении, симметричном относительно среднего, имеем:

$$F(z) = 1 \quad F(\quad z),$$

что позволяет определить F (z) при отрицательном значении z.

Иногда мы должны использовать нормальную таблицу в обратном направлении, чтобы найти значение z, соответствующее данной вероятности. Мы обычно обозначаем через $z = z\alpha$ значение случайной величины Z, для которого

$$P(Z \quad z\alpha) \text{ знак равно } 1 - \alpha .$$

Machine Translated by Google

ПРИМЕРЫ

См. Приложение G.

Примерыобычного использования таблицы:

$$P(Z \le 2{,}5) = 0{,}9938,$$

$$P(1 \le Z \le 2) = P(Z \le 2) - P(Z \le 1) = 0{,}9772 - 0{,}8413 = 0{,}1359,$$

$$P(Z \ge -1) = 1 - P(Z \le 1) = 1 - 0{,}8413 = 0{,}1587,$$

$$P(-0{,}5 \le Z \le 1{,}5) = P(Z \le 1{,}5) - P(Z \le -0{,}5) = P(Z \le 1{,}5) - [1 - P(Z \le 0{,}5)] = 0{,}9332 - [1 - 0{,}6915] = 0{,}9332 - 0{,}3085 = 0{,}6247.$$

И наоборот, мы можем использовать нормальную таблицу для определения предела z_α, для которого вероятность того, что случайная величина Z примет значение, меньшее этого предела (z_α), равна фиксированному значению $(1 - \alpha)$, то есть:

$$P(Z \le z_\alpha) = 0{,}95 \Rightarrow z_\alpha = 1{,}645 \quad P(Z \le z_\alpha) = 0{,}975 \Rightarrow z_\alpha = 1{,}960.$$

Это позволяет нам при проверке гипотезы определить критическое значение z_α относительно заданного уровня значимости α.

Пример применения. В случае одностороннего критерия, если уровень значимости α равен 5%, то можно определить критическое значение $z_{0{,}05}$ следующим образом:

$$P(Z \le z_{0{,}05}) = 1 - \alpha = 1 - 0{,}05 = 0{,}95 \Rightarrow z_{0{,}05} = 1{,}645.$$

В случае двустороннего теста нам нужно найти значение $z_{\alpha/2}$. При уровне значимости 5% критическое значение $z_{0{,}025}$ получается следующим образом:

$$P(Z \le z_{0{,}025}) = 1 - \frac{\alpha}{2} = 1 - 0{,}025 = 0{,}975 \Rightarrow z_{0{,}025} = 1{,}960.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Нормальное распределение
- Односторонний тест
- Статистическая таблица
- Двусторонний тест

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Гринвуд, Дж. А., Хартли, Х.О.: Руководство по таблицам математической статистики. Принстон University Press, Принстон, Нью-Джерси (1962)

Лаплас, PS de: Mémoire sur la probabilité des причины par les événements. Мем. акад. Рой. наук. (представлено различными учеными) 6, 621–656 (1774) (или Laplace, PS de (1891) Œuvres complètes, vol 8. Gauthier-Villars, Париж, стр. 27–65).

Лаплас, PS de: Mémoire sur les probabilités.Mem. акад. Рой. наук. Paris, 227–332 (1781) (или Laplace, PS de (1891) Œuvres complètes, vol. 9. Gauthier-Villars, Paris, pp. 385–485.)

Национальное бюро стандартов: Руководство по Таблицы нормального интеграла вероятности. Министерство торговли США. Серия прикладной математики, 21 (1952)

Пирсон, Э.С., Хартли, Х.О.: Таблицы биометрии для статистиков, 1 (2-е изд.). Издательство Кембриджского университета, Кембридж (1948)

Machine Translated by Google

Шеппард, В.Ф.: Новые таблицы интеграла вероятности. Биометрика 2, 174–190 (1903)

Sheppard, WF: Таблица отклонений от нормальной кривой. Биометрика 5, 404–406 (1907).

Нормализация

Нормализация – это преобразование нормально распределенная случайная величина случайная величина, подчиняющаяся стандартному нормальному распределению Позволяет рассчитать и сравнить значения, принадлежащие нормальным кривым среднего и различных отклонения на основе ссылки или mal распределение, то есть стандартное нормальное распределение.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть случайная величина X следует нормальному распределению N μ, σ2 . Его нормализация дает нам случайную величину

$$Z = \frac{X - \mu}{\sigma}$$

которое следует стандартному нормальному распределению H (0, 1).

Каждое значение распределения N μ, σ2 может преобразовать в стандартную переменную Z, каждый Z представляет собой отклонение от среднее, выраженное в единицах стандартного отклонения.

ПРИМЕРЫ

Студенты профессионального училища имеют было два экзамена. Каждый экзамен оценивался шкале от 1 до 60, а оценки считаются реализацией двух случайных величин.

нормального распределения. Давайте сравним оценки, полученные учащимися за эти два экзамена.

Вот средства и стандартная девиа ции каждого экзамена, рассчитанные по всему учебным вмятины:

Испытание 1: μ1 = 35 , σ1 = 4 ,

Испытание 2: μ2 = 45 , σ2 = 1,5 ,

Студент по имени Марк получил следующее Результаты:

Экзамен 1: X1 = 41 ,

Экзамен 2: X2 = 48 .

Вопрос в том, чтобы знать, какой экзамен Марк набрала больше очков по сравнению с другим вмятины. Мы не можем напрямую сравнивать результаты двух экзаменов, потому что результаты принадлежат дистрибутивы с разными средствами и стан резкие отклонения.

Мы просто исследуем разницу каждого отметьте и сравните это со средним значением его распределение, получаем:

Экзамен 1: X1 μ1 = 41 35 = 6 ,

Экзамен 2: X2 μ2 = 48 45 = 3 .

Обратите внимание, что оценка Марка была на 6 баллов выше. чем среднее по экзамену 1 и все его 3 балла выше среднего на экзамене 2. Поспешный вывод наводит нас на мысль, что Марк лучше на экзамене 1, чем на экзамене 2, относительно другие студенты.

Этот вывод учитывает только отличие каждого результата от среднего. Это не учитывает рассредоточение студентов оценки вокруг среднего. Разделим разницу от среднего на стандартное отклонение, чтобы сделать результаты сопоставимыми:

$$\text{Экзамен 1: } Z1 = \frac{X1 \quad \mu1}{\sigma1} \text{ "=" } \frac{6}{_} = 1{,}54 ,$$

$$\text{Экзамен 1: } Z2 = \frac{X2 - м2}{\sigma2} \text{ "=" } \frac{3}{1{,}5} = 2 .$$

Этим расчетом мы нормализовали оценки X1 и X2. Мы можем сделать вывод, что

Machine Translated by Google

нормализованное значение больше для экзамена 2 (Z2 = 2), чем для экзамена 1 (Z1 = 1,5), и что Марк лучше справился с экзаменом 1 по сравнению с другими учащимися.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Нормальное распределение

Нулевая гипотеза

При выполнении проверки гипотез нулевая гипотеза — это гипотеза, подлежащая проверке. Обозначается H0. Противоположная гипотеза называется альтернативной гипотезой. Это альтернативная гипотеза, которая будет принята, если проверка приведет к отклонению нулевой гипотезы.

ИСТОРИЯ

См. гипотезы и проверка гипотез.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

При выполнении проверки гипотез о параметре генеральной совокупности нулевой гипотезой обычно является предположение о предполагаемом значении этого параметра. Тогда нулевая гипотеза будет представлена в виде:

$$H0 : \theta = \theta 0 ,$$

θ — неизвестный параметр популяции, θ0 — предполагаемое значение этого параметра. Этот параметр может быть, например, средним значением генеральной совокупности. Когда проверка гипотез направлена на сравнение двух популяций, нулевая гипотеза обычно предполагает, что параметры равны:

$$H0 : \theta 1 = \theta 2 ,$$

где θ1 — параметр генеральной совокупности, из которой поступила первая выборка, θ2 — совокупность, из которой произошла вторая выборка.

ПРИМЕРЫ

Мы собираемся проверить нулевую гипотезу на трех примерах проверки гипотезы. 1. Проверка гипотезы на проценте выборки. Кандидат, баллотирующийся на должность, хочет знать, получит ли он более 50% голосование.

Нулевая гипотеза для этой задачи может быть сформулирована следующим образом:

$$H0 : \pi = 0,5 ,$$

где π – процент населения, подлежащий оценке.

2. Проверка гипотезы о среднем значении генеральной совокупности Производитель хочет проверить точность новой машины, которая должна производить детали диаметром 8 мм. Мы можем сформулировать следующую нулевую гипотезу:

$$H0 : \mu = 8 ,$$

где μ — среднее значение оцениваемой совокупности.

3. Проверка гипотезы о сравнении средств двух популяций Страховая компания решила оборудовать свои офисы компьютерами. Он хочет покупать компьютеры у двух разных поставщиков, если между надежностью двух марок нет существенной разницы. Было отобрано по два образца ПК от каждой марки. Затем он записывает, как много времени уходило на то, чтобы каждый образец имел проблему на ПК. Согласно нулевой гипотезе среднее время, прошедшее с момента возникновения первой проблемы, одинаково для каждой марки:

$$H0 : \mu 1 \quad \mu 2 = 0 ,$$

Machine Translated by Google

где μ1 и μ2 —соответствующие средние значения двух совокупностей.
Эту гипотезу также можно записать в виде:

$$H0 : \mu1 = \mu2 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Альтернативная гипотеза
- Дисперсионный анализ
- Гипотеза
- Проверка гипотезы

Machine Translated by Google

О

Наблюдение

Наблюдение – это результат научного изучение сбора информации о статистической единице, принадлежащей данной совокупности.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Данные
- Выброс
- Население
- Образец

Шансы и отношение шансов

Шансы определяются как отношение вероятности события к вероятности дополнительное событие, то есть отношение вероятность наступления события (в частном случае эпидемиологии болезни) и вероятность того, что оно не произойдет. Отношение шансов, как и относительный риск, является мерой относительного эффекта причины. Более точнее, это отношение шансов, рассчитанное в популяциях, подверженных данному риску. фактором в разной степени.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Коэффициенты, связанные с событием «болезнь» формально представляют собой следующее соотношение, где p обозначает вероятность заболевания:

$$\text{Коэффициенты} = \frac{p}{1-p}.$$

Пусть существует фактор риска, к которому группа E подвергается воздействию, а группа NE не подвергается воздействию Соответствующее отношение шансов, OR, является отношение шансов группы E и шансов группа CB:

$$OR = \frac{\text{шансы группы E}}{\text{шансы группы NE}}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Возьмем в качестве примера расчет отношение шансов из данных в следующем таблица, сравнивающая две группы людей атаку от такие же потенциально опасные болезнь; одна группа («леченная» группа) получает лечение, в то время как другой не ("контрольная группа):

Группа	Коэффициенты выживших после смерти ИЛИ
Обработано 5% 95%	0,053 0,47
Ссылка 10% 90%	0,111 1,0*

*) По определению ОШ равно 1 для группы ссылка.

Если мы назовем здесь риском риск, связанный с событие «смерть», то ОШ является хорошей аппроксимацией относительного риска при условии, что риски смерти одинаковы или, если они сильно отличаются, то малы. Мы обычно рассматривают риск менее 10%, как в данном случае, как маленький. Отношение шансов не остается приближением относительного

Machine Translated by Google

риск, потому что существует смещение ОШ, которое появляется при небольших рисках. То есть:

$$\text{ИЛИ} = \frac{\text{риск в группе лечения выживаемость в группе лечения}}{\text{риск в контрольной группе выживаемость в группе сравнения}} = \text{относительный риск} \times \frac{\text{риск в группе сравнения}}{\text{выживаемость в группе сравнения}} \times \frac{\text{выживаемость в группе лечения}}{\text{выживаемость в группе лечения}} = \text{погрешность отношения шансов}$$

с:

$$\text{смещение ИЛИ} = \frac{\text{показатель выживаемости в контрольной группе}}{\text{показатель выживаемости в группе лечения}}.$$

Оценки относительных и шансов

коэффициенты используются для измерения силы причинного

Связь между воздействием фактора риска и заболеванием. Два соотношения могут принимать значения от 0 до бесконечности, которые оправдываются одной и той же общей интерпретацией.

Мы можем выделить два явно различных случая: 1.

Значение ОШ больше 1 указывает на риск или шансы заболевания, намного более высокие, когда исследуемая группа подвергается воздействию определенного фактора риска. В таком случае есть положительная ассоциация.

2. Меньшее значение означает меньшее воздействие фактора риска. В этом случае есть негативная ассоциация.

Доверительный интервал при 95% отношения шансов

Техника и обозначения, используемые для доверительного интервала 95% отношения шансов, следующие, те же, что и для относительного риска:

$$\exp(\ln \text{ОШ}) \pm 1{,}961 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

a, b, c и d должны быть большими. Если доверительный интервал включает значение 1,0, то отношение шансов не является статистически значимым.

ПРИМЕРЫ

Рассмотрим исследование взаимосвязи между фактором риска «курение» и раком молочной железы. Вероятность рака молочной железы представляет собой отношение вероятность развития этого рака и вероятность отсутствия этой болезни в течение данного периода. Вероятность развития рака молочной железы для каждой группы определяется уровнем воздействия курения в группе. Следующий пример иллюстрирует этот пример:

Группа	Рак (100000/2 года)	Нет рак (100000/2 года)	Шансы грудь рак (/2 года)	или
Не курящий	114	99886	0,00114	1,0
Пассивный курильщики	252	99748	0,00252	2,2
Активный курильщики	276	99724	0,00276	2,4

В этом примере вероятность рака молочной железы у группы, не подвергавшейся воздействию, составляет 0,00114 (114:99886), у пассивных курильщиков — 0,00252 (252:99748), а у активных курильщиков — 0,00276 (276:99274). Принимая в качестве контрольной группы не подвергавшихся воздействию субъектов, отношение шансов рака молочной железы для активных курильщиков составляет:

$$\frac{\frac{276}{99724}}{\frac{114}{99886}} = 2{,}4 = \frac{276}{114} \cdot \frac{99886}{99724}.$$

Поскольку риск развития рака молочной железы в течение 2 лет очень мал, выживаемость как активных курильщиков, так и некурящих близка к 100%, поэтому коэффициент выживаемости (смещение отношения шансов) практически равен 1,0, что интерпретируется как почти полное отсутствие какой-либо ошибки ИЛИ. В этом примере отношение шансов рака молочной железы идентично

Machine Translated by Google

риск рака молочной железы у пассивных курильщиков и активные курильщики. Так как рак молочной железы редкая болезнь, отношение шансов является хорошей аппроксимацией относительного риска.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Атрибутивный риск

Предотвратимый риск

Причина и следствие в эпидемиологии

Заболеваемость

Уровень заболеваемости

Шансы и отношение шансов

Распространенность

Уровень распространенности

Относительный риск

Риск

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Лилиенфельд, А.М., Лилиенфельд, Д.Э .: Основы эпидемиологии, 2-е изд. Кларендон, Оксфорд (1980)

MacMahon, B., Pugh, TF: Эпидемиология: Принципы и методы. Маленький Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Морабия, А .: L'Epidémiologie Clinique. Издания «Que sais-je?». Presses Universitaires de France, Париж (1996)

Официальная статистика

Прилагательное официальный означает исходящий от правительство, администрация; уполномоченный надлежащим органом. Таким образом, слово «официальный» в сочетании со словом «статистика» указывает на связь с государством. Более того, этимология слова статистика, от латинский статус (государство), означающий быть управляемым государственная статистика данных, собранных государственными учреждениями или частными агентствами работает на правительство. Термин официальный статистика относится к очень общему понятию, которое можно получить разными способами.

ИСТОРИЯ

Перепись, наиболее широко известная официальная статистика, является практикой почти столь же старой, как социальная организация людей. С древности, цари Вавилона и Египта, а также китайский, греческий, иврит, персидский, и римские вожди пытались определить количество солдат под их командованием и товары, которыми они владели. Первые статистические исследования датируются более чем 6000 лет назад. но это были лишь спорадические попытки подсчета людей и товаров. Статистики, в современном понимании этого термина, появляются только в конце 17 века, с созданием современного государства и потребности цвета государственной администрации во Франции и Англии. В XVIII веке в энциклопедии Дидро и Аламбера «арифметическая политика» определяется как «применение арифметических расчетыс политическими предметами и обыкновениями». Французская революция, а затем рождение демократия в США оправдывала новые изменения в официальной статистике. новых государств, знать политический вес различных их сущностях, необходимо знать о их популяции. Вот почему первые регулярные десятилетние переписи населения начались в США в 1790.

Современ Второй мировой войны империи, колонии и племенные общества постепенно уступили место национальным государствам. Организовывать выборов и собраний и содействовать развитие современных корпораций, эти новые государства должны установить статистические системы которые были существенными условиями их развития вариант. Организация Объединенных Наций, в частности,

Machine Translated by Google

Уполномоченная его Статистической комиссией, приложила значительные усилия для популяризации и развития официальной статистики в различных штатах. Под их эгидой в 1950 г. была организована первая всемирная перепись населения. В 1960 г. последовала еще одна.

Сейчас официальная статистика везде существ. Организация экономического сотрудничества и развития (ОЭСР) хочет производить статистические данные, которые могли бы помочь в формулировании точной, всеобъемлющей и сбалансированной оценки состояния основных слоев общества. Эти статистические данные, называемые показателями, с 1960-х годов являются предметом многочисленных исследований в области экономики, здравоохранения и образования. Цель состоит в том, чтобы создать для политиков систему информации, позволяющую им определять последствия своей экономической, социальной и образовательной политики.

ПРИМЕРЫ

Официальная статистика охватывает многие области, наиболее важными из которых являются:

- Население (семьи, домашние хозяйства, рождение, усыновление, признание, смерть, браки, разводы, миграции и т. д.)
- Окружающая среда (земля, климат, вода, фауна, флора). и т. д.)
- Занятость и активная жизнь (предприятия и организации, занятость, условия труда и т. д.)
- Национальные счета (национальный учет, платежный баланс и т. д.)
- Цена (ценыи индексыцен)
- Производство, торговля и потребление (производительность компаний, производство, оборот, внешняя торговля и т. д.)
- сельское хозяйство и лесоводство (использование ресурсов, рабочая сила, машины, леса, рыболовство и т. д.)
- энергия (баланс, производство, потребление и т. д.)
- Строительство и жилье (здания и жилищные конструкции, условия проживания и т. д.)
- Туризм (гостиницы, поездки и т. д.)
- Транспорт и связь (автомобили, сооружения, дорожное движение, дорожно-транспортные происшествия и т. д.)
- Финансовые рынки и банки (банковская система, бухгалтерский учет и т.д.)
- Страхование (страхование от болезней, частное страхование и т.д.)
- Здоровье (система здравоохранения, причины смертности и заболеваемости, медицинские работники, оборудование, расходы на здравоохранение и т.д.)
- Образование и наука (первичное и среднее образование, высшее образование, исследования и разработки и т. д.)
- Спорт, культура и уровень жизни (язык, религия, культурные мероприятия, спорт и т. д.)
- Политика (выборы, участие в политической жизни и т. д.)
- Государственные финансы (федеральные и местные государственные доходы и расходы, налогообложение и т. д.)
- Правовая система (уголовный кодекс, суды, рецидивизм и т. д.)

ОБЛАСТИ И ОГРАНИЧЕНИЯ

История показывает взаимозависимость между статистикой и эволюцией

состояние. Эта взаимозависимость объясняет, по крайней мере частично, трудности, с которыми сталкиваются статистики при определении четких границ в своей дисциплине. На протяжении всей истории и даже сегодня политическая власть определяет пределы изучения официальной статистики, расширяя или ограничивая содержание в зависимости от своих потребностей.

Политический выбор играет роль во многих областях официальной статистики: в определениях

такие понятия, как безработица, бедность, благополучие и др.; в определениях переменных, таких как

профессиональные категории, уровень

Machine Translated by Google

заработной платы и т. д. Эта проблема выбора приводит к двум другим проблемам:

1. Сопоставимость статистики

Потребности и задачи государства диктуют выбор в этой области. Потому что потребности не всегда одинаковы, выбор будет разным в разных местах и на международном уровне. статистика не всегда будет сопоставима. Разница в средствах разных штатов для сбора и использования данных также объясняет разница в результатах между странами.

2. Объективность

Официальная статистика требует объективности. Даже если политическая воля не может быть полностью эволюция статистических теорий и усовершенствование инструментов позволяют гораздо важнее и надежнее Сбор данных. Объем обработанного объем данных значительно увеличился, а информация стала более объективной.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Перепись
- Данные
- Индекс
- Индикатор
- Население

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Duncan, JW, Gross, AC: статистика 21 век: предложения по улучшениюстатистики для лучшего принятия решений. Irwin ProfessionalPublishing, Chicago (1995)

Дюпакье, Ж и Дюпакье, М.: Histoire de демография. Академия Перрен, Париж (1985)

INSEE: Pour une histoire de la statistique. Том. 1: Взносы. INSEE, Economica, Париж(1977)

INSEE: Pour une histoire de la statistique. Том. 2: Материалы. Дж. Аффичард (ред.) INSEE, Economica, Париж(1987)

ISI: Будущее статистики: интернационал Альтернатива. Ворбург (1994)

Стиглер, С.: История статистики, Измерение неопределенности до 1900. Белкнап, Лондон (1986).

Развлечение Статистической комиссии ООН Основные принципыофициальной статистики

ОЖИВ

Огив – это графическое изображение кумулятивное частотное распределение. Это тип частотного графика, который также называется кумулятивный полигон частот. Это служит дать число (или пропорцию) наблюдений, меньших или равных конкретному ценить.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Оживал построен на системе перпендикулярных осей. Наносим на горизонтальную ось границы ранее определенных интервалов классов. На основе каждого из этих предельных значений, определяем ординату высота равна кумулятивной частоте соответствующий этому значению Соединение полинии сегментируя определенные таким образом последовательные точки, мыполучаем линию, называемую оживой. Мы обычно помещают кумулятивную частоту в левая вертикальная ось и процент накопленная частота на правой вертикали ось.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Огивыособенно полезныдля оценки центилей в распределении. Например, мы может знать центральную точку так, что 50%

Machine Translated by Google

наблюдения будут ниже этой точки и 50% выше. Для этого проводим линию от точка 50% на оси процентов до пересечения с кривой.

спроецировать пересечение на горизонтальную ось. Последнее пересечение дает нам желаемое значение. Полигон частот и ogive используются для сравнения двух статистических наборов число которых может быть другим.

ПРИМЕРЫ

Из таблицы, представляющей распределение из дневного оборота 75 магазинов построим оживу.

Распределение дневного оборота 75 магазинов (в г. долларов):

Оборот (сорт интервалы)	Фре качество	Ку му поздно частый су	Процент уровень накопленный частота
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
Общий	75		

В этом примере получаем оценку что у 50% магазинов оборот меньше больше или равно примерно US&289.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение частоты
- Полигон частот
- Графическое представление

Олкин, Ингрэм

Олкин Ингрэм родился в 1924 году в Уотербери, штат Коннектикут, США. Он получил диплом в 1947 в Городском колледже Нью-Йорка и продолжил обучение в Колумбийском университете городе, где получил степень магистра статистика (1949). Он получил докторскую степень по статистической математике в университете Северной Каролины в 1951 году. Затем он преподавал в Университете штата Мичиган и в Университете Миннесоты, прежде чем профессор статистики Стэнфордского университета, должность, которую он занимает в настоящее время.

Ингрэм Олкин работал редактором Анналы статистики, заместитель главного редактора Психометрика, Образовательный журнал Статистика, журнал Американской статистической ассоциации и линейная алгебра and its Applications, среди других математических и статистических журналов. Его исследование фокусируется на многомерном анализе, неравенствах, теории и применении метаанализа в социальных науках и биологии. Он является соавтором многих работ, в том числе Теория неравенства мажорации и ее Приложения (1979 г.), Выбор и упорядочивание популяций (1980 г.), Вероятностные модели. and Applications (1980) и, совсем недавно, «Статистические методы в метаанализе». (1985).

Избранные произведения и статьи Ингрэма Олкин:

1979 (совместно с Маршаллом А.) Неравенства: теория мажорации и ее приложения. Академик, Нью-Йорк.

1984 (совместно с Марсалья, Г.) Создание матриц корреляции. СИАМ Журнал Научные и статистические вычисления 5, стр. 470–475.

Machine Translated by Google

1985 (совместно с Хеджесом, Л.В.) Статистические методы метаанализа. Академик, НьюЙорк.

1994 (с Глесером, Л.Дж и Дерманом, К.)
Вероятностные модели и приложения.
Макмиллан, НьюЙорк. 1995 г.

(совместно с Мохером Д.) Метаанализ
рандомизированных контролируемых испытаний. Афера
cern для стандартов. Варенье. Мед.
доц. 274, 1962–1963. 1996

(совместно с Беггом К., Чо М., Иствудом С.,
Хортоном Р., Мохером Д., Питкиньм Р.,
Рени Д., Шульцом К.Ф, Симелом Д.,
Строупом Д.Ф.) Повышение качества
отчетности по рандомизированные контролируемые испытания.
Заявление CONSORT. Варенье.
Мед. доц. 276(8), 637–639.

Односторонний тест

Односторонний или односторонний тест параметра совокупности —это тип проверки гипотез, в котором значения, для которых мы можем отклонить нулевую гипотезу, обозначаемую H0, полностью расположены в одном хвосте распределения вероятностей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Односторонний тест на выборке - это тип проверки гипотез, при котором гипотезы имеют форму:

$$(1)\ H0 : \theta \geq \theta 0$$
$$H1 : \theta < \theta 0$$

или

$$(2)\ H0 : \theta \leq \theta 0$$
$$H1 : \theta > \theta 0,$$

где H0 —нулевая гипотеза, H1 —альтернативная гипотеза, θ —неизвестный параметр генеральной совокупности, θ0 —предполагаемое значение этого параметра.

В случае одностороннего испытания двух образцов гипотезы следующие:

$$(1)\ H0 : \theta 1 \geq \theta 2$$
$$H1 : \theta 1 < \theta 2$$

или

$$(2)\ H0 : \theta 1 \leq \theta 2$$
$$H1 : \theta 1 > \theta 2,$$

где θ1 и θ2 —неизвестные параметры двух сравниваемых популяций.

Критическая область, которая определяется как набор значений тестовой статистики, для которых нулевая гипотеза H0 отвергается, так как односторонний критерий представляет собой набор значений, меньших критическое значение теста или набор значений, превышающих критическое значение теста.

ПРИМЕРЫ

Односторонний тест среднего значения популяции Производитель духов хочет убедиться, что их флаконы действительно содержат не менее 40 мл духов. Образец из 50 бутылок дает средний объем 39 мл со стандартным отклонением 4 мл.

Гипотезы следующие:

Нулевая гипотеза H0 : мк 40 ,

Альтернативная гипотеза H1 : μ < 40 .

Если размер выборки достаточно велик, выборочное распределение среднего может быть аппроксимировано нормальным распределением. Для уровня значимости α = 1% получаем значение zα в нормальной таблице: zα = 2,33. Критическое значение нулевой гипотезы вычисляется по выражению

$$\mu\bar{x} \quad z\alpha \cdot \sigma\bar{x},$$

где $\mu\bar{x}$ – среднее выборочного распределения средних

$$\mu\bar{x} = \mu$$

О

Machine Translated by Google

$\sigma_{\bar{x}}$ — стандартное отклонение выборочного распределения средних или стандартная ошибка среднего

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Таким образом, мы получаем следующее критическое значение:

$$40 - 2{,}33 \cdot \frac{4}{\sqrt{50}} = 38{,}68$$

Поскольку выборочное среднее $\bar{x} = 39$ больше критического значения для одностороннего левого теста, производитель не может отвергнуть нулевую гипотезу, и поэтому он может считать, что его флаконы с духами действительно содержат по крайней мере 40 мл духов, среднее значение выборки распределено случайным образом.

Односторонний тест на процентное соотношение двух групп населения

Автор законопроекта о дорожном движении полагает, что его законопроект будет более охотно принят городским населением, чем сельским. Были обследованы две выборки по 100 человек из городской и сельской местности. В городской среде (население 1) положительно отнеслись к его законопроекту 82 человека, в то время как в сельской среде (население 2) положительно отреагировали только 69 человек. Чтобы подтвердить или опровергнуть подозрение автора предложения, мы формулируем тест гипотезы, выдвигая следующие гипотезы:

Нулевая гипотеза H0 : $\pi_1 \quad \pi_2$ или $\pi_1 \quad \pi_2 \quad 0$ Alt. гипотеза H1 : $\pi_1 > \pi_2$ или $\pi_1 \quad \pi_2 > 0$

где π_1 и π_2 представляют благополучную долю городского и сельского населения соответственно.

В зависимости от измеренного процента двух образцов ($p_1 = 0{,}82$ и $p_2 = 0{,}69$), мы можем оценить стоимость эталона отклонение выборочного распределения процентной разницы или стандартная ошибка процентной разницы:

$$\hat{\sigma}_{p_1 \quad p_2} = \sqrt{\frac{p_1 \cdot (1 - p_1)}{n_1} \quad \frac{p_2 \cdot (1 - p_2)}{n_2}}$$

$$\text{"="} \quad \sqrt{\frac{0{,}82 \cdot (1 \quad 0{,}82)}{100} \quad \frac{0{,}69 \cdot (1 \quad 0{,}69)}{100}}$$

$$= 0{,}06012 .$$

Если мы разработаем тест с уровнем значимости α, равным 5%, то значение z_α в обычной таблице будет равно 1,645. Критическое значение нулевой гипотезы рассчитывается по формуле:

$$\pi_1 \quad \pi_2 + z_\alpha \cdot \sigma_{p_1 \quad p_2} = 0 + 1{,}645 \cdot 0{,}06012 = 0{,}0989 .$$

Поскольку разница между наблюдаемым пропорции $p_1 \quad p_2 = 0{,}82 \quad 0{,}69 = 0{,}13$ больше критического значения, мы должны отвергнуть нулевую гипотезу в пользу альтернативной гипотезы, тем самым подтвердив предположение высказавшего.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Область приема
- Альтернативная гипотеза
- Проверка гипотезы
- Нулевая гипотеза
- Область отторжения
- Выборочное распределение
- Уровень значимости
- Двусторонний тест

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Миллер, Р.Г., младший: Одновременный статистический вывод, 2-е изд. Springer, Берлин, Хайдельберг, НьюЙорк (1981)

Machine Translated by Google

Однофакторный дисперсионный анализ

Однофакторный дисперсионный анализ фактора с t- уровнями (или t- обработок) —это метод, который помогает определить, существует ли достоверная разница между t обработками (популяциями). Принцип одностороннего дисперсионного анализа заключается в сравнении изменчивость в образцах , взятых из популяции с изменчивостью между этими образцы. Источник изменчивости называется «ошибка» соответствует изменчивости в образцах, а источник изменчивости названный «эффект» соответствует изменчивости между образцами.

ИСТОРИЯ

См. дисперсионный анализ .

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Линейная модель фактора с t уровнями _ (лечение) заключается в следующем:

$$Yij = \mu + \tau i + \varepsilon ij ,$$
$$i = 1, 2,...,t, j = 1, 2,...,ni ,$$

где Yij представляет наблюдение jполучение лечение i, μ —общее среднее общее для всех видов лечения τi представляет собой фактический эффект лечения с упоминанием наблюдений, а εij представляет собой экспериментальную ошибку наблюдения Yij.

Эта модель подвергается основным гипотез , связанных с дисперсионным анализом, если предположить, что ошибки ij равны независимые случайные величины следующие нормальное распределение N(0, σ2).

Чтобы увидеть, существует ли разница между При лечении проводится проверка гипотезы. Следующая нулевая гипотеза:

$$H0 : \tau 1 = \tau 2 = \cdots = \tau t$$

означает, что t обработок идентичны.

Альтернативная гипотеза формулируется как следует:

H1 : не все значения τi (i = 1, 2 , ...,t) идентичны.

Принцип дисперсионного анализа состоит в том, чтобы сравнить изменчивость в каждом образце с изменчивостьюсреди образцов. Для При этом используется критерий Фишера , в котором соотношение формируется числитель которого является оценкой дисперсии между обработками (выборками) и чей знаменатель является оценкой разница в лечении.

Это отношение, обозначенное F, соответствует закону Фишера. распределения с t 1 и N t степенями свободы(N —общее количество наблюдений). Нулевая гипотеза H0 будет отклоняется на уровне значимости α, если F соотношение выше или равно значениюв Таблица Фишера, что означает, если

$$F \quad Ft \quad 1,N \quad t,\alpha .$$

Если нулевая гипотеза H0 не может быть отвергнута, то можно считать, что образцы приходят из того же населения.

Расчет дисперсии между методами лечения

Для дисперсии между обработками сумма квадратов среди обработок рассчитывается как следует:

$$SSTr = \sum_{я=1}^{т} ni(\bar{Y}i. \quad \bar{Y}..)^2 ,$$

где $\bar{Y}i.$ является средним значением i-го лечения. Число степеней свободы, связанных с этой суммой, равно t 1, где t количество образцов (или групп). Разница между методами лечения одинакова к:

$$c^2_{Tp} \text{ "=" } \frac{CCTp}{т - 1} .$

O

Machine Translated by Google

Расчет дисперсии внутри лечения

Для дисперсии внутри обработок сумма квадратов в процедурах, также называемых ошибка, должны быть рассчитаны следующим образом:

$$SSE = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 .$$

Число степеней свободы, связанных с этой суммой, равно N t, где N общее количество наблюдений и t число количество образцов. Тогда дисперсия внутри вариантов лечения равна:

$$s_E^2 = \frac{SSE}{N - t} .$$

Общая сумма квадратов (SST) представляет собой сумму квадраты отклонений каждого наблюдения Yij от общего среднего $\bar{Y}_{..}$:

$$TST = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 .$$

Число степеней свободы, связанных с этой суммой, равно N 1, где N общее количество наблюдений. Мы также можем найти SST следующим образом:

$$SST = SSTr + SSE .$$

Тест Фишера

Теперь мы можем рассчитать коэффициент F и сдержать мой, если образцы можно считать взятыми из одной и той же популяции.

Отношение F равно оценке _ дисперсия между методами лечения, разделенная на оценка дисперсии в рамках лечения, значение

$$F = \frac{s_{Tr}^2}{s_E^2} .$$

Если F больше или равно значению в Таблица Фишера для t 1 и N t степеней свободы, то можно сделать вывод, что разница между образцами обусловлена экспериментальные методы лечения и не только случайность.

Таблица дисперсионного анализа

принято обобщать информацию Анализ дисперсии в таблице называется Таблица дисперсионного анализа, представленная следующим образом:

Источник вариация	Степени из свобода	Сумма квадраты	Среднее значение квадраты	F
Среди обращаться менты	т-1	SSTr	s_{Tr}^2	$\frac{s_{Tr}^2}{s_E^2}$
В пределах обращаться менты	Nt	SSE	s_E^2	
Общий	N-1	SST		

Попарное сравнение средних

Когда тест Фишера отклоняет нулевую гипотезу, что означает наличие существенной разницы между средними значениями выборок, мы можно задаться вопросом, где эта разница. Сейчас были разработаны универсальные тесты, чтобы ответить этот вопрос. Существует тест наименьшего значимого различия (LSD) (по-французски test de la différence minimale), что делает сравнения средних, взятых попарно. Для чтобы использовать этот тест, отношение F должно указывать на значительную разницу между означает.

ПРИМЕРЫ

Во время приготовления круассаны впитывают жир в переменные количества. Мы хотим посмотреть, будет ли поглощаемое количество зависит от типа

Machine Translated by Google

жир. Готовятся четыре разных вида жира и готовятся шесть круассанов каждого типа. смазки.
Данные приведены в следующей таблице (числа обозначают количество смазки, поглощаемой на круассан):

Смазка			
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

Нулевая гипотеза состоит в том, что между методами лечения нет различий, а это означает, что количество жира, впитавшегося во время приготовления не зависит от типа смазки:

$$H0 : \tau1 = \tau2 = \tau3 = \tau4 .$$

Для проверки этой гипотезы используется тест Фишера. используется, для которого необходимо определить отношение F. Поэтому дисперсия между методами лечения и дисперсия внутри методов лечения должны быть рассчитаны

Различия между методами лечения

Общее среднее $\bar{Y}..$ и среднее значение четыре множества $\bar{Y}i. = \frac{1}{ni}\sum_{j=1}^{ni} Y1i$ (i = 1, 2, 3, 4) имеют быть рассчитаны. Поскольку все ni равны 6 и N равно 24, получаем:

$$\bar{Y}.. = \frac{64 + 72 + \ldots + 70 + 68}{24} = \frac{1770}{24} = 73,75 .$$

Мы также рассчитываем:

$$\bar{Y}1. = \frac{1}{6}\sum_{j=1}^{6} Y1j = \frac{432}{6} = 72 ,$$

$$\bar{Y}2. = \frac{1}{6}\sum_{j=1}^{6} Y2j = \frac{510}{6} = 85 ,$$

$$\bar{Y}3. = \frac{1}{6}\sum_{j=1}^{6} Y3j = \frac{456}{6} = 76 ,$$

$$\bar{Y}4. = \frac{1}{6}\sum_{j=1}^{6} Y4j = \frac{372}{6} = 62$$

Затем мы можем вычислить сумму квадратов среди процедур:

$$SSTr = \sum_{i=1}^{4} ni(\bar{Y}i. - \bar{Y}..)^2$$
$$= 6 (72 - 73,75)^2 + 6(85 - 73,75)^2 + 6(76 - 73,75)^2 + 6(62 - 73,75)^2$$
$$= 1636,5 .$$

Число степеней свободы, связанных с этой суммой, равно 4 1 = 3.
Разница между методами лечения одинакова к:

$$c^2_{Tp} = \frac{CCTp}{т - 1} = \frac{1636,5}{3} = 545,5 .$$

Разница в лечении

Сумма квадратов в обработках или ошибках равно:

$$CCЭ = \sum_{i=1}^{4}\sum_{j=1}^{6} (Yij - \bar{Y}i.)^2$$
$$= (64 - 72)^2 + (72 - 72)^2 + \ldots + (70 - 62)^2 + (68 - 62)^2$$
$$= 2018 .$$

Число степеней свободы, связанных с этой суммой, равно 24 4 = 20.

Machine Translated by Google

Дисперсия внутри лечения остается равной к:

$$\sigma^2_E = \frac{SSE}{N - t} = \frac{2018}{20} = 100{,}9 .$$

Тест Фишера

Все элементы, необходимые для расчета F, теперь известны:

$$F = \frac{\sigma^2_{Tr}}{\sigma^2_E} = \frac{545{,}5}{100{,}9} = 5{,}4063 .$$

Нам нужно найти значение таблицы Фишера с 3 и 20 степенями свободы при уровне значимости $\alpha = 0{,}05$:

$$F_{3.20,0.05} = 3.10 .$$

Мы замечаем, что

$$F \quad F_{3,20,0,05},$$

$$5{,}4063 \quad 3{,}10 \quad ,$$

что означает, что H0 должен быть отвергнут. Это означает, что существует значительная разница между методами лечения. Поэтому количество впитавшейся смазки зависит от типа используемой смазки.

Таблица дисперсионного анализа

Эта информация обобщена вслед.

Нижняя таблица дисперсионного анализа:

Источник из вариация	Степени из свобода	Сумма квадраты	Среднее значение квадраты	Ф
Среди обращаться менты	3	1636,5	545,5	5.4063
В пределах обращаться менты	20	2018	100,9	
Всего	23	3654,5		

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Контраст
- Таблица Фишера
- Тест Фишера
- Тест на наименьшее существенное различие
- Двусторонний дисперсионный анализ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кокс, Д.Р., Рид, Н.: Теория дизайна экспериментов. Чепмен и Холл, Лондон (2000)

Монтгомери, округ Колумбия: Дизайн и анализ эксперименты, 4-е изд. Уайли, Чичестер (1997)

Шеффе, Х.: Дисперсионный анализ. С'п'Н (1959)

Исследование операций

Исследование операций – это область прикладного математика, использующая научные методы для обеспечить необходимую основу для принятия решений. Обычно применяется в комплексе проблемы организации людей и оборудования, чтобы найти наилучшее решение для достижения какая-то цель. Входит в исследование операций методы — моделирование, линейное программирование, математическое программирование (нелинейное), теория игр и т. д.

ИСТОРИЯ

Первоначальная цель исследования операций должен был найти военное применение в Соединенном Королевстве в 1930-х годах. Эти военные Затем приложения были в значительной степени расширены, и после Второй мировой войны использование исследования операций было распространено на военные общественные приложения и даже к частный сектор. Влияние на операции

Machine Translated by Google

исследования быстрого развития компьютеров были очень важны.

ПРИМЕРЫ

Оптимизация, линейное программирование и моделирование являются примерами операций исследовать.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Линейное программирование
- Оптимизация
- Моделирование

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Bealle, EML: Введение в оптимизацию. Уайли, НьюЙорк (1988)

Коллатц, Л., Веттерлинг, В.: Оптимизация проблемы. Springer, Берлин Гейдельберг НьюЙорк (1975)

Оптимальный дизайн

Оптимальный план – это экспериментальный дизайн, удовлетворяющий определенному критерию оптимальности (например, минимизация ошибки, связанные с оценками, которые мы должен сделать). Поскольку лечение применялось к различные экспериментальные единицы фиксируются экспериментатор, мы можем попытаться улучшить качество результатов. Мы также можем увеличить точность оценок, уменьшить доверительные интервалы, или увеличить мощность проверки гипотез, выбирая оптимальный дизайн внутри категории, к которой она принадлежит.

ИСТОРИЯ

Теория оптимальных планов была разработана Кифером Джеком (1924–1981). Один из основные положения его работы основаны на статистическая эффективность дизайна. В своей статье 1959 г. он представляет основные критерии получения оптимального дизайна. Кифера сборник статей в области планирования экспериментов был опубликован Springer с сотрудничество Института математики математическая статистика.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Поскольку мы не можем свести к минимуму все ошибки на Второе время оптимальный дизайн будет определяться выбором критерие в оптимальности. Эти различные критерии выявляются в Статья Кифера 1959 года.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- План экспериментов

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Браун, Л.Д, Олкин, И., Сакс, Дж, Винн, Х.П.: Джек Карл Кифер, Сборник статей III: Дизайн экспериментов. Спрингер, Берлин Гейдельберг, НьюЙорк (1985)

Федоров В.В. Теория оптимальных экспериментов. Академик, НьюЙорк (1972)

Кифер, Дж: Оптимальные экспериментальные планы. Дж Рой. Стат. соц. сер. Б 21, 272–319 (1959)

Оптимизация

Большое количество статистических задач может решать с помощью методов оптимизации. Мы используем методы оптимизации в статистических процедурах, таких как метод наименьших квадратов, максимальное правдоподобие, оценка L1 и т. д. во многих других областях статистики, таких как регрессионный анализ, проверка гипотез, планирование эксперимента и т. д., оптимизация играет скрытую, но важную роль.

Machine Translated by Google

Методы оптимизации можно разделить на следующим образом:

1. Классические методы оптимизации (дифференциальный расчет, множители Лагранжа).
2. Методы математического программирования (линейное программирование, нелинейное программирование, динамическое программирование)

ИСТОРИЯ

Проблемы оптимизации были сформулированы Евклидом, но только с развитием дифференциального исчисления и вариационное исчисление в 17-м и 18-м веках были математические инструменты, способные решения таких проблем имеются. Эти методы были впервые применены в решение некоторых задач оптимизации в геометрии и физике. Задачи такого типа были сформулированы и изучали такие известные математики, как как Эйлер, Леонард Бернулли, Якоб, Якоби, Карл Густав Якоби и Лагранж, Жозеф Луи, которым мы обязаны множителем Лагранжа.

Открытие важных приложений метод оптимизации в военных задачах в годы Второй мировой войны и достижения в области технологий (компьютер) позволили для решения все более сложных задач и оптимизации в самых разных поля.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Множитель Лагранжа
- Линейное программирование
- Исследование операций

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Артханари, Т.С., Додж, Ю: Математика Программирование в статистике. Уайли, Нью Йорк (1981)

Выброс

Выброс — это наблюдение, которое хорошо отделено от остальных данных.

Выбросы определяются относительно предполагаемого базового распределения или теоретической модель. Если они изменятся, наблюдение может не быть более отчужденным.

ИСТОРИЯ

Проблема выбросов была решена многими учеными с 1750 г. Бошкович, Роже Жозеф де Лаплас, Пьер Симон, и Лежандр, Адриан Мари. По словам Стиглера, Стивена М. (1973), Лежандр А. М. предложил в 1805 г. отвергнуть выбросы, а в 1852 г. Пирс и Бенджамин установили первый критерий для определения выбросов. Появились критические замечания, особенно со стороны Эйри, Джорджа Бидделла (1856 г.), против использования такого критерия.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В рамках статистического исследования мы можем сформулируйте вопрос о выбросах двумя способами:

- Первый состоит из адаптации методов, которые устойчивы к наличию выбросов в образец.
- Второй пытается устранить выбросы. После идентификации удалите их из выборку перед проведением статистического анализа.

Удаление выбросов из набора данных является мошенническим спорная проблема, особенно в небольшом наборе данных. Проблема выбросов привела к исследованию новые статистические методы, называемые надежным методом раз. Оценки, не чувствительные к выбросам, говорят крепкий.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Анализ остатков
Матрица шляпы
оценка L1
Точка рычага
Наблюдение
Надежная оценка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Эйри, Великобритания: письмо профессора Эйри, королевского астронома, редактору. Астроном. Журнал 4, 137–138 (1856 г.)

Барнетт, В., Льюис, Т.: Выбросы в статистических данных. Уайли, НьюЙорк (1984)

Пирс, Б.: Критерий отклонения сомнительных наблюдений. Астрон. Ж. 2, 161–163 (1852).

Стиглер, С.: Саймон Ньюкомб, Перси Дэниелл и история надежных оценок 1885–1920 гг. Варенье. Стат. доц. 68, 872–879 (1973)

Machine Translated by Google

Индекс Пааше

Индекс Пааше является составным индексом. число цен, полученное методом взвешенной суммы. Этот номер индекса соответствует соответствует отношению суммы цен фактического периода n и суммы цен отчетного периода 0, эти суммы взвешенные по соответствующим количествам фактический период. Индекс Пааше отличается от индекс Ласпейреса только по выбору метод весов. Действительно, в Ласпейресе индекс, веса даны с использованием количеств отчетного периода Q0 , а не текущего периода Qn.

ИСТОРИЯ

В середине 19 века немецкий статистик Пааше Герман разработал формулу для порядковый номер, который носит его имя. Пааше Х. (1874 г.) работал над ценами, записанными в Гамбурге.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Индекс Пааше рассчитывается следующим образом:

$$B/0 = \frac{Пн \cdot Qн}{P0 \cdot Qn},$$

где Pn и Qn – соответственно цены и объемы продаж в текущем периоде и P0 и Q0 —цены и объемы продаж в отчетном периоде . Суммы относятся к считаются товарами и выражаются в базовой 100.

$$B/0 = \frac{Пн \cdot Qн}{P0 \cdot Qn} \cdot 100 .$$

Модель Пааше также может быть применена для расчета количественного индекса (также называемого объемным). индекс). В данном случае речь идет о ценах. постоянные и переменные величины:

$$B/0 = \frac{Qн \cdot Пн}{Q0 \cdot Pн} \cdot 100 .$$

ПРИМЕРЫ

Рассмотрим следующую таблицу, указывающую соответствующие цены трех продуктов питания в отчетном году 0 и в текущем году n, а также как объемы продаж в текущем году:

Продукт	Количество продано в 1988 году (Qn (тысячи)	Цена (евро) 1970 (P0)	1988 (Пн)
Молоко	85,5	0,20	1,20
Хлеб	50,5	0,15	1.10
Масло	40,5	0,50	2.00

Из следующей таблицы имеем:

$$PnQn = 239,15 и$$

$$P0Qn = 44,925 .$$

Machine Translated by Google

Продукт	PnQn	п0Qn
Молоко	102,60	17.100
Хлеб	55,55	7,575
Масло	81.00	20.250
Общий	239,15	44.925

Затем мы можем найти индекс Пааше:

$$P0 = \frac{Pn \cdot Qn}{P0 \cdot Qn} \cdot 100$$

$$"=" \frac{239,15}{44.925} \cdot 100 = 532,3 .$$

Другими словами, согласно Пааше Индекс цен на эти товары за рассматриваемый период вырос на 432,3% (532,3 100).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Составной порядковый номер
- Индекс Фишера
- Индекс
- Индекс Ласпейреса
- Простой порядковый номер

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Пааше, Х.: Über die Preisentwicklung der letzten Jahre nach den Hamburger Börsen-notirungen. Ярб. Натл. Стат. 23, 168–178 (1874 г.)

Пара случайных величин

Пара переменных, значения которых определяются случайным экспериментом, называется пара случайных величин. Есть два виды пар:

- Пара случайных величин дискретна, если набор значений, принимаемых каждой из случайных величин, является конечным или бесконечным числом способный набор.
- Пара случайных величин непрерывна если множество значений, принимаемых каждой из случайных величин, является бесконечным несчетным набор.

В случае пары случайных величин совокупность всех возможных результатов эксперимента составляет выборочное пространство, т.е. находится в двумерном пространстве.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пара случайных величин, как правило, обозначаются заглавными буквами, такими как X и Y. Это функция двух переменных с истинными значениями в 2.

Если A является подмножеством выборочного пространства, то:

$$P(A) = \text{Вероятность того, что } (X, Y) \quad A$$

$$= P[(X, Y) \quad A] .$$

Случай пары случайных величин может можно обобщить на случай n случайных величин: если для описания необходимо n чисел осуществление случайного эксперимента, такая реализация представлена n случайными переменные X1, X2,...,Xn. Пример пространства находится в n-мерном пространстве. Если A является подмножеством выборочного пространства, то мы можем скажи это:

$$P(A) = P[(X1, X2,...,Xn) \quad A] .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Определенные случайные ситуации заставляют сидите не одно, а несколько числовых сущностей галстуки одновременно. Рассмотрим, например, система из двух элементов в ряду и случайных периодов жизни X и Y; изучение продолжительность жизни всего множества должна учитывать события, касающиеся X и Y одновременно. время. Или рассмотрим цилиндр, изготовленный машина диаметром X и высотой Y с

Machine Translated by Google

эти значения случайны. Вводимые допуски обычно учитывают обе сущности одновременно; структура их вероятностных зависимость должна быть известна.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Ковариация
- Совместная функция плотности
- Функция предельной плотности
- Случайная переменная

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Хогг, Р.В., Крейг, А.Т.: Введение к математической статистике, 2-е изд. Макмиллан, Нью-Йорк (1965)

Парный Т-тест Стьюдента

Парный критерий Стьюдента является проверкой гипотез, который используется для сравнения средств двух популяции, когда каждый элемент популяции связан с элементом из другой один.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть xij —наблюдение j для пары i(j = 1, 2 и i = 1, 2,...,n). Для каждой пары наблюдений вычисляем разность

$$\text{ди знак равно } xi2 - xi1 .$$

Ищем расчетный норматив ошибка среднего значения di , обозначаемая d:

$$sd^{-} = \frac{1}{\sqrt{n}} \cdot Сд ,$$

где Sd —стандартное отклонение di :

$$Сд = \sqrt{\frac{\sum_{я=1}^{н} (ди - д^{-})^2}{п - 1}} .$$

Результирующий статистический тест определяется следующим образом:

$$T = \frac{A^{-}}{сд^{-}} .$$

Гипотезы

Двусторонний тест «Парные тесты студента» .

Гипотезы:

H0 : δ = 0 (нет разницы среди процедур)

H1 : δ = 0 (есть отличие среди процедур) ,

где δ – разница между средними двух популяций (δ = μ1 μ2).

Правила принятия решений

Мы отвергаем нулевую гипотезу о значимости уровень α, если

$$|T| > tn \quad 1, \alpha_{\frac{}{2}} ,$$

где tn 1, α $\frac{}{2}$ это ценность студента _ таблица с n 1 степенями свободы.

ПРИМЕРЫ

Предположим, что к десять пар наблюдений. Полученные данные и соответствующие разности обозначаются по di представлены в следующей таблице:

Пара я	Уход		ди "=" xi 2- xi 1
	1	2	
1	110	118	8
2	99	104	5
3	91	85	6
4	107	108	1
5	82	81	1
6	96	93	3
7	100	102	2
8	87	101	14
9	75	84	9
10	108	111	3

П

Machine Translated by Google

Среднее значение равно:

$$\bar{д} = \frac{1}{10}\sum_{я=1}^{10} ди = \frac{32}{10} = 3{,}2 .$$

Стандартное отклонение рассчитывается следующим образом:

$$Сд = \sqrt{\frac{\sum_{я=1}^{10} (ди - \bar{д})^2}{10 - 1}} = \sqrt{\frac{323{,}6}{9}} = 6{,}00 ,$$

и стандартная ошибка:

$$\bar{sd} = \frac{1}{\sqrt{n}} \cdot Сд$$

$$= \frac{1}{\sqrt{10}} \cdot 6{,}00 = 1{,}90 .$$

Тогда статистический тест равен:

$$T = \frac{\bar{д}}{\bar{сд}} = \frac{3.2}{1{,}90} = 1{,}69 .$$

Если мы выберем уровень значимости α = 0,05, значение t9,0,025 равно 2,26 . Таким образом, нуль гипотеза H0 : δ = 0 не может быть отвергнута потому что |T| < t9,0,025.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Ученический стол
- Студенческий тест

Панель

Панель – это разновидность переписи , которая повторяется периодически. Он основан на постоянном (или полупостоянная) выборка лиц, домохозяйства и т. д., которым регулярно задают вопросы об их поведении или мнении. панель предлагает преимущество следования индивидуальное поведение допрашиваемых или единиц измерения любых изменений в их поведение во времени. Сбор информации ed в панели часто богаче, чем в простой переписи, сохраняя при этом ограниченные затраты. Мы адаптировать скорость обновления образца в зависимости от цель расследования. Когда цель эксперт должен следить за эволюцией людей с течением времени (например, следуя карьерному пути), образец остается прежним; мы называем это лонгитюдинальным исследованием. Когда цель выявляется служить способом оценки характеристик населения в разные периоды, образец будет частично обновляться время от времени время; мы называем это сменной панелью. Таким образом данные панели представляют собой источник очень важную информацию, поскольку они содержат одновременное измерение личности и времени хорошо

ИСТОРИЯ

История панельных исследований восходит к 1759 г., когда французский граф Жюмонбей Лард филибер Гено начал записывать свои рост сыновей с шестимесячным интервалом от рождения до 18 лет. Его записи имеют мало общего с современными панельными исследованиями, если не считать от систематического характера его неоднократных наблюдения. Текущая концепция панели исследование было начато примерно в 1920-х годах и 1930-е годы, когда несколько исследований человека начался рост и развитие . Поскольку В середине 20-го века панельные исследования получили широкое распространение в социальных науках. Несколько факторов способствовали росту панельных исследования в области социальных наук. Большой рост в 1960-х и 1970-х годах пришли в ответ к увеличению федерального финансирования панельных исследований в Соединенных Штатах. Например, озабоченность политикой в отношении работы и экономического поведения привела к финансированию текущих исследований, таких как Национальные лонгитюдные исследован (NLS) в 1966 г. и панельное исследование доходов.

Machine Translated by Google

Динамика 1968 г. Последняя, например, начиналась с 4802 семей с перечнем переменных более 5000; выборка бедных домохозяйств была избыточной. В опросе NLS приняли участие 5020 пожилых мужчин, 5225 молодых мужчин, 5083 зрелых женщины, 5159 молодых женщин и 12686 молодых людей.

Технологические достижения облегчили сбор данных в национальном масштабе и увеличили способность исследователей управлять, анализировать и обмениваться большими и сложными наборами данных. Кроме того, достижения в области статистических методов анализа лонгитюдных данных побудили исследователей не только собирать панельные данные, но и задавать новые вопросы о существующих наборах данных. Рост панельных исследований очевиден в других промышленно развитых странах,

слишком.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

На практике панели в строгом смысле этого слова встречаются редко. При этом следует учитывать следующие проблемы панели: сложность набора определенных категорий экспертов, усталость участников, которая является одной из основных причин отсутствия ответов, и эволюция населения , вызывающая деформации выборки от генеральной совокупности, которую она должна представлять. На практике мы часто пытаемся устранить эту проблему, предпочитая сменную панель строгой.

Существуют также специфические ошибки измерения в панелях:
(1) Эффект телескопа: ошибка датировки, часто допускаемая новыми экспертами при указании даты события, о котором их спрашивают.

(2) эффект группы или предвзятость обусловленности: изменение в поведении членов группы, которое может проявиться со временем в результате повторных опросов. Участников дискуссии спрашивают об их поведении, и, наконец, они меняют его.

ПРИМЕРЫ

Запрос о здоровье и медицинском обслуживании: опрос, проведенный INSEE примерно в 10 000 домохозяйств. За каждым домом следят

3 месяца, в течение которых каждые 3 недели члены домохозяйства опрашиваются исследователем об их медицинском обслуживании.

Панель потребителей Sofres: Панель Metascope компании Sofres основана на выборке из 20 000 домохозяйств. Каждый месяц участники дискуссии получают по почте анкету для самостоятельного заполнения по различным покупкам. Эти анкеты в основном используются компаниями, которые хотят знать, что их клиенты

хотеть.

Панель аудитории: Панель Mediamat основана на выборке из 2300 домохозяйств, оснащенных аудиометрами с кнопками.
Это позволяет исследователям измерять аудиторию телестанций.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Предвзятость

Население

Образец

Опрос

ССЫЛКИ

Дункан, Г.Дж.: Панельные исследования домохозяйств: перспективы и проблемы. Методология социальных наук, Тренто, Италия (1992 г.)

Каспршик Д., Дункан Г. и соавт. (ред.): Панельные обзоры. Нью-Йорк, Чичестер, Брисбен, Торонто, Сингапур. Уайли, Нью-Йорк (1989)

Роуз, Д.: Панельные исследования домохозяйств: обзор. Инновации 8 (1), 7–24 (1995)

Диггл, П.Дж., П. Хигерти, К.-Ю. Лян и С.Л. Зегер.: Анализ продольных данных, 2-е изд. Издательство Оксфордского университета, Оксфорд (2002)

П

Machine Translated by Google

Параметр

Параметр характеризует количественный аспект совокупности .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Параметры популяции часто неизвестны. Однако мы можем оценить параметр по статистике , рассчитанной по выборке с использованием метода оценки.

ПРИМЕРЫ

Параметр обычно обозначается греческой буквой:

μ: среднее значение популяции
σ: стандартное отклонение популяции π: процентное отношение к популяции

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Оценка
- Оценщик
  - Наименьших квадратов
- Максимальная вероятность
- Момент
- Статистика

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Леманн, Э.Л.: Теория точечной оценки, 2-е изд. Уайли, Нью-Йорк (1983)

Параметрический тест

Параметрический тест — это форма проверки гипотез , при которой делаются предположения о лежащем в основе распределении наблюдаемых данных.

ИСТОРИЯ

Одним из первых параметрических тестов был критерий хи-квадрат, введенный Пирсоном К. в 1900 году.

ПРИМЕРЫ

Тест Стьюдента является примером параметрического теста. Он направлен на сравнение средних значений двух нормально распределенных совокупностей. Среди наиболее известных параметрических тестов мы упомянем t-критерий Стьюдента и критерий Фишера. тест.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Биномиальный тест
- Тест Фишера
- Проверка гипотезы
  - Студенческий тест

Частичная автокорреляция

Частичная автокорреляция при задержке k — это автокорреляция между Yt и Yt k , которая не учитывается задержками от 1 до k — 1.

См. автокорреляцию.

Частичная корреляция

Частная корреляция между двумя переменными определяется как корреляция двух переменных способен, контролируя треть или более других переменных. Мера частичной корреляции между переменными X и Y имеет третий переменная Z как мера прямого отношения связь между X и Y , которая не принимает учитывать последствия линейных отношений этих двух переменных с Z.

расчет частной корреляции между X и Y при заданном Z служат для определения того, в какой мере линейная связь между X и Y может быть связано с корреляцией двух переменных с З.

ИСТОРИЯ

Интерпретация анализа коэффициентов корреляции между двумя переменными стала

Machine Translated by Google

важный вопрос в статистике во время все более широкого использования корреляционных методов в начале 1900-х годов. Пирсона, Карл знал, что большая корреляция между двумя переменными может быть связана с их корреляцией с третьей переменной. Это явление не было признано до 1926 г., когда Юль Джордж Удни доказал его на примере, получив коэффициенты корреляции между временными рядами.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Из выборки из n наблюдений (x1, y1, z1), (x2, y2, z2),..., (xn, yn, zn) из неизвестного распределения трех случайных величин X, Y и Z коэффициент частной корреляции, который мы обозначим через rxy.z, определяется как коэффициент корреляции, рассчитанный

междуxi andyi c

$$xi = \beta\, 0x + \beta\, 1xzi\,,$$

$$yi = \beta\, 0y + \beta\, 1yzi\,,$$

где β 0x и β 1x — оценки методом наименьших квадратов , полученные путем регрессии xi на zi, а β 0y и β 1y — оценки методом наименьших квадратов, полученные путем регрессии yi на zi. Таким образом, по определению имеем:

$$rxy.z = \frac{(xi - \bar{x})(yi - \bar{y})}{\sqrt{(x - x)^{-2}}\ \sqrt{(й - y)^{-2}}}\,.$$

Мы можем доказать следующий результат:

$$rxy.z = \frac{rxy - rxz \cdot ryz}{\sqrt{1 - r2_{xз}}\ \cdot\ \sqrt{1 - r2_{yz}}}\,,$$

где rxy, rxz и ryz — коэффициенты корреляции соответственно между xi и yi, xi и zi, yi и zi . Обратите внимание, что когда xi и yi не коррелируют с zi, то есть когда rxz = ryz = 0, мы имеем:

$$rxy.z = rxy\,,$$

т. е. коэффициент корреляции равен коэффициенту частной корреляции, или, другими словами, величина корреляции между xi и yi вовсе не обусловлена наличием zi. С другой стороны, если важны корреляции с zi , то rxy.z = 0, что указывает на то, что наблюдаемая корреляция между xi и yi обусловлена только корреляцией между этими переменными с

зи. Таким образом, мы говорим, что между xi и yi нет прямой связи (а скорее косвенная связь через zi).

В более общем случае мы можем вычислить частную корреляцию между xi и yi с zi1 и zi2 относительно двух переменных Z1 и Z2 следующим образом:

$$rxy.z1z2 = \frac{rxy{\cdot}z1 - rxz2.z1 \cdot ryz2.z1}{\sqrt{1 - r2_{xz2.z1}}\ \cdot\ \sqrt{1 - r2_{yz2.z1}}}\,.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Мы должны настаивать на том, что корреляция измеряет только линейную связь между двумя переменными, не принимая во внимание функциональные модели или прогнозирующую способность данной модели.

ПРИМЕРЫ

Мы наблюдали, например, сильную положительную корреляцию между количеством зарегистрированных инфарктов в течение определенного периода и количеством проданного за тот же период мороженого. Мы не должны делать вывод, что чрезмерное потребление мороженого провоцирует инфаркты или, наоборот, что инфаркт провоцирует особое желание мор

пле, большое количество инфарктов не связано с количеством проданного мороженого

а от жары. Таким образом, мы имеем сильную корреляцию между количеством инфарктов

и температура. Большое количество

П

Machine Translated by Google

продажи мороженого также объясняется жарой, что подразумевает сильную корреляцию между количеством проданного мороженого и темпом температура. Наблюдается сильная корреляция между количеством инфарктов и количество проданного мороженого - это conse следствие двух сильных корреляций. Третья переменная скрыта за видимой связью двух предыдущих переменных.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент корреляции

ССЫЛКИ

Yule, GU (1926) Почему мы иногда получаем бессмысленные корреляции между временными рядами? Исследование выборки и природы временных рядов. Дж. Рой. Стат. соц. (2) 89, 1– 64

Частичное наименьшее абсолютное

Регрессия отклонения

Частичная регрессия с наименьшим абсолютным отклонением (частичная LAD) — это метод регрессии, который линейно связывает вектор отклика с набором предикторов с использованием производных компонентов. Это метод моделирования регрессии L1 , хорошо подходящий для ситуаций, когда количество параметров в модели линейной регрессии превышает количество наблюдений. Он в основном используется для целей прогнозирования, а не для вывода по параметрам. Метод частичной регрессии LAD расширяет частичную регрессию наименьших квадратов до нормы L1 , связанной с регрессией LAD, вместо нормы L2 , которая связана с частичными методами наименьших квадратов. Частичная LAD-регрессия следует структуре одномерного алгоритма частичной регрессии по методу наименьших квадратов и извлекает компоненты (обозначаемые t) из направлений (обозначаемых w), которые зависят от переменная отклика. Направления определяются ковариационной оценкой Гнанадесикана–Кеттернинга (GK), которая заменяет обычную дисперсию, основанную на норме L2 , на MAD, среднее абсолютное отклонение, основанное на L1. Поэтому мы используем обозначение $wmad_k$ для частичных направлений LAD.

ИСТОРИЯ

Метод частичной регрессии LAD был представлен Dodge, Y. et al. (2004). Дальнейшее развитие он получил в Dodge, Y. et al. (2004) и протестировано с помощью бутстрапина Kondylis, A. и Уиттакер, Дж. (2005).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Для i = 1,...,n и j = 1,...,p, обозначающих единицы наблюдения и предикторы, алгоритм частичной LAD-регрессии приведен ниже: 1. Центрировать или стандартизировать как X , так и y.

2. Для k = 1,...,kmax:

вычислить $wmad_k$ по формуле

$$wmad_{j,k} = \frac{1}{4}\,mad2(x_{j,k} + y) - mad2(x_{j,k} - y),$$

и увеличьте wmad до 1.k

Извлечь компонент

$$t_k = \sum_{j=1}^{p} wmad_{j,k}\, x_{j,k},$$

где

$$wmad_k = wmad_{1,k},\ldots,wmad_{p,k}.$$

Ортогонализуем каждый $x_{j,k}$ относительно t_k: $x_{j,k} = x_{j,k} - E(x_j|t_k)$.

3. Задайте результирующую последовательность вектор y_k подбора = $T_k q_k$, где T_k = (t1,..., tk) — матрица оценок, а qk = (q1,...,qk) — вектор коэффициентов регрессии LAD.

Machine Translated by Google

4. Восстановите подразумеваемую пластину частичного восстановления
коэффициенты по β LAD. "=" W безумный k qk, где матрица W ее столбцами являются векторы wmad k сумасшедший k бассейны в выраженный в условия исходного xj.

Одномерная частичная регрессия методом наименьших квадратов и метод частичной LAD имеют общее следующие свойства:

1. y = q1t1 + ... + qktk + 2. X = p1t1 + ... + pktk + f ,

3. cor(ti, tj) = 0 при i = j,

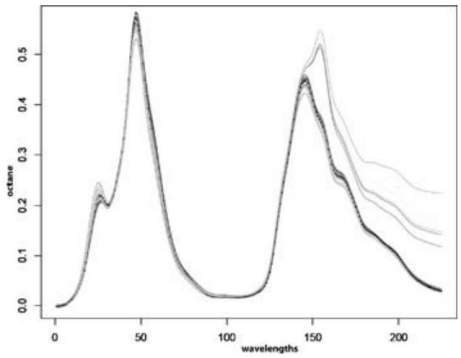
где cor(,) обозначает корреляцию Пирсона коэффициент, а f соответствуют остаточному условия, и pk X -загрузка.

Частичный метод LAD строит регрессионную модель, связывающую предикторы с ответ согласно:

$$y_k^{\text{плед}} = \sum_{j=1}^{n} \beta_{дж}^{\text{плед}} x_{дж}.$$

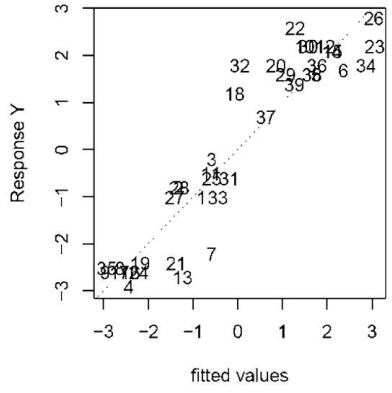
ПРИМЕР

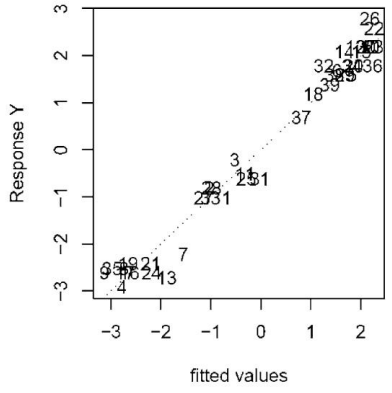
Мы приводим здесь пример из ближнего инфракрасного диапазона. эксперименты по спектроскопии. Мы используем набор данных об октановом числе, состоящий из 39 образцов бензина, для которых октановое число были измерены на 225 длинах волн (в нанометров). Регрессоры представляют собой сильно мультиколлинеарные спектры при различном количестве длин волн, а их количество превышает размер выборки н.



На рисунке мы даем линии для 39 пробы бензина по всей длине волны.

Мы используем метод частичной регрессии LAD. построить линейную модель на основе двух производных компоненты. Получившийся в результате сюжет





Модель частичной регрессии по методу наименьших квадратов потребуется дополнительный компонент в

П

Machine Translated by Google

окончательная модель для получения хороших прогностических результатов в виде частичной LAD-регрессии. модель. Это связано с группой из шести выбросов в наборе октановых данных. это обсерва 25, 26, 36, 37, 38 и 39, которые содержали спирт. Они видны для длин волн более 140 нм на линейном графике. выше.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Регрессия наименьшего абсолютного отклонения
- Частичная регрессия методом наименьших квадратов

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Додж Ю., Кондилис А., Уиттакер Дж.:
Расширение регрессии PLS1 до PLAD и использование нормы L1 в софтмоде
Эллинг КОМПСТАТ 2004, стр. 935–942, Physica/Verlag-Springer (2004)

Додж Ю., Уиттакер Дж., Кондилис А.:
Прогресс на PLAD и PQR. В: Материалы 55-й сессии ISI 2005 г., стр. 495–503 (2005 г.)

Кондилис, А., Уиттакер, Дж.: Использование ремня Boot для регрессии PLAD. Пожалуйста и Связанные методы. В: Материалы Международный симпозиум PLS'05 2005, стр. 395–402. Издание Aluja et al. (2005)

Частичный метод наименьших квадратов

Регрессия

Частичная регрессия по методу наименьших квадратов (PLSR) это статистический метод, который связывает два данных матрицы X и Y, обычно называемые блоками, через латентно-линейная структура. Применяется либо на вектор одного ответа y (одномерный PLSR) или матрица отклика Y (многомерный PLSR). PLSR обычно используется при записи переменные сильно коррелированы. это очень подходящий метод в тех случаях, когда количество переменныепревышают количествополезных возможных наблюдений. Это связано с тем, что вместо этого использует ортогональные производные компоненты исходных переменных. Использование нескольких ортогональные компоненты вместо многочисленных коррелированные переменные гарантируют уменьшение задачи регрессии на малом подпространстве, которая часто стабилизирует изменчивость оцениваемых коэффициентов и обеспечивает более точные прогнозы.

ИСТОРИЯ

Первоначально PLSR был разработан в рамках алгоритм NIPALS, хотя он был реализованы в различных алгоритмах, включая ортогональные оценки и ортогональную загрузку PLSR, алгоритм SIMPLS для PLSR, алгоритм Хелланда PLSR и Алгоритм ядра PLSR.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Методы PLSR решают следующую задачу максимизации:

$$\max_{w,q} \{\mathrm{cov}(X_{k-1}w, Y_{k-1}q)\}$$

при условии

$$\max_{k} w = 1, \quad q^T \quad k \, q_k = 1$$

и

$$t_k \quad t_j, \; u_k \quad u_j \text{ для каждого } k = j,$$

где cov(·, ·) обозначает ковариацию и w_k, q_k и t_k, u_k загружают векторы и компоненты (или векторы оценок) для X и Y соответственно. Использование индекса k в X и Y показывает, что PLSR дефлирует данные при каждом извлеченная размерность k и использует остатки как новые данные для следующего измерения.

Machine Translated by Google

Одномерный PLSR гораздо проще интерпретировать.
прет. На самом деле это обобщение множественной
линейной регрессии , которое уменьшает
коэффициенты регрессии в направлениях низких ковариаций.
связь между переменными и ответами. Учитывая
следующее отношение:

$$\text{cov}(Xw, y) \quad \text{cor}(Xw, y)\text{var}(X)^{1/2},$$

где cor и var обозначают корреляцию и
дисперсии, соответственно легко проверить
что критерий максимизации в PLSR
является компромиссом между методом наименьших квадратов
регрессия и регрессия по принципалу
компоненты. Последние методы максимизируют
cor(Xw, y) и var(X) соответственно.
Одномерный PLSR включает в себя следующее
шаги:

1. Для (X, y)общецентрированных:
2. Для к = 1,...,р:
   - Сохраните cov(xj,k 1, yk 1) в векторе wk.
   - Пул wk в матрице W(p×k).
   - Извлеките компонент tk как tk = Xj,k 1wk.
   - Ортогонализуйте Xk 1 и yk 1 с помощью
     уважение к тк.
3. В качестве новых данных используйте остатки по методу наименьших квадратов.
   для k k + 1.

После извлечения компонентов
регрессия по методу наименьших квадратов используется для регрессии
их на вектор отклика y. Номер
компонентов, которые должны быть в конечном итоге
сохранено определяется с помощью бутстрапа и
методы перекрестной проверки.

ПРИМЕР

Метод PLSR представляет собой обобщение
множественной линейной регрессии. Это оправдано
используя задачу регрессии, в которой не происходит
корреляции между предикторами, т. е.
ортогональный дизайн. В таблице ниже приведены
результаты ортогонального плана с
две объясняющие переменные и один ответ.

Таблица: Ортогональный дизайн

y	x1	x2
18	2	4
12	1	3
10	0	6
16	1	5
11	2	3
9	0	3
11	1	6
8	1	1
7	1	0
12	1	1

Множественный линейный регрессионный анализ для
данные в Таблице 1 (y как ответ и два
объясняющие переменные x1 и x2) приводят к
следующие оценки методом наименьших квадратов:

$$y = 11{,}40 - 1{,}8571x1 + 0{,}1408x2 .$$

С помощью метода PLSR (данные были изначально
центрированы и масштабированы) и регрессии
ответ на производные компоненты мы
привели к точно таким же оценкам только с
используется одна компонента, т. е. byy = t1 ·q1.
Мы наконец получаем

$$y = T1q1 = Xw0q1 = X\beta \text{ PLS} .$$

Таким образом, для стандартизированных данных
коэффициенты регрессии

$$q\beta \text{ PLS} = (\quad 0{,}68,\ 0{,}164),$$

что при обратном преобразовании к исходному
масштабу равно (11, 40, 1,8571, 0,1408).
Последние представляют собой множественную линейную регрессию.
оценки.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Широко используются методы PLSR.
в ближних инфракрасных экспериментах по спектроскопии.
Они уже давно используются химометрами для
многомерной калибровки и

П

Machine Translated by Google

для измерения концентрации химических веществ. ПЛСР
методы в равной степени использовались во многих
статистических приложениях с большим количеством переменных, помимо
наблюдения, например, в области экологической метрики,
в экспериментах с микрочипами в биостатистике,
и в функциональном анализе данных.
Методы PLSR сосредоточены в основном на прогнозировании
точности и в основном используются для построения
хороших прогностических моделей.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Начальная загрузка

Метод наименьших квадратов

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Тененхаус, М.: La régression PLS. Теори
и практика. Технип, Париж (1998)

Франк И., Фридман Дж.: Статистический взгляд
некоторых инструментов хемометрической регрессии.
Технометрика 35, 109–135 (1993).

Мартенс, Х., и Нэс, Т.: Многомерная калибровка. Уайли,
Нью-Йорк (1989)

Хелланд: О структуре регрессии PLS, Comm. в Стат-
Симул. и комп.,
17, 581–607 (1988)

Пирсон, Эгон Шейп

Пирсон, Эгон Шейп (1895–1980), сын
статистик Пирсон Карл изучал математику в Тринити-
колледже в Кембридже. В
В 1921 г. он поступил на факультет статистики
Университетского колледжа в Лондоне. Он
познакомился с Нейманом, Ежи, с которым начал
сотрудничать во время визита последнего в Лондон в
1924–1925 гг., а также сотрудничал с Госсетом,
Уильямом Сили. Когда
его отец бросил Университетский колледж в 1933 году, он
перешел в отдел прикладной статистики и в 1936 г.
стал заведующим журналом
нал Биометрика после смерти его
отец, основатель журнала.
Помимо работ, опубликованных им в соавторстве с
Нейманом, Дж. и приведших к
развитие теории проверки гипотез, Пирсон, Э.С.

проблемы контроля качества и операций
исследовать.

Избранные статьи Пирсона, Э.С.:

1928 г. (совместно с Нейманом, Дж.) Об использовании и
интерпретация определенных тестовых критериев
для целей статистического вывода.
Биометрика 20A, 175–240, стр. 263–
295.

1931 г. Проверка значимости коэффициента
корреляции. Варенье. Стат. доц.
26, 128–134.

1932 г. Установлены процентные пределы
распределения образцов ренджина из
нормальной популяции (n 100). Биометрия
24, 404–417.

1933 г. Нейман, Дж. И Пирсон, Э.С.
проверка статистических гипотез в
отношение к вероятности априорно. проц.
Камб. Филос. соц. 29, 492–510.

1935 г. Применение статистического метамфетамина.
ods к промышленной стандартизации и
Контроль качества. Брит. Стан. 600.
Британский институт стандартов, Лондон.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Проверка гипотезы

Пирсон, Карл

Пирсон, Карл (1857–1936) родился в Лондоне.
известен своим многочисленным вкладом в

Machine Translated by Google

статистика. После учебы в Королевском колледже в Кембридже он был назначен в 1885 году заведующим кафедрой прикладной математики Университетского колледжа в Лондоне. Всю свою карьеру он провел в этом университете, где в 1911 году стал заведующим кафедрой евгеники. В 1901 году с помощью Гальтона Фрэнсиса он основал журнал «Биометрика» и был его главным редактором до своей смерти в 1936 год.

В 1906 году он пригласил на 1 год в свою лабораторию Госсета Уильяма Сили, с которым он решал проблемы, связанные с образцами малых размеров. Он вышел на пенсию в 1933 году, и его отделение в университете было разделено на два: отделение евгеники было поручено Фишеру и Рональду Эйлмеру, а отделение статистики — его собственному сыну Пирсону, Эгону Шейпу.

Избранные статьи Пирсона, Карла:

1894 1948. Вклад в математику.

математическая теория эволюции. I. С. 1–40. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, Кембридж. Впервые опубликовано как: О анализе асимметричных частотных кривых, Philos. Транс. Рой. соц. Лонд. сер. А 185, 71–110.

1895 1948. Вклад в математику.

математическая теория эволюции. II. Косые вариации в однородном материале. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, Кембридж, стр. 41–112. Впервые опубликовано в Philos. Транс. Рой. соц. Лонд. сер. А 186, 343–414.

1896 1948. Математические вклады в

теория эволюции. III. Регрессия, наследственность и панмиксия. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, Кэм мост, стр. 113–178. Впервые опубликовано в Philos. Транс. Рой. соц. Лонд. сер. А 187, 253–318 в Philosophical Magazine, 5-я серия, 50, стр. 157–175.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение хи-квадрат
- Коэффициент корреляции

Процент

Процент — понятие меры, позволяющей описать долю лиц или статистических единиц, обладающих определенной характеристикой в совокупности, оцениваемой на основе 100.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Процент обычно обозначается буквой p , когда он измеряется на выборке , и π, когда он касается генеральной совокупности.

Пусть имеется выборка размера, если k индивидуумов обладают определенной характеристикой; процент особей определяется статистикой

$$p = \frac{k}{n} \cdot 100\% .$$

Расчет того же процента на популяции размером N дает нам параметр π:

$$\pi = \frac{k}{N} \cdot 100\% .$$

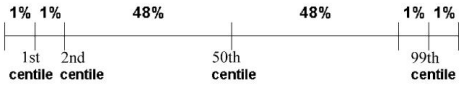
Процентиль

Процентили - это измерения местоположения вычисляется в наборе данных. Мы называем процентилями значения, которые делят распределение на 100 равные части (каждая часть содержит одинаковое количество наблюдений). X -й процентиль

Machine Translated by Google

значение Cx такое, что x% наблюдений меньше и (100 x)% наблюдений больше, чем Cx. Например, 20-й процентиль — это значение, которое отделяет 20% значений, которые ниже его, от 80% значений.

которые выше его. Тогда у нас будет 99 процентилей для данного распределения:



Обратите внимание, например:

Процентиль 10 = первый дециль

Процентиль 25 = первый квартиль
Процентиль 50 = медиана

Это понятие является частью семейства квантилей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Процесс вычисления аналогичен медиане , квартилям и децилям. Когда все необработанные наблюдения даны, процентили рассчитываются следующим образом:

1. Организуйте n наблюдений в виде частотного распределения.
2. Процентили соответствуют наблюдениям, для которых относительная кумуляция частота превышает соответственно 1%, 2%, 3%, ..., 98%, 99%.

   Некоторые авторы предлагают следующую формулу, которая позволяет с точностью определить значение различных процентилей:

   Расчет j-го процентиля: Рассмотрим i целую часть $j\cdot(n+1)$ k дробную часть $j\cdot(n+1)$ и $\frac{}{100}$.

   Рассмотрим xi и xi+1 значения данных наблюдений , соответственно классифицированных в i-й и (i + 1)-й позициях (когда n наблюдений отсортированы по возрастанию) в порядке).

   j -й процентиль равен:

$$Cj = xi + k \cdot (xi+1 \quad xi) .$$

Когда у нас есть наблюдения , которые сгруппированные в классы, процентили определяются следующим образом:

1. Определяем класс, в котором желаемый процентиль:
   - 1-й процентиль: первый класс, для которого относительная кумулятивная частота превышает 1%.
   - 2-й процентиль: первый класс, для которого относительная кумулятивная частота превышает 2%.
   - 99-й процентиль: первый класс, для которого относительная кумулятивная частота превышает 99%.
2. Рассчитайте значение процентилей в зависимости от гипотезы , согласно которой наблюдения равномерно распределены в каждом классе:

$$\text{процентиль} = L1 + \frac{(n \cdot q) \quad finf}{fcentile} \cdot c,$$

где L1 — нижняя граница в классе процентиля, n — общее количество наблюдений, q — центиль, q — 99 , q — $\frac{}{1\ 100}$ за 1-го человека 100 , сумма $\frac{2}{100}$ для 2 - го процентиля, ..., для частот — 99-го процентиля, finf ниже , чем класс процентиля, fpercentile — частота класса процентиля, c – размерность интервала класса процентилей.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Расчет процентилей имеет истинное значение только для количественной переменной , которая может принимать свои значения в заданном интервале. На практике процентили можно вычислить только при большом количестве наблюдений , так как этот расчет состоит в делении

Machine Translated by Google

набор наблюдений на 100 частей. Это понятие применяется редко, потому что в большинстве описательных анализов определение децилей или даже квартилей достаточно для интерпретация результатов.

ПРИМЕРЫ

Рассмотрим пример расчета процентили частотного распределения непрерывной переменной , где наблюдатель ции группируются в классы. Следующая таблица частот представляет прибыль (в тысячах евро) 2000 г. бактерии:

Выгода (тысячи евро)	Фре quen cies	Куму поздно частоты	Родственник кумулятивный Эд Фре качество
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
Общий	2000 г.		

Класс, содержащий первый процентиль, представляет собой класс 100–150 [тот, для которого относительная кумулятивная частота превышает 0,01 (1%)]. Учитывая, что наблюдения равномерно распределены в каждом классе, получаем для первого процентиля следующее значение:

1-й процентиль

$$= 100 + \frac{2000 \cdot \frac{1}{100} - 0}{160} \cdot 50$$

$$= 106{,}25 .$$

Класс, содержащий 10-й процентиль, класс 150–200. Значение 10-го процентиля равно

10-й процентиль

$$= 150 + \frac{2000 \cdot \frac{10}{100} - 160}{200} \cdot 50$$

$$= 160 .$$

Мы можем вычислить все процентили в так же. Тогда мы можем заключить, например, что 1% из 2000 пекарен имеют прибыль от 100 000 до 106 250 евро, или что 10% имеют прибыль от 100000 до 160000 евро и т.д.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- дециль
- Мера местоположения
- медиана
- квантиль
- Квартиль

П

Перестановка

Термин перестановка является предметом комбинаторного анализа. Он относится к договоренности или упорядоченный из n объектов. Теоретически мы следует различать случай, когда n объекты полностью идентифицированы, а те, где они идентифицированы лишь частично.

ИСТОРИЯ

См. комбинаторный анализ.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

1. Количество перестановок различных объекты

Machine Translated by Google

Количество возможных перестановок
n объектов равно n факториалу:

Перестановки

$$= n! = n \cdot (n - 1) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 .$$

2. Количество перестановок частично различных объектов.
Количество возможных перестановок
n объектов , среди которых n1, n2,..., nr
среди них не различаются равные:

$$\text{количество перестановки} = \frac{n!}{n1! \cdot n2! \cdot \ldots \cdot №!},$$

где n1 + n2 + ... + nr = n.

ПРИМЕРЫ

Если у нас есть три объекта A, B и C, то возможные перестановки следующие:

азбука
АСВ
БАК
ВСА
ТАКСИ
ЦБ

Таким образом, у нас есть шесть возможных перестановок.
Не перечисляя их, можно найти
количество перестановок, используя следующие формула:

$$\text{Количество перестановок} = n! ,$$

где n — количество объектов. Это дает мы в нашем примере:

$$\text{Количество перестановок} = 3! = 3 \cdot 2 \cdot 1 = 6 .$$

Представьте, что мы составляем сигналы, выравнивая некоторые флаги. Если у нас есть пять флажков разного цвета, количество сигналов, которые мы может составить будет:

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120 .$$

С другой стороны, если среди флагов есть два красных, два белых и один черный флаг, мы можем составить:

$$\frac{5!}{2! \cdot 2! \cdot 1!} = 30$$

разные сигналы.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Договоренность
- Комбинация
- Комбинационный анализ

Пиктограмма

Пиктограмма — это символ, представляющий концепцию, объект, место или событие посредством иллюстрации. Как и гистограмма, пиктограмма используется либо для сравнения категорий качественной переменной, либо для сравнения наборов данных, происходящих из разных лет или из разных мест. Пиктограммы в основном используются в журналах как рекламные графики для сравнения и
не представляют большого статистического интереса, поскольку дают только грубые приближения.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

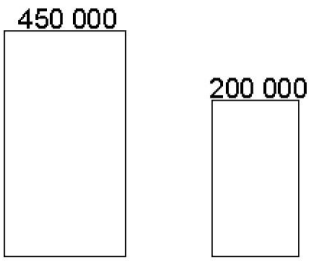
Фигура, выбранная для иллюстрации различных величин на пиктограмме, должна, по возможности, делиться на четыре части, чтобы они
может представлять половинки и четверти. Масштаб _ выбирается в зависимости от имеющегося пространства. Поэтому цифра не обязательно равна
к единству.

ПРИМЕРЫ

Рассмотрим производство пшеницы в двух странах. A и B. Производство страны A равно
равен 450 000 центнеров, а страны Bis – 200 000 центнеров (1 центнер равен
до 100 кг).

Machine Translated by Google

Получаем следующую пиктограмму:



Поверхность каждой фигуры пропорциональна представленное количество.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Гистограмма
- Графическое представление
- Количественный график

Круговая диаграмма

Круговая диаграмма — это тип количественного графа. Он состоит из круга, разделенного на сектора, каждый сектор с углом, пропорциональным изображаемая величина.

ИСТОРИЯ

См. графическое представление.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Чтобы построить круговую диаграмму, вычисляется сумма представленных частот , а затем рассчитываются относительные частоты, представляющие разные сектора. Чтобы нарисовать эти секторов по кругу, относительные частоты преобразуются в градусы следующим образом трансформация:

$$\alpha = f \cdot 360 \quad ,$$

где α представляет собой угол в центре и f относительная частота.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Круговые диаграммы используются для визуального представления данных, формирующих различные части целое население. Они часто встречаются в средства массовой информации (телевидение, газеты, журналы, д.), где они служат для объяснения в простой и синтетическим образом концепция или ситуация, которая трудно понять только по цифрам.

ПРИМЕРЫ

Мы представим распределение семейного положения в Австралии на 30 июня 1981 г., используя круговая диаграмма. Данные следующие :

Семейное положение в Австралии на 30 июня 1981 г. (в тысячи)

Семейное положение	Частота	Относительное частота
Холостяк	6587,3	0,452
Женатый	6836,8	0,469
В разводе	403,5	0,028
вдова	748,7	0,051
Общий	14576,3	1.000

Источник: Австралийское статистическое бюро, Австралийский карманный ежегодник, 1984 г., с. 11

Сначала относительные частоты преобразуются в градусы:

а) $\alpha = 0{,}452 \cdot 360$	$= 162, 72$
б) $\alpha = 0{,}469 \cdot 360$	$= 168, 84$
в) $\alpha = 0{,}028 \cdot 360$	$= 10, 08$
г) $\alpha = 0{,}051 \cdot 360$	$= 18, 36$

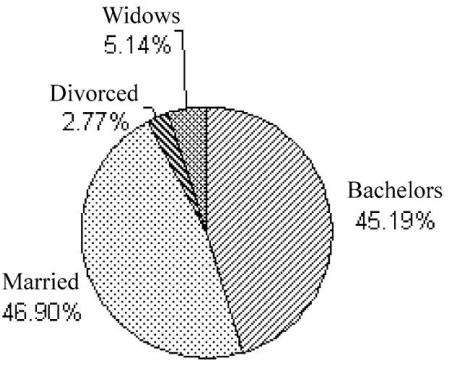
На следующей странице показана круговая диаграмма.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Гистограмма
- Графическое представление
- Пиктограмма
- Количественный график

П

Machine Translated by Google



Питман, Эдвин Джеймс Джордж

Питман, Эдвин Джеймс Джордж родился в Мельбурн (Виктория), Австралия, 1897 г. Он ходил в школу в Кенсингтонской государственной школе и Южный Мельбурнский колледж. Во время его На последнем году обучения ему была присуждена стипендия для изучения математики в Выселяске. и в Диксоне, а также в колледже Ормонд. Он получил степень в области письма (аналогично степень бакалавра) (1921), степень бакалавра наук (похоже на степень бакалавра) (1922 г.) и степень магистра степень (1923). Затем он был назначен временным профессором математики в Кентербери. Колледж Университета Новой Зеландии (1922–1923). вернулся в Австралию, где был назначен доцентом в Тринити и Ормонд-колледж, а также преподавал физику на полставки в Мельбурнском университете (1924–1924 гг.). 1925). В 1926 году Питман, EJG был назван профессор математики в Университете Тасмании, где он оставался до выхода на пенсию. в 1962 г. Умер в 1993 г. в Кингстоне (Тасмания).

Большой вклад Pitman, EJG в теория вероятностей занимается изучением поведение характеристической функции в окрестность нуля.

Избранные произведения Питмана, Эдвина Джеймса Джордж:

1937 Тесты значимости, которые могут быть применяется к выборкам из любых популяций. Дополнение, Дж. Рой. Стат. соц. сер. Б 4, 119–130.

1937 Тесты значимости, которые могут быть применяется к выборкам из любых популяций. II. Коэффициент корреляции тест. Дополнение, Дж. Рой. Стат. соц. сер. Б 4, 225–232.

1938 Тесты значимости, которые могут быть применимы к выборкам из любых популяций. III. Дисперсионный анализ тест. Биометрика 29, 322–335.

1939 Оценка местоположения и масштаба параметры. Биометрика 30, 391–421.

1939 г. Проверка гипотез о параметрах положения и масштаба. Биометрия 31, 200–215.

Оценка точки

Точечная оценка параметра населения позволяет получить уникальное значение , рассчитанное из образца. Это значение будет считаться оценка неизвестного параметра .

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим популяцию , в которой признак X изучал; Предположим, что форма распределения X известна, но параметр θ , от которого зависит это распределение, есть неизвестный.

Для оценки θ выборка размером n (X1, X2,...,Xn) берется из совокупности и функция G(X1, X2,...,Xn) создается так, что для конкретной реализации (x1, x2,...,xn) предоставит уникальное значение

Machine Translated by Google

как оценку значения параметра θ .

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В общем случае точечная оценка параметра популяции не совсем точно соответствует значению этого параметра. Там существует ошибка выборки , возникающая из-за тот факт, что часть населения была опущено. Эту ошибку можно измерить вычисление дисперсии или стандартного отклонения оценки , которая использовалась для оценить параметр населения.

ПРИМЕРЫ

Компания по производству лампочек хочет изучить среднюю продолжительность жизни его лампочки. Требуется выборка размером n (X1, X2,...,Xn) от производства света луковицы.

Функция G(X1, X2,...,Xn) , которую она будет использовать для оценкипараметр θ, соответствующий среднему значению μ совокупности, равен следует:

$$X^{-} = \sum_{i=1}^{n} \frac{Xi}{n} .$$

Эта оценка X^{-} μ является случайной переменная, и для конкретной реализации (x1, x2,...,xn) выборки принимает точное значение , обозначаемое x^{-}. Поэтому , например, при n = 5 можно получено (в часах):

(812, 1067, 604, 918, 895) .

Оценщик x^{-} становится:

$$x^{-} = \frac{812 + 1067 + 604 + 918 + 895}{5}$$

$$= \frac{4296}{5} = 859,2 .$$

Значение 859,2 является точечной оценкой средний срок службы этих лампочек.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Ошибка
- Оценка
- Оценщик
- Образец
- Выборка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

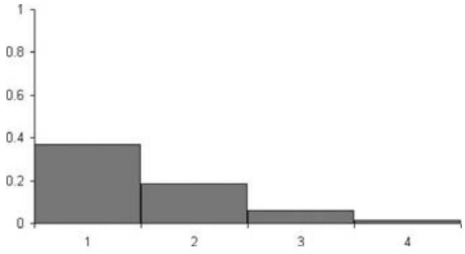
Леманн, Эрих Л., Казелла, Г.: Теория балльная оценка. Springer-Verlag, Нью-Йорк Йорк (1998)

Распределение Пуассона

Распределение Пуассона представляет собой дискретное распределение вероятностей. Это особенно полезно явления счета в единицу времени или космос.

Случайная величина X, соответствующая количество элементов, наблюдаемых на единицу время или пространство следует распределению Пуассона параметра θ, обозначаемого P(θ), если его функция вероятности :

$$(X = x) = x! \qquad \exp(\quad \theta) \cdot \theta P \ ^{\text{инс}} .$$



Распределение Пуассона, θ = 1

ИСТОРИЯ

В 1837 г. Poisson, SD опубликовал дистрибутив, носящий его имя. Он достиг этого распределения с учетом пределов биномиальное распределение.

П

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Ожидаемое значение распределения Пуассона по определению:

$$E[X] = \sum_{x=0} x \cdot P(X = x)$$

$$= \sum_{x=0} x \cdot \frac{e^{-\theta}\theta^{x}}{x!}$$

$$= \sum_{x=1} x \cdot \frac{e^{-\theta}\theta^{x}}{x!}$$

$$= \theta \cdot e^{\theta} \sum_{x=1} \frac{\theta^{x-1}}{(x-1)!}$$

$$= \theta \cdot e^{\theta} \cdot e^{\theta}$$

$$= \theta .$$

Дисперсия X равна : _

$$Var(X) = E[X^2] - (E[X])^2$$

$$= E[X(X-1) + X] - (E[X])^2$$

$$= E[X(X-1)] + E[X] - (E[X])^2 .$$

С

$$E[X(X-1)] = \sum_{x=0} \theta^{x} (x-1) \frac{e^{-\theta} \cdot \theta^{x}}{x!}$$

$$= \sum_{x=2} \theta^{x} (x-1) \frac{e^{-\theta} \cdot \theta^{x}}{x!}$$

$$= e^{-\theta} \cdot \theta^{2} \cdot \sum_{x=2} \frac{\theta^{x-2}}{(x-2)!}$$

$$= e^{-\theta} \cdot \theta^{2} \cdot e^{\theta}$$

$$= \theta^{2} ,$$

мы получаем:

$$Вар(X) = \theta^{2} + \theta - \theta^{2} = \theta .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Распределение Пуассона используется для определения следующие события, в том числе:

- Количество частиц, испускаемых радиоактивным веществом.
- Количество телефонных звонков, записанных центром тер.
- Количество несчастных случаев с застрахованное лицо.
- Количество прибытий на стойку.
- Количество бактерий в микроскопическом подготовка.
- Количество растений или животных на поверхности в природе определяется наблюдателем.

Когда θ стремится к бесконечности, пуассоновская распределение можно аппроксимировать нормальным распределением со средним значением θ и переменными анс θ.

ПРИМЕРЫ

Секретарь делает в среднем две ошибки на страницу. Какова вероятность иметь три ошибки на одной странице? Количество ошибок на странице, X, следует распределение Пуассона параметра θ = 2, и его функция вероятности тогда:

$$(X = x) = x! \quad \frac{\exp(-\theta) \cdot \theta^{x}}{x!}$$

$$= \frac{\exp(-2) \cdot 2^{x}}{x!} .$$

Вероятность совершения трех ошибок на страница равна:

$$P(X = 3) = \frac{e^{-2} \cdot 2^{3}}{3!} = 0{,}1804 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Биномиальное распределение
- Дискретное распределение вероятностей
- Нормальное распределение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Poisson, SD: Recherches sur la probabilité des jugements en matière criminelle

Machine Translated by Google

et en matière civile. In: Procédés des Règles Générales du Calcul des Probabil ités. Башелье, Imprimeur-Libraire pour les Mathématiques, Париж (1837 г.)

Пуассон, Симеон-Дени

Пуассон, Симеон-Дени (1781–1840), несмотря на то, что он происходил из скромной провинциальной семьи, смог продолжить учебу в Политехнической школе в Париже, где впоследствии стал профессором.

С 1815 года он также преподавал в Сорбонне и в том же году был избран в Академию наук.

Пуассона интересовали исследования различных области: механика, физика и теория вероятностей. В 1837 году он опубликовал статью под названием «Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités». В своих многочисленных работах он представил основы статистического метода для социальных наук.

Избранные произведения Симеона-Дени Пуассона:

1824 Sur la probabilité des résultats moyens des наблюдения. Connais sance des temps pour l'an 1827, стр. 273–302.

1829 г. Suite du Mémoire sur la probabilité du résultat moyen des Observations, inséré dans la connaissance des temps de l'année 1827. Con naissance des temps pour l'an 1832, стр. 3–22. 1836a Note sur la loi des grands nombres.

Comptes rendus hebdomadaires des séances de l'Académie des Sciences 2: стр. 377–382.

1836b Note sur le calcul des probabilités. Comptes rendus hebdomadaires des сеансы Академии наук 2: стр. 395–400.

1837 г. Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités. Башелье, Imprimeur Libraire pour les Mathématiques, Париж.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

распределение Пуассона

Объединенная дисперсия

Объединенная дисперсия используется для оценки значения дисперсии двух или более совокупностей, когда соответствующие дисперсии каждой совокупности неизвестны, но могут рассматриваться как равные.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

На основе k выборок размерностей n1, n2,..., nk объединенная дисперсия S2 определяется оценивается по формуле:

$$S2_{п} = \frac{\sum_{я=1}^{К} \sum_{Дж=1}^{ни} xij - \bar{x}i.2}{\sum_{я=1}^{К} ни - к},$$

где xij — j -е наблюдение выборки i и x¯i. среднее арифметическое выборки i.

Зная соответствующие дисперсии каждой выборки, объединенную дисперсию также можно определить по формуле:

$$S2_{п} = \frac{\sum_{я=1}^{К} (ни - 1) \cdot S2_{я}}{\sum_{я=1}^{К} ни - к},$$

П

Machine Translated by Google

где $S2_{я}$ — дисперсия выборки i:

$$S2_{я} = \frac{\sum_{j=1}^{ни} (xij - \bar{x}i.2)}{ни - 1}.$$

ПРИМЕРЫ

Рассмотрим два образца компьютеров разных марок, для которых мы отметили время (в часов) до первой проблемы:

Бренд 1	Бренд 2
2800	2800
2700	2600
2850	2400
2650	2700
2700	2600
2800	2500
2900	
3000	

Количество выборок k равно 2, и размерность марок выборки равна, соответственно:

$$п1 = 8 ,$$

$$п2 = 6 .$$

Расчет дисперсии первого образец дает нам

$$S2_{1} = \frac{\sum_{j=1}^{n1} (x1j - \bar{x}1.)^2}{п1 - 1},$$

где

$$\bar{x}1. = \frac{\sum_{j=1}^{n1} x1j}{n1} = \frac{22400}{8} = 2800$$

$$S2_{1} = \frac{95000}{8 - 1} = 13571,428 .$$

Для второго примера получаем:

$$S2_{2} = \frac{\sum_{j=1}^{п2} (x2j - \bar{x}2.)^2}{п2 - 1}$$

с

$$\bar{x}2. = \frac{\sum_{j=1}^{п2} x2j}{п2} = \frac{15600}{6} = 2600$$

$$S2_{2} = \frac{70000}{6 - 1} = 14000 .$$

Если предположить, что неизвестные дисперсии две популяции $\sigma2_{1}$ и $\sigma2_{2}$ тождественны, то можно вычислить оценку $\sigma2 = \sigma2_{1} = \sigma2_{2}$ по объединенной дисперсии:

$$S2_{п} = \frac{\sum_{я=1}^{2} (ни - 1) \cdot S2_{я}}{\sum_{я=1}^{2} ни - 2}$$

$$= \frac{(8 - 1)\cdot 13571,428 + (6 - 1)\cdot 14000}{8 + 6 - 2}$$

$$= 13750 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Дисперсия

Население

Население определяется как совокупность статистических единиц одного и того же характера, количественная информация о которых нас интересует.

Население составляет эталонную вселенную при изучении данной статистической задачи.

ПРИМЕРЫ

Государство гражданина, группа деревьев в лесу, рабочие на фабрике и цены на потребительские товары — все это образует дискретное население.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Образец

Распространенность

См. показатель распространенности.

Уровень распространенности

Распространенность заболевания – это количество лиц, затронутых в статистической совокупности в данный момент времени. Это понятие близко к понятию «акции».

Коэффициент распространенности болезни – это доля больных в популяции в данный момент времени.

ИСТОРИЯ

Фарр, Уильям, пионер в использовании статистики в эпидемиологии и создатель концепций уровня смертности и вопросов и ответов, показали, что распространенность заболевания эквивалентно произведению инцидентности и длительность болезни.

MacMahon, B. and Pugh, TF (1970) проиллюстрировали взаимосвязь уровня заболеваемости и коэффициент распространенности с использованием данных, собранных между 1948 и 1952 гг. лейкемия среди белого населения Бруклин, Нью-Йорк. Уровень заболеваемости составил 32,5 случая на миллион жителей в год. в год, а показатель распространенности составил 6,7 случая на миллион жителей. Продолжительность таким образом, острый лейкоз должен был составлять 0,21 года или 2,5 месяца:

$$\text{Длительность острый лейкоз} = \frac{\text{распространенность ела}}{\text{уровень заболеваемости}} = \frac{6.7}{32,5} \text{ год} = 0,21 \text{ года}.$$

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Коэффициент распространенности определяется следующим образом:

$$\text{Коэффициент распространенности} = \frac{\text{количество страдающих}}{\text{размер Население}}.$$

Распространенность и заболеваемость заболевания примерно связаны с между собой средней продолжительностью болезни (то есть средней продолжительностью выживания или средняя продолжительность лечения). Это отношение написано:

$$\text{Уровень распространенности} \approx (\text{заболеваемость}) \cdot \text{средняя продолжительность болезни}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Распространенность отличается от риска. Риск количество, имеющее прогностическую ценность, поскольку оно касается фиксированного периода наблюдения и содержит, благодаря этому, информацию, однородную во времени. Напротив, распространенность не имеет отношения к истории какой-либо болезни в популяции: она не может дать никаких указаний на риск в той мере, в какой он агрегируется, при данном момент, недавно диагностированные случаи и другие в далеком прошлом. Обратите внимание, что чем больше средняя выживаемость, если болезнь неизлечима, или чем дольше период выздоровления, если болезнь излечим, тем выше распространенность болезнь.

Наконец, обратите внимание, что уровень распространенности часто относится не к четко определенной болезни, а к интервал значений биологического параметра. Например, показатель распространенности гиперхолестерина может быть определен как доля лиц, имеющих уровень холестерина выше 6,5 ммоль/л в данной популяции в данный момент. Таким образом, процентили биологического параметра, рассматриваемого как измерение доли особей, для которых параметры имеют значение

П

Machine Translated by Google

внутри определенного интервала, можно интерпретировать как коэффициент распространенности. Например, показатель распространенности по отношению к значениям меньше или равный медиане (или процентилю 50) биологического параметра, составляет 50%.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Атрибутивный риск
- Предотвратимый риск
- Причина и следствие в эпидемиологии
- Уровень заболеваемости
- Шансы и отношение шансов
- Относительный риск
- Риск

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Корнфилд, Дж.: Метод оценки сравнительных показателей на основе клинических данных. Применение при раке легких, молочной железы и шейки матки. Дж. Натл. Рак инст. 11, 1269–75 (1951)

Лилиенфельд, А.М., Лилиенфельд, Д.Э.: Основы эпидемиологии, 2-е изд. Кларен Дон, Оксфорд (1980)

МакМахон, Б., Пью, Т. Ф.: Эпидемиология: принципы и методы. Литтл Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Морабия, А .: L'Epidémiologie Clinique. Издания «Que sais-je?». Presses Universitaires de France, Париж (1996)

Вероятность

Мы можем определить вероятность события либо с помощью относительных частот, либо с помощью аксиоматического подхода.

В первом подходе мы предполагаем, что случайный эксперимент многократно повторяется в те же условия.ForeacheventAdefined в выборочном пространствемы определяем nA как количество раз, когда событие A происходило в течение первых n повторений эксперимента . В этом случае вероятность события A, обозначаемая P (A), определяется по формуле:

$$P(A) = \lim_{n} \frac{nA}{n},$$

что означает, что P (A) определяется как предел относительно числа раз, когда событие A произошло относительно общего числа повторений.

Во втором подходе для каждого события A мы принимаем, что существует вероятность A, P (A), удовлетворяющая следующим трем аксиомам:

1. 0 P (A) 1, 2. P () = 1, 3. Для каждой последовательности взаимоисключающих событий A1, A2,... (то есть событий Ai Aj = φ, если i = j) :

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

ИСТОРИЯ

Первые азартные игры знаменуют собой начало истории вероятности, и мы можем утверждать, что они восходят к появлению Homo sapiens.

Происхождение слова «опасность» вызывает меньше сомнений. Согласно Кендаллу, Морису Джорджу (1956), он был завезен в Европу во время Третьего крестового похода и производный от него. от арабского слова «аль зхар», что означает «кость».

Бензекри, JP (1982) дает ссылку на Франка д'Ориента по имени Гийом де Тир, согласно которому слово «опасность» было связано с названием замка в Сирии. Этот

Machine Translated by Google

замок был объектом осады, во время которой захватчики изобрели игру в кости, давшую название этим играм.

Следы азартных игр можно найти в древних цивилизациях. Рисунки и предметы, найденные в египетских могилах времен Первой династии (3500 г. до н.э.), показывают, что в то время уже играли в эти игры.

Согласно греческому историку Геродоту, азартные игры были изобретены Паламедом во время осады Трои.

Игра в астрагалы для собак была одной из первых азартных игр. Таранная кость небольшая кость стопы, симметричная относительно вертикальной оси. Греки, а позже и римляне высоко ценили эту игру, возможно, согласно Kendall, MG (1956), изначальная игра в кости. Очень старые астрагалы были найдены в Египте, но более древние, изготовленные из терракоты, были обнаружены в северном Ираке.

Они датируются третьим тысячелетием до нашей эры. Другой, тоже интертеррakотовый, датируемый тем же временем, был найден в Индии.

Существует множество доказательств, демонстрирующих Как в древности играли в игры с костями.

Для Дэвида, Ф.Н. (1955), две гипотезы помогают объяснить этот разрыв. Согласно первой гипотезе игральные кости вообще были деформированы. Согласно второй гипотезе, игра в кости носила религиозный характер. Можно было задавать богам вопросы: результат данного броска костей открывал ответы на вопросы, которые были заданы богам.

Среди азартных игр весьма древнее происхождение имеют и игральные карты. Мы знаем, что ими пользовались в Китае, Индии, Аравии и Египте. На Западе мы находим первые свидетельства существования карт в Венеции в 1377 г., в Нюрнберге в 1380 г. и в Париже в 1397 году. Карты Таро — самые древние игральные карты.

До 15 века, по словам Кендалла М.Г. (1956), церковь и короли боролись против практики игры в кости.

В качестве доказательства этого можно упомянуть законы Людовика IX, Эдуарда III и Генриха VIII. Карточные игры были добавлены к списку незаконных занятий, и мы находим примечание о запрете карточных игр, изданное в Париже в 1397 году. Несмотря на запреты, азартные игры присутствовали без перерыва, от римлян до эпохи Возрождения, во всех социальные классы.

Неудивительно, что они лежат в основе математических работ о вероятности. На Западе первые сочинения на эту тему приходят из Италии. Они были написаны Кардано в его трактате Liber de Ludo Aleae, опубликованном посмертно в 1663 году, и Галилео Галилеем в его Sopra le Scoperte dei Dadi. В работах Галилея мы находим

Принцип равновероятности игры в кости.

В остальной Европе эти проблемы, связанные с опасностями, также интересовали математиков. Гюйгенс из Нидерландов опубликовал в 1657 году работу под названием De Ratiociniis in Aleae.

Хорошо , что, согласно Холгейту П. (1984), интерес к вероятностям распространился на другие области; Теория игр должна была бы развиваться параллельно этому в связи с динамикой тела. Работы Гюйгенса оказали значительное влияние на работы двух других известных математиков: Бернулли, Якоба , и де Муавра, Абрахама.

Но именно во Франции, благодаря Паскалю, Блезу (1623–1662) и де Ферма, Пьеру (1601–1665), теория вероятностей действительно оформилась. Согласно Тодхантеру, И. (1949), Паскаль, Б., получил от знаменитого игрока Гомбо, Антуана, рыцаря Мере, просьбу решить следующую задачу: в игре

П

Machine Translated by Google

есть два противника Первый побеждает n из m игр, в то время как второй выигрывает p из m игр, что является шансами одного из них выиграть игру (при условии, что он выигрывает первый, сыгравший m игр).
Паскаль, Б. связался с deFermat,P., который нашел решение. Паскаль сам открыл рекуррентная формула, показывающая тот же результат.
Согласно Бензекри, Дж. П. (1982), в то время, когда Запад был свидетелем падения Римской империи и наступления христианской эры, Восток был вовлечен в интенсивную научную и художественную деятельность. В этих обстоятельствах восточные ученые, такие как Омар Хайям, вероятно, открыли вероятностные правила.

Существует аспект вероятности, представляющий особый интерес для математиков: проблемы комбинаторного анализа.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Аксиоматическая основа вероятности

Рассмотрим случайный эксперимент с выборочным пространством , содержащим n элементов:

$$= \{x1, x2,\ldots,xn\} .$$

Каждому простому событию xi можно сопоставить вероятность P (xi) , обладающую следующими свойствами:

1. Вероятности P (xi) неотрицательны:

$$P(xi) \quad 0 \text{ для } i = 1, 2,\ldots,n.$$

2. Сумма вероятностей xi для i перехода от 1 к n равна 1:

$$\sum_{я=1}^{н} P(кси) = 1 \quad .$$

3. Функция вероятности Пизы, которая присваивает число от 0 до 1 каждому простому событию случайного эксперимента:

$$П: \quad [0, 1] .$$

Свойства вероятностей 1.

Вероятность выборочного пространства — это наибольшая вероятность, которую можно связать к событию:

$$P() = 1 \quad .$$

2. Вероятность невозможного события равна 0:

$$\text{Если } A = \phi , \quad \text{тогда } P(A) = 0 .$$

3. Пусть A^{-} — дополнение A в , затем вероятность A^{-} равна 1 минус вероятность A:

$$P\,A^{-} = 1 \quad P(A).$$

4. Пусть A и B — два несовместных события (A B = φ). Вероятность A B равна сумме вероятностей A и B:

$$P(A \quad B) = P(A) + P(B).$$

5. LetA и Bbetдва события. Вероятность A B равна:

$$P(A \quad B) = P(A) + P(B) \quad P(A \quad B).$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Условная возможность
- Непрерывное распределение вероятностей
- Дискретное распределение вероятностей
- Событие
- Случайный эксперимент
- Образец пространства

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бензекри, JP: Histoire et préhistoire de l'analyse des données. Дюно, Париж (1982)

Machine Translated by Google

Дэвид, FN: Игра в кости и игры (примечание о история вероятности). Биометрика 42, 1–15 (1955)

Холгейт, П.: Влияние работы Гюйгенса в динамике о его вкладе в вероятность. Междунар. Стат. 52 , 137–140 (1984)

Кендалл, MG: Исследования по истории вероятность и статистика: II. Начало исчисления вероятностей. Биометрия 43, 1–14 (1956).

Стиглер, С.: История статистики, Измерение неопределенности до 1900. Белкнап, Лондон (1986).

Тодхантер, И.: История математического теория вероятностей. Издательство Челси Компания, Нью-Йорк (1949)

Распределение вероятностей

Синоним функции вероятности, но используется гораздо чаще, чем «вероятность функция» , чтобы описать распределение, дающее вероятность появления значения случайной величины X при значении x.

ИСТОРИЯ

Именно в XVII веке началось систематическое изучение проблем, связанных со случайными начались явления. Выдающийся физик Галилей уже пытался изучить ошибки физических измерений, учитывая эти как случайные и оценивая их вероятность. В этот период теория страхования также возник и был основан на анализе законов, управляющих случайными явлениями. таких как заболеваемость, смертность, несчастные случаи и т. Однако сначала необходимо было изучить более простые явления, такие как азартные игры. Они обеспечивают особенно простой и понятный модели случайных явлений, позволяющие один наблюдать и изучать конкретные законы которые управляют ими; более того, возможность многократное повторение одного и того же эксперимента позволяет экспериментально проверить эти законы.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

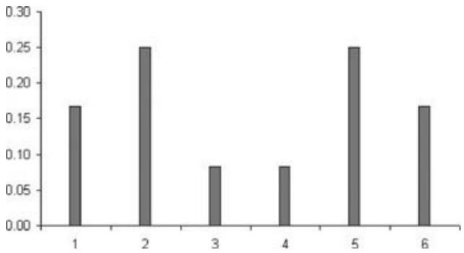
Функция P(b) = P(X = b), где b варьируется в зависимости от возможных значений дискретная случайная величина X, называется функция вероятности X. Поскольку P(X = b) есть всегда положительна или равна нулю, функция вероятности также положительна или равна нулю.

Функция вероятности представлена на система осей. Различные значения b X равны _ по оси абсцисс, изображения P(b) по оси ординат. Вероятность P (b) представлена прямоугольниками шириной равной единице и высота, равная вероятности b. C X должен принимать хотя бы одно из значений b, сумма P(b) должна быть равна 1, что означает что сумма поверхностей прямоугольников должен быть равен 1.

ПРИМЕРЫ

Рассмотрим случайный эксперимент , заключающийся в броске фиксированной кости. Рассмотрим случайную величину X , соответствующую количеству полученных точек. Функция вероятности P(X = b) задается вероятностями связанные с каждым значением X :

б	1	2	3	4	5	6
P(б)	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{6}$



П

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Непрерывное распределение вероятностей
Функция плотности
Дискретная функция распределения
Дискретное распределение вероятностей
Функция распределения
Вероятность
Функция вероятности
Случайная переменная

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Джонсон, Н.Л., Коц, С.: Распределения в статистике: дискретные распределения. Уайли, Нью-Йорк (1969)

Джонсон, Н.Л., Коц, С.: Распределения в статистике: непрерывное одномерное распределение буции, тт. 1 и 2. Уайли, Нью-Йорк (1970).

Джонсон, Н.Л., Коц, С., Балакришнан, Н.: Дискретные многомерные распределения, Джон Вили (1997)

Ротшильд, В., Логотетис, Н.: Распределения вероятностей. Уайли, Нью-Йорк (1986)

Функция вероятности

Функция вероятности дискретной случайной величины – это функция, которая связывает каждое значение этой случайной величины к ее вероятности.

См. распределение вероятностей.

ИСТОРИЯ

См. вероятность, распределение вероятностей.

п -Ценить

Значение p определяется как вероятность, рассчитанная по нулевой гипотезе, того, что результат, столь же экстремальный, как наблюдаемое значение в выборке , или вероятность получения результата, по крайней мере столь же экстремального, как данная точка данных, при условии, что точка данных была результатом только случайности.

ИСТОРИЯ

Значение p было введено Гиббенсом и Праттом в 1975 году.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Давайте проиллюстрируем случай проверки гипотезы , сделанной для оценки среднего ; тот же принцип применяется для любой другой оценки; только обозначения разные.
Предположим, что мы хотим проверить следующие гипотезы:

$$H0 : \mu = \mu0$$

$$H1 : \mu > \mu0 ,$$

где μ представляет собой среднее значение нормально распределенной совокупности с известным стандартным отклонением σ. Выборка размерности n дает наблюдаемое среднее значение $\bar{x}$.
Рассмотрим случай, когда $\bar{x}$ превышает μ0 , и вычислим вероятность получения оценки $\hat{\mu}$ больше или равной $\bar{x}$ при нулевой гипотезе μ = μ0. Этой вероятности соответствует значение p . Таким образом:

$$p = P(\hat{\mu} \quad \bar{x} | \mu = \mu0) .$$

Стандартное случайное значение Z , определяемое как

$$Z = \text{следует нормальному распределению} \quad \frac{\hat{\mu} \quad \mu0}{\sigma \sqrt{n}}$$

среднее 0 со стандартным отклонением 1. Введение подставляя эту переменную в выражение p, находим:

$$p = PZ \quad \frac{\bar{x} \quad \mu0}{\sigma \sqrt{n}} ,$$

Machine Translated by Google

который в этой форме может быть прочитан в обычном стол.

Для уровня значимости α сравнение p с α позволяет принять решение об окончательном отклонении нулевой гипотезы. Если: • p α: мы отвергаем нулевую гипотезу H0 в пользу альтернативной гипотезы H1; • p > α: мы не отвергаем нулевую гипотезу H0.

Мы также можем вычислить верхнюю границу допустимой области H0, зная значение Cα, такой как

$$P(\hat{\mu} \quad C\alpha \,|\mu = \mu 0) = \alpha .$$

В зависимости от стандартной нормальной случайной величины zα значение Cα выражается как:

$$C\alpha = \mu 0 + z\alpha \cdot \quad \frac{\sigma}{\sqrt{n}} \cdot$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ Мы часто используем значение p , чтобы понять результат проверки гипотезы.
Он используется для одностороннего теста, когда гипотезы имеют вид:

$$H0 : \mu = \mu 0$$

$$H1 : \mu > \mu 0 ,$$

или

$$H0 : \mu = \mu 0$$

$$H1 : \mu < \mu 0 ,$$

или

$$H0 : \mu \quad \mu 0$$

$$H1 : \mu > \mu 0 ,$$

или

$$H0 : \mu \quad \mu 0$$

$$H1 : \mu < \mu 0 .$$

Значение p можно интерпретировать как наименьший уровень значимости , для которого нельзя отвергнуть нулевую гипотезу .

ПРИМЕРЫ

Предположим, что мы хотим провести одностороннюю проверку гипотезы.

$$H0 : \mu = 30$$

против $H1 : \mu > 30$

для нормально распределенной совокупности со стандартным отклонением σ = 8. Выборка размером n = 25 дает наблюдаемое среднее значение 34 . Поскольку $\bar{x}$ (= 34) явно превышает μ = 30, мы должны попытаться принять альтернативную гипотезу, согласно которой μ > 30.
Рассчитайте вероятность получения наблюдаемого среднего значения 34 при нулевой гипотезе μ = 30. Это значение p :

$$p = P(\bar{X} \quad 34 \,|\mu = 30)$$

$$= \Pi \quad \frac{\bar{X} \quad 30}{\frac{8}{5}} \quad \frac{34 \quad 30}{\frac{8}{5}}$$

$$= P(Z \quad 2{,}5) ,$$

где $Z = \frac{\bar{X} \quad \mu}{\frac{\sigma}{\sqrt{n}}}$, нормально распределено

со средним значением 0 и стандартным отклонением 1; p = 0,0062 по обычной таблице.
Для уровня значимости α = 0,01 мы отклоняем нулевую гипотезу для альтернативной гипотезы μ > 30. Мы можем вычислить верхнюю границу допустимой области H0 , Cα , например

$$P(\bar{X} \quad C\alpha \,|\mu = 30) = 0{,}01$$

8 и находим $C\alpha = 30 + 1{,}64 \cdot = 32{,}624$. 5

П

Machine Translated by Google

Это подтверждает, что мы действительно должны отвергнуть нулевую гипотезу μ = 30, когда наблюдаемое среднее равно 34.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Область приема
- Проверка гипотезы
- Обычный стол
- Уровень значимости

ССЫЛКИ

Гиббенс, Р. Дж., Пратт, Дж. В.: Интерпретация P-значений и методология. Являюсь. Стат. 29 (1), 20–25 (1975)

Machine Translated by Google

Сюжет QQ
(Квантиль к квантильному графику)

График QQ, или график от квантиля к квантилю, представляет собой график, который проверяет соответствие между эмпирическим распределением и заданным теоретическим распределением.

Одним из методов, используемых для проверки нормальности ошибок регрессионной модели, является построение QQ-графика остатков. Если точки выровнены на линии x = y, то данные распределены нормально.

ИСТОРИЯ

Метод графического представления , известный как график QQ, появился в начале 1960-х годов. С тех пор он стал необходимым инструментом в анализе данных и/или остатков, потому что он обеспечивает очень много информации и легко интерпретируется.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

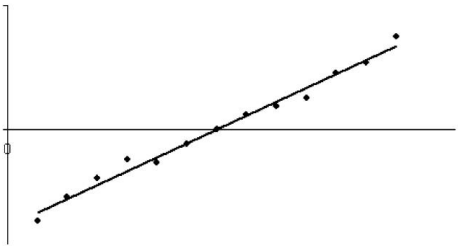
Чтобы построить график QQ, мы делаем два шага.

1. Расположите данные x1, x2,...,xn в порядке возрастания:

   x[1] x[2] ... x[n] .

2. Свяжите каждую точку данных x[i] с i/(n+ 1)-квантилем qi стандартного нормального распределения. Нанесите на график по оси ординат упорядоченные данные xi и по оси абсцисс квантили qi.

Если переменные имеют одинаковое распределение, то графическое представление между квантилями первой переменной относительно квантилей второго распределения будет линией наклона 1. Так, если данные xi распределены нормально, точки на графике должны быть почти выровнены по линии уравнения xi = qi.



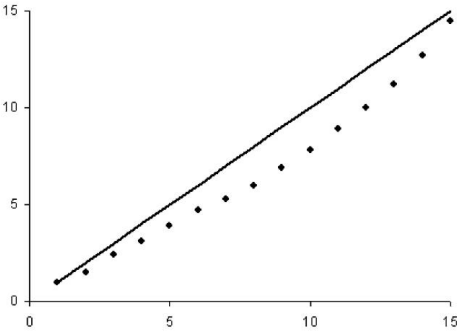
ОБЛАСТИ И ОГРАНИЧЕНИЯ График

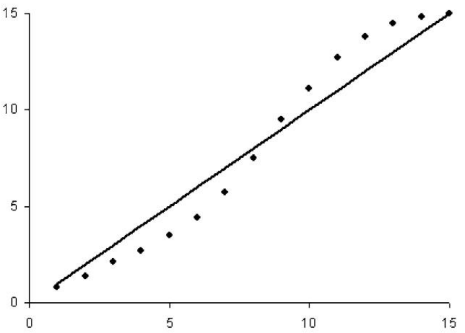
QQ используется для проверки того, следуют ли данные определенному распределению или имеют ли два заданных набора данных одинаковое распределение. Если распределения одинаковые, график представляет собой линию. Чем дальше полученный результат от диагонали 45 , тем дальше эмпирическое распределение от теоретического.

крайние точки имеют большую изменчивость, чем те, что в центре распределения. Таким образом U-образный график означает, что одно распределение смещено относительно другого. S-образный график указывает на то, что распределения представляют большее влияние экстремальных значений на другое распределение (длинный хвост).

Machine Translated by Google

Частое использование графиков QQ, когда кто-то хочет знать поведение остатков во время простой или множественной линейной регрессии , точнее, чтобы узнать, является ли их распределение нормальным.





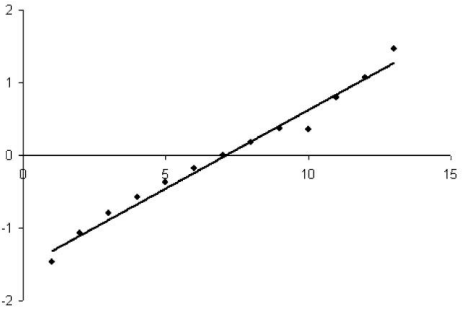
ПРИМЕРЫ

Следующая таблица содержит остатки (в порядке возрастания) линейной регрессии с 13 наблюдений. Мы связываем их с quantilesqi , как описано выше.

График QQ представлен на следующем рисунке. Мы видим, что кроме двух точек, вероятно выбросы, точки хорошо выровнены.

Таблица: остатки и квантили

я	эй	Ци
1	1,62	1,47
2	0,66	1,07
3	0,62	0,79
4	0,37	0,57
5	0,17	0,37
6	0,12	0,18
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



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Анализ остатков
- Анализ данных
- Исследовательский анализ данных
- Множественная линейная регрессия
- Распределение вероятностей
- квантиль
- Остаток
- Простая линейная регрессия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Чемберс, Дж. М., Кливленд, В. С., Кляйн Эр, Б., Тьюки, П.А.: Графические методы для Анализ данных. Уодсворт, Белмонт, Калифорния (1983)

Хоаглин, Д.С., Мостеллер, Ф., Тьюки, Дж.В. (ред.): Понимание надежных и Исследовательский анализ данных. Уайли, Нью Йорк (1983)

Machine Translated by Google

Уилк, М.Б., Ганадесикан, Р.: Вероятность
методы построения графиков для анализа данных.
Биометрика 55, 1–17 (1968).

Качественная категориальная переменная

Качественная категориальная переменная — это переменная с модальностями в форме категории
гори, такие как, например, «мужчина» и «женщина» переменной «пол»; или категории «красный», «оранжевый», «зеленый», «синий», «синий» и «фиолетовый» переменной «цвет».
Модальности качественного категориального переменная может быть представлена номинальным шкале или в порядковой шкале.

Примером качественной категориальной переменной, имеющей порядковую шкалу, является качественная профессиональная переменная с категориями «квалифицированные», «полуквалифицированные» и «неквалифицированные».

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Категория
- Переменная

квантиль

Квантиль измеряет положение (или центральное тенденция) и не обязательно пытаться определить центр распределения наблюдений.

ции, а для описания конкретной позиции.
Это понятие является расширением понятия медианы (которая делит распределение наблюдения на две части). Наиболее часто используемые квантили:

- Квартили, разделяющие совокупность наблюдения на четыре части,
- Децили, разделяющие совокупность наблюдения на десять частей,
- Centiles, которые отделяют набор наблюдения на сто частей.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Расчет квантилей имеет смысл
только для количественной переменной , которая может принимать значения на определенном интервале.

Понятие квантиля указывает на разделение распределения наблюдений в
произвольное количество частей. Обратите внимание, что чем больше наблюдений, тем больше
сложное разделение распределения может быть.

Обычно квартили можно использовать для любого распределения. Вычисление децилей и, в конечном счете, центилей требует относительно большого количества наблюдений для получения достоверной интерпретации.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- дециль
- Мера местоположения
- медиана
- Процентиль
- Квартиль

Количественный график

Количественные графики используются для представления и суммировать поступающую числовую информацию из изучения категориального количественного переменная.

Наиболее часто используемые типы количественных графиков:

- Гистограмма
- Пиктограмма
- Круговая диаграмма

ИСТОРИЯ

Бенигер, младший и Робин, Дж. Л. (1978) возвращаются историческое развитие количественных графики с 17 века. Шмид, CF
и Schmid, SE (1979) обсуждают многочисленные формы количественных графиков.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Гистограмма

Графическое представление

Пиктограмма

Круговая диаграмма

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бенигер, Дж. Р., Робин, Д. Л.: Количественные графики в статистике: краткая история. Являюсь. Стат. 32, 1–11 (1978)

Шмид, К.Ф., Шмид, С.Э.: Справочник по графическому представлению, 2-е изд. Уайли, Нью-Йорк (1979)

Количественная переменная

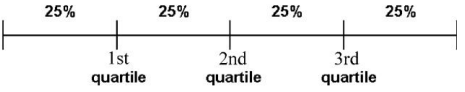
Количественная переменная - это переменная с числовыми модальностями. Например, вес, размер, возраст, скорость и время являются количественными переменными. Мы различаем дискретные переменные (например, количество детей в семье) от непрерывных переменных (например, длина прыжка).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Переменная

Квартиль

Квартили – это меры местоположения распределения наблюдений. Квартили разделяют распределение на четыре части. Таким образом, для данного распределения имеется три квартиля.
Между каждым квартилем мы находим 25% от общего числа наблюдений:



Обратите внимание, что второй квартиль равен медиане .

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Расчет квартиля аналогичен расчету медианы . Когда у нас есть все наблюдатели

В зависимости от вариаций квартили рассчитываются следующим образом: 1. n наблюдений необходимо расположить в виде частотного распределения.

2. Квартили соответствуют наблюдениям, для которых относительная кумулятивная частота превышает 25%, 50% и 75%. Некоторые авторы предлагают следующую формулу, позволяющую с точностью определить значение разных квартилей:
Вычисление j-го квартиля: Пусть i – целая часть j·(n+1) – дробная часть j·(n+1) $\frac{}{4}$ и к $\frac{}{4}$.
Пусть xi и xi+1 — значения наблюдений соответственно в i-й и (i + 1)-й позициях (при расположении наблюдений в порядке возрастания).

j -й квартиль

$$Qj = xi + k \cdot (xi+1 \quad xi) .$$

Когда у нас есть наблюдения, сгруппированные в классы, квартили определяются следующим образом:

1. Определите класс, в котором находится квартиль. найденный:
   - Первый квартиль: класс, для которого относительная кумулятивная частота превышает 25%.
   - Второй квартиль: класс, для которого относительная кумулятивная частота превышает 50%.
   - Третий квартиль: класс, для которого относительная кумулятивная частота превышает 75%.
2. Вычислить значение квартиля в зависимости от предположения, согласно которому наблюдения равномерно распределены в каждом классе; j -й квартиль Qj равен:

$$Qj = Lj + \frac{н \cdot 4 _ \quad конечная}{fj} \cdot cj ,$$

Machine Translated by Google

где Lj — нижняя граница класса квартиль Qj, n – общее количество наблюдений, finf – сумма частот ниже класса квартиля, fj – частота класса квартиля Qj, и cj – размер интервала класса квартиль Qj.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Расчет квартилей имеет смысл только для количественной переменной , которая может принимать значения на определенном интервале.

Квартиль подобен медиане . Это также на основе ранга наблюдения, а не значений. Выброс будет иметь лишь небольшое влияние по квартильным значениям.

ПРИМЕРЫ

Давайте сначала возьмем пример, когда у нас есть десять наблюдений (n = 10):

1244555679

Несмотря на то, что квартили не могут быть рассчитаны для небольшого числа наблюдений (мы должны быть очень осторожны при их интерпретации), мы изучим этот случай, чтобы понимать принципы расчета.

Первый квартиль Q1 находится на позиции = $\frac{n+1}{4}$ 2,75. Q1 – три четверти расстояния между вторым и третьим наблюдателями.

ции (которые мы будем называть x2 и x3) . Мы можем рассчитать Q1 следующим образом:

$$Q1 = x2 + 0{,}75 \cdot (x3 \quad x2)$$
$$= 2 + 0{,}75 \cdot (4 \quad 2)$$
$$= 3{,}5 .$$

Второй квартиль Q2 (который является медианой) равен находится на позиции 2 · $\frac{n+1}{4}$ (или $\frac{n+1}{2}$), который равно для нашего примера 5.5.

$$Q2 = x5 + 0{,}5 \cdot (x6 - x5)$$
$$= 5 + 0{,}5 \cdot (5 \quad 5)$$
$$= 5 .$$

Третий квартиль Q3 находится в позиции 3 · $\frac{n+1}{4}$ = 8,25. Q3 равно:

$$Q3 = x8 + 0{,}25 \cdot (x9 - x8)$$
$$= 6 + 0{,}25 \cdot (7 \quad 6)$$
$$= 6{,}25 .$$

Значения 3,5, 5 и 6,25 разделяют наблюдения на четыре равные части.
Для второго примера рассмотрим следующую таблицу частот , представляющую число детей в семье на 200 семей составляет:

Ценить (число из дети)	Родственник (число из семьи)	Родственник частый су	Накопленный частота
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
Общий	200	1,00	

Первый квартиль равен этим наблюдениям которые имеют относительную кумулятивную частоту превышает 25%, что соответствует двум детей (поскольку относительная совокупная частота для двоих детей идет от 22 до 47%, в том числе 25%). Второй квартиль равен трем детям, потому что относительная кумулятивная частота для трех детей составляет от 47 до 74%, в том числе

Machine Translated by Google

50%. Третий квартиль равен четырем детям.
потому что относительная кумулятивная частота для
четырех детей увеличивается с 74 до 95%, что
включает 75%.

Квартили 2, 3 и 4 делят 200 семей на
четверти. Мы можем сгруппировать 50 из 200 семей.
семей в первый квартал с нулем, единицей,
или двое детей, 50 во втором квартале
с двумя или тремя детьми, 50 в третий
четверть с тремя-четырьмя детьми и 50
в четвертую четверть с четырьмя, пятью или шестью
дети.

Теперь рассмотрим пример с вычислением
квартилей по частоте
распределение непрерывной переменной, где
наблюдения группируются в классы.
Следующая таблица частот представляет
прибыль (в тысячах франков) 100 магазинов:

Выгода (тысячи франков)	Куму поздно качество	Куму поздно качество	Родственник частота
100–200	10	10	0,1
200–300	20	30	0,3
300–400	40	70	0,7
400–500	30	100	1,0
Общий	100		

Класс, содержащий первый квартиль, является
класс 200–300 (с относительной кумулятивной
частотой 25%).
Учитывая, что наблюдения равномерно
распределены в каждом классе, получаем
для первого квартиля следующие значения:

1-й квартиль

$$= 200 + \frac{100 \cdot \frac{1}{4} - 10}{20} \cdot 100$$

$$= 275 .$$

Класс, содержащий второй квартиль,
класс 300-400. Стоимость второго _
квартиль равен:

2-й квартиль

$$= 300 + \frac{100 \cdot \frac{2}{4} - 30}{40} \cdot 100$$

$$= 350 .$$

Класс, содержащий третий квартиль, является
класс 400–500. Значение третьего квартиля
равно:

3-й квартиль

$$= 400 + \frac{100 \cdot \frac{3}{4} - 70}{30} \cdot 100$$

$$= 416{,}66 .$$

Мы можем сделать вывод, что 25 из 100 магазинов
иметь прибыль от 100000 до 275000
франков, 25 имеют прибыль от 275000
и 350000 франков, 25 имеют прибыль между
350000 и 416066 франков, а 25 имеют
прибыль от 416066 до 500000 франков.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- децил
- Мера местоположения
- медиана
- Процентиль
- квантиль

Кетле, Адольф

Кетле, Адольф родился в 1796 г.
Гент, Бельгия. В 1823 году министр
общественное образование попросило его контролировать
строительство обсерватории для

Machine Translated by Google

город Брюссель. С этой целью он был послан в Париж для изучения астрономии и подготовки планов строительства обсерватории. Во Франции Бувар Алексис поручил свое воспитание. Кетле, Адольф познакомился, среди прочих, с Пуассоном, Симеоном-Дени и де Лапласом, Пьером Симоном, которые убедили его в важности вероятностных расчетов. В течение следующих 4 лет он много путешествовал, посещая различные обсерватории и собирая инструменты. Он написал популярные работы по науке: Astronomie élémen.

taire, «Astronomie populaire et Positions de physique» и «Instructions populaires sur le calcul des probabilités» в 1828 году. В 1828 году он был назначен астрономом Брюссельской обсерватории и быстро поднялся до должности директора обсерватории, которую он занимал в течение следующих 25 лет.

В 1836 году он стал наставником принцев Эрнеста и Альберта Саксен-Кобург-Готских, что позволило ему продолжить интерес к расчету вероятностей.

Его уроки для мальчиков были опубликованы в 1846 году под названием Lettres à SAR le.

правящий герцог Саксен-Кобург-э-Готский; они обратились к теории вероятностей применительно к этическим и политическим наукам.
Кетле умер в 1874 году.

Основные произведения Кетле, Адольф:

1826 г. Элементарная астрономия. Мальтер, Париж.

1828 г. Физические положения. Брюссель.

1828 г. Популярные инструкции по исчислению вероятностей.

1846 Letters à SAR le Duc régnant de Saxe-Cobourg et Gotha sur la théorie des probabilités, appliquée aux sci ences мораль и политика, Brus sels.

Выборка квоты

Квотная выборка — это неслучайный метод выборки . Выборка выбирается таким образом, чтобы она воспроизводила изображение , максимально близкое к генеральной совокупности. Квотный метод основан на известном распределении населения по определенным признакам (полу, возрасту, социально-экономическому классу и т. д.).

ИСТОРИЯ

См. образцы.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим популяцию , состоящую из N особей. Известно, что распределение населения имеет характер

характеристики A, B, . . . , Z с соответствующими модальностями A1, A2,...,Aa; B1, B2,...,Bb; ...; Z1, Z2,...,Zz, искомая выборка получается путем умножения количества различных модальностей всех контрольных признаков.

скоростью зондирования q, что означает процент населения, которое нужно зондировать .

Выборка будет иметь объем n = N ·q и будет иметь следующее распределение:

$$q \cdot XA1 , q \cdot XA2 ,...,q \cdot XAa , q \cdot XB1 , q \cdot XB2 ,...,q \cdot XBb ,..., q \cdot XZ1 , q \cdot XZ2 ,...,q \cdot XZz ,$$

где X... представляет количество различных модальностей для всех контрольных признаков.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Контрольный признак, который будет использоваться в качестве основы для распределение выборки должно соответствовать следующие правила: (1) Он должен быть в тесной корреляции с изучаемыми переменными. (2) Это должно иметь известное распределение для всего Население.

Обычно используемыми контрольными признаками являются пол, возраст, социально-профессиональный класс, регион и др.

К преимуществам квотной выборки относятся следующий:

- В отличие от случайной выборки квотная выборка не требует наличия звучащая основа .
- Стоимость намного ниже, чем на рандом выборка.

Основные недостатки:

- Квотный метод не имеет достаточного теоретического обоснования. Он основан на следующей гипотезе: правильный распределение контрольных признаков обеспечивает репрезентативность распределения наблюдаемые черты.
- Метод квот не может рассчитать точность полученных оценок из образца. Поскольку следователи выбирают людей для опроса, невозможно узнать, с какой вероятностью любой данный индивидуум населения может быть частью выборки. Поэтому невозможно применить расчет вероятности того, что в случае случайного выборки, может ассоциировать с каждой оценкой меру ошибки , которая могла было совершено.

ПРИМЕРЫ

Предположим, мы хотим изучить рыночные возможности потребительского товара в городе X. В качестве контрольных переменных выбраны пол, возраст, и социально-профессиональный класс. Население распространение (в тысячах человек) следующий:

Пол	Возраст	Социально-профессиональный класс	
Мужчины 162	0–20 70	Начальник	19.25
Женщины 188	21–60 192,5	Независимая дент проф.	14,7
	61 87,5	Менеджеры 79,1	
		Рабочие 60,9	
		Унем занятый	176,05
Всего 350	Всего 350	Всего	350

Мы выбираем скорость звучания, равную $\frac{1}{300}$. выборке, поэтому имеем следующий размер и распространение:

$$n = N \cdot q = 350000 \cdot = 1167 . \quad \frac{1}{300}$$

Пол	Возраст	Социально-профессиональный класс	
Мужской	540 0–20 233	Начальник	64
Женщины 627	21–60 642	Независимая дент проф.	49
	61 292	менеджера 264	
		Рабочие 203	
		Унем занятый	587
Всего 1167	Всего 1167	Всего	1167

Затем выбор индивидуумов осуществляется простая случайная выборка в каждой категории.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Оценка
- Простая случайная выборка
- Выборка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Сукхатме, П.В.: Теория выборки обследований с приложениями. Издательство государственного университета Айовы, Эймс, Айова (1954 г.)

Machine Translated by Google

Случайный эксперимент

Эксперимент , исход которого нельзя предсказать заранее , называется случайным экспериментом.

Случайный эксперимент можно охарактеризовать следующим образом:

1. Можно описать множество всех возможных результатов (называемое выборочным пространством случайного эксперимента).
2. Невозможно с уверенностью предсказать результат.
3. Каждый возможный исход можно связать с вероятностью появления.

ОБЛАСТИ И ОГРАНИЧЕНИЯ Случайные явления, к которым применимы вероятности , можно встретить во многих различных областях, таких как экономические науки, социальные науки, физика, медицина, психология и т. д.

ПРИМЕРЫ

Вот несколько примеров случайных экспериментов, характеризуемых их действием и набором возможных результатов (их выборочным пространством).

Действие	Образец пространства
Подбрасывание монеты	= {Орел, решка}
Бросаем кубик	= {1, 2, 3, 4, 5, 6}
Вытягивание мяча из контейнера одним зеленый шар, один синий шар и один желтый шар	"=" {Синий, зеленый, желтый}
Количество осадков в определенный период	= набор неотрицательные действительные числа (бесконечное несчетное множество)

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Событие
- Вероятность
- Образец пространства

Случайное число

Случайное число называется реализацией равномерно распределенная случайная величина в интервале [0, 1].

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Если наложенные выше условия не проверяются строго, т. е. если распределение, откуда взят образец, не является точным,

Machine Translated by Google

Если равномерное распределение и последовательные розыгрыши не обязательно независимы, то мы говорим, что набор случайно полученных чисел образует серию псевдослучайных чисел.

ПРИМЕРЫ

Контейнер содержит 10 шаров от 0 до 9. Если наш Цель состоит в том, чтобы получить случайное число от 0 до 1 до m знаков после запятой, нам нужно нарисовать m шаров с заменой, стремящейся сформировать число из m цифр. Затем считаем эти розыгрыши так, как если бы было только одно число, разделив его на 10м.
В качестве примера предположим, что мы последовательно вытягиваем следующие шары: 8, 7, 5, 4 и 7. Соответствующее случайное число равно 0,87547.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Генерация случайных чисел
- Равномерное распределение

Генерация случайных чисел

См. генерацию случайных чисел.

Случайная переменная

Переменная , значение которой определяется результатом случайного эксперимента, называется случайной величиной.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Случайная величина обычно обозначается одной из последних букв алфавита (в верхнем регистре). Это функция с действительным значением, определенная на выборочном пространстве . Другими словами, реальная случайная величина X представляет собой отображение на :

X : .

Мы ссылаемся на:

- Дискретные случайные величины, когда набор всех возможных значений случайной величины либо конечен, либо бесконечен.

способный;

- Непрерывные случайные величины, когда множество всех возможных значений случайной величины бесконечно несчетно.

ПРИМЕРЫ

Примеры дискретной случайной величины

Если мы бросим два кубика одновременно, 36 возможных результатов будут следующими:

= {(1, 1), (1, 2),..., (1, 6),
(2, 1), (2, 2),..., (2, 6),..., (6, 1), (6, 2),..., (6, 6)} .

Случайная величина X = «сумма двух игральных костей» равна дискретная случайная величина. Она принимает свои значения в конечном множестве E = {2, 3, 4,..., 11, 12}, то есть сумма двух игральных костей может быть равна 2, 3, 4,... или 12. В этом Например, X — полностью положительная случайная величина, а операция выполняется из E.

Можно приписать вероятность различным значениям случайной величины X.
могу взять.

Тогда значение «сумма равна 2» получается только по результату (1, 1). С другой стороны, значение «сумма равна 8» получается по результатам (2, 6), (3, 5), (4, 4), (5, 3) и (6, 2).
Поскольку содержит 36 результатов с равной вероятностью, мы можем утверждать, что «шанс» или «вероятность» получения «суммы, равной 1 2», равен 36. Вероятность получения «суммы равно 8” $\frac{5}{36}$, и т. д.

Пример непрерывной случайной величины

Такие случайные величины характеризуют меня например, период времени

Machine Translated by Google

телефонного звонка, определенного в интервале]0, [, направление ветра, определяемое в интервале [0, 360[и т. д.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Функция плотности
- Вероятность
- Случайный эксперимент
- Образец пространства
- Ценить

Рандомизация

Рандомизация – это процедура, с помощью которой обработки случайным образом распределяются по экспериментальным единицам. Он используется, чтобы избежать субъективного лечебные аффекты. Он позволяет устранить систематическая ошибка, вызванная неконтролируемым факторы, которые сохраняются даже при повторных экспериментах. Рандомизацию можно использовать, чтобы избежать любое предубеждение, которое может исказить данные. Наконец, рандомизация необходима, если мы стремимся оценить ошибку эксперимента.

ИСТОРИЯ

В сентябре 1919 года Фишер Рональд Эйлмер начал свою карьеру на Ротамстедской экспериментальной станции, где проводились сельскохозяйственные исследования. происходило. Он начал с исторических данных, а затем принял во внимание данные сельскохозяйственных испытаний, для которых он разработал дисперсионный анализ. Эти испытания привели его к разработке экспериментальных конструкций. Он разработал высокоэффективные экспериментальные планы, используя принципы рандомизации, латинского квадрата, рандомизированного блока и факторные договоренности. Все эти методы можно найти в его работах. Статистические методы для научных работников (1925). Его статья 1926 г. о постановке сельскохозяйственных опытов стала книга План экспериментов (1935).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Полностью рандомизированный дизайн
- План экспериментов
- Эксперимент
- Экспериментальная установка
- Уход

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Статистические методы для Исследовательские работники. Оливер и Бойд, Эдин Бург (1925)

Фишер, Р.А.: Организация полевых экспериментов. Дж. Министерство сельского хозяйства. 33, 503–513 (1926)

Фишер, Р.А.: План экспериментов. Оливер и Бойд, Эдинбург (1935)

Фишер, Р.А.: Статистические методы, план эксперимента и научные выводы. Переиздание Статистических методов для научных работников «Дизайн Эксперименты», «Статистические методы и научное заключение». Беннетт, Дж. Х., Йейтс, Ф. (ред.) Oxford University Press, Оксфорд (1990)

Рандомизированный дизайн блоков

Рандомизированный блочный план — это экспериментальный план, в котором экспериментальные единицы находятся в группах, называемых блоками. Лечение случайным образом распределяются между экспериментальными единиц внутри каждого блока. Когда все процедуры появляются хотя бы один раз в каждом блоке, мы имеем полностью рандомизированный дизайн блоков. В противном случае имеем неполный рандомизированный блочная конструкция.

Такая конструкция используется для минимизации последствия систематической ошибки. Если экспериментатор сосредотачивается исключительно на различиях между методами лечения, эффекты,

Machine Translated by Google

к различиям между различными блоками следует устранить.

ИСТОРИЯ

См. экспериментальный план.

ПРИМЕРЫ

Фермер владеет пятью участками земли, на которых он хочет выращивать кукурузу. Он хочет бежать эксперимент, так как у него есть два вида кукурузы и два вида удобрений. Кроме того, он знает, что его участки весьма неоднородны в отношении солнечного света, и поэтому может возникнуть систематическая ошибка, если солнечный свет действительно облегчить выращивание кукурузы.

Фермер делит землю на пять блоков и случайным образом приписывает каждому блоку следующие четыре обработки:

A: кукуруза 1 с удобрением 1
B: кукуруза 1 с удобрением 2
C: кукуруза 2 с удобрением 1
D: кукуруза 2 с удобрением 2

Таким образом, он устраняет источник систематической ошибки из-за фактора солнца. Выполняя одну и ту же процедуру случайным образом внутри каждого блока, он не отдает предпочтения какому-то особому сорту кукурузы по сравнению с другими, так и особому виду удобрений по сравнению с другими. В следующей таблице показано случайное расположение четырех процедур в пяти блоках:

Блок 1	DCAB
Блок 2	ЦБАД
Блок 3	ABCD
Блок 4	ACBD
Блок 5	DCAB

Дисперсионный анализ позволяет сделать выводы после устранения блокирующих эффектов о том, существует ли значимое различие между обработками.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- План экспериментов

Диапазон

Размах является самой простой для вычисления мерой дисперсии .

Рассмотрим набор наблюдений относительно количественная переменная X. Диапазон определяется как разница между самым высоким наблюдаемым значением и самым низким наблюдаемым значением. ценить.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим набор наблюдений относительно количественная переменная X.

Если мы обозначим Xmax значение самого высокого наблюдения в наборе наблюдений и Xmin самое низкое значение, то диапазон определяется как:

$$\text{диапазон} = X_{max} - X_{min} .$$

Когда наблюдения сгруппированы в классы, диапазон равен разности между центрами двух крайних классов э. Пусть δ1 — центр первого класса, а δk — центр последнего класса. Диапазон равен:

$$\text{диапазон знак равно } \delta_k - \delta_1 .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Диапазон — это мера дисперсии , которая используется не очень часто, потому что она дает только указание на диапазон значений , но не дает никакой информации о распределении в пределах этого диапазона.

Тем не менее, он имеет большое преимущество в том, что его легко вычислить. Однако следует быть очень осторожным с его толкованием.

Machine Translated by Google

Диапазон — это мера дисперсии , которая учитывает только два крайних значения. Таким образом, этот метод может привести к неправильным интерпретациям, если количество детей в семи семьях равно 0, 0, 1, 1, 2, 3 и 6, диапазон равен 6, тогда как все семьи, кроме одной, имеют от 0 до 3 детей. В этом

В этом случае изменчивость относительно низка, тогда как диапазон велик.

Использование диапазона для измерения дисперсии (или изменчивости) представляет еще одно неудобство. Она может увеличиваться только при увеличении количества наблюдений . это дисадван

Это связано с тем, что в идеале мера изменчивости не должна зависеть от числа наблюдений.

Нужно проявлять крайнюю осторожность, когда интерпретация диапазона, потому что могут быть выбросы. Они оказывают прямое влияние на значение диапазона.

ПРИМЕРЫ

Вот примеры двух групп людей, посещающих вечерние курсы:

Группа 1:	33 34 34 37 40 40 40 43 44 45
Группа 2:	22 25 31 40 40 41 44 45 50 52

Среднее арифметическое $\bar{x}$ этих двух наборов наблюдения идентичны:

$$\bar{x} = \frac{390}{10} = 39.$$

Даже если среднее арифметическое возрастов этих двух групп одинаково, 39 лет, дисперсия вокруг среднего очень разная.

Диапазон для первой группы равен 12 (45 - 33), тогда как диапазон для второй группы равен 52 - 22 = 30 лет.

Обратите внимание, что диапазон дает только ориентировочную меру изменчивости значений ; это не указывает на форму распределения в пределах этого интервала.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Межквартильный диапазон
- Мера рассеивания

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Андерсон, Т.В., Слов, С.Л.: Введение в статистический анализ данных. Хоутон Миффлин, Бостон (1978)

Ранг матрицы

Ранг матрицы – это максимальное число бер линейно независимых строк или столбцов.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть A — матрица размера m × n . Мы определяем его ранг, обозначаемый rA, как максимальное число линейно независимых строк или столбцов.

В частности:

$$rA \quad \text{мин}(m, n) .$$

Если rA = min (m, n), то говорят, что матрица A имеет полный ранг.

Если ранг квадратной матрицы порядка равен n, то ее определитель отличен от нуля; другими словами, эта матрица обратима.

ПРИМЕРЫ

Пусть A будет матрицей 2 × 3:

$$A = \begin{matrix} 732 \\ 104 \end{matrix} .$$

Его ранг меньше или равен min(2, 3).

Количество линейно независимых строк или столбцов равно 2. Следовательно, rA = 2.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- определитель
- Матрица

p

Machine Translated by Google

Рао, Кальямпуди Радхакришна

Рао Радхакришна родился в 1920 г.
Хадагали, Карната, Индия. он получил свое
Кандидат наук. в 1948 году в Кембриджском университете.
С 1989 года он получил звание почетного доктора
многих университетов по всему миру.
мира, особенно Невшательского университета,
в Швейцарии.

Его вклад в развитие статистической теории и ее
приложений сравним с работами Фишера, Рональда
Эйлмера,
Нейман, Ежи и другие важные современные
статистики. Среди его вкладов
наиболее важными являются теорема Фишера-Рао,
теорема Рао–Блэквелла, теорема Крамера–
Неравенство Рао и U-критерий Рао. Он
соавтор 12 книг и более 250
научные статьи. Его книга «Статистика и
Правда: «Пусть шанс работает» касается
фундаментальной логики и статистических
приложений и был переведен на многие языки.
датчики

Избранные произведения и публикации Рао,
Кальямпуди Радхакришна:

1973 Линейный статистический вывод и его
Приложения, 2-е изд. Уайли, Нью
Йорк.

1989 (совместно с Shanbhag, DN) Дальнейшее
расширение Choquet-Deny и Deny
теоремы с приложениями в теории
характеристик. QJ Math., 40,
333–350.

1992 (совместно с Чжао, Л.К.) Линейное представление
R-оценок в линейных моделях.
Может. Журнал статистики, 220:359–368.

1997 Статистика и правда: шанс
на работу, 2-е изд. Всемирная научная,
Сингапур.

1999 (совместно с Toutenburg, H.) Linear Models.
Наименьшие квадраты и альтернативы,
2-е изд. Серия Спрингера по статистике.
Springer, БерлинHeidelbergНовинка
Йорк.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Теорема Гаусса – Маркова
- Обобщенный обратный

Регрессивный анализ

Регрессионный анализ — это метод, который
позволяет изучать и измерять отношение
между двумя или более переменными. Начиная с
данных, зарегистрированных в выборке,
регрессионный анализ направлен на определение оценки
математическая связь между двумя или более переменными
способен. Цель состоит в том, чтобы оценить стоимость
одна переменная как функция одной или нескольких других
э переменные. Оцениваемая переменная называется
зависимой переменной и обычно обозначается Y.
Напротив, переменные
которые объясняют изменения Y , называются
независимыми переменными, и они обозначаются
по X.
Когда Y зависит только от одного X, мы имеем
простой регрессионный анализ, но когда Y зависит
по более чем одной независимой переменной, мы
иметь множественный регрессионный анализ. Если
связь между зависимой и независимой переменными
линейна, то мы имеем линейный регрессионный
анализ.

ИСТОРИЯ

Пионер линейного регрессионного анализа,
Боскович, Роджер Джозеф, астроном
а также врач, был одним из первых
найти способ определения коэффициентов линии
регрессии. Для получения строки

Machine Translated by Google

проходя максимально близко ко всем наблюдениям, следует определить два параметра таким образом, чтобы соблюдались два условия Босковича, RJ:

1. Сумма отклонений должна равняться нуль.
2. Сумма отклонений по абсолютной величине должен быть минимальным.

Боскович Р.Дж. установил эти условия около 1755–1757 гг. Он использовал метод геометрическое разрешение, которое он применил, когда измерение длины пяти земных меридианы.

В 1789 году де Лаплас, Пьер Симон в своей работа Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pend ule, приняли два условия, налагаемые Боскович, Р. Дж., и установил алгебраический метод решения проблемы Босковича. алгоритм.

Два других математика, Гаусс, Карл Фридрих и Лежандр, Адриан Мари, кажется, открыли, не сообщая друг с другом методом наименьших квадратов . Кажется, Гаусс, CF, использовал его. с 1795 года. Но именно Лежандр А.М. опубликовал метод наименьших квадратов в 1805 г. в своей работе Nouvelles méthodes pour la определение орбит комет, где мы находим приложение Sur la méth ode des moandres carrés. Споры об истинном авторе открытия, вспыхнувшего между двумя мужчинами в 1809 году. Гаусс, К.Ф. опубликовал свой метод со ссылками на его предыдущие работы. Обратите внимание, что в том же период, американец по имени Адриан Роберт, который ничего не знал о работах Гаусса и Легандра, также работал с наименьшими затратами. квадраты.

XVIII век ознаменовался развитием методом наименьших квадратов. Согласно Стиглер, С. (1986), это развитие тесно связан с тремя проблемами с того периода. Первый касался математическое представление и определение движений Луны.

второй был связан с объяснением векового неравенства, наблюдаемого в движении Юпитер и Сатурн. Наконец, третье касалось определения формы земля.

По мнению того же автора, в произведения Гальтона, Фрэнсиса происхождение термин линейная регрессия. Его исследования были сосредоточены о наследственности: возникло понятие регрессии из анализа наследственности в семье члены.

В 1875 г. Гальтон, К. Ф. провел эксперимент с мелким горошком, распределенным по семи частям. группы по массе. Для каждой группы он взял десять семян и попросил семерых друзей культивировать оценить их. Таким образом, он вырастил 490 зародышей. Он обнаружил, что каждая группа семян, классифицированная по весу, следует нормальному распределению, а кривые, центрированные по разным весам, распределяются одинаково нормально; изменчивость разные группы были одинаковыми. Гальтон Ф. говорил о «реверсии» (в биологии «реверсия» означает возврат к примитивному типу). Это была линейная реверсия, Гальтон, Ф. был свидетелем, так как семь семенных групп были нормально распределены, по сравнению не с весом родителей, а до значения, близкого к общему населению иметь в виду.

Человеческая наследственность была предметом исследования и Галтона Ф., который пытался определить соотношение между длиной родительских семян и у их потомства. В 1889 г. он опубликовал свою работу под названием «Природное свойство». это. В этой области с 1885 г. он использовал термин регрессия, а в своих последних работах он остановился на понятии корреляции.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

См. регрессию LAD, множественную линейную регрессию и простую линейную регрессию.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Цель регрессионного анализа состоит не только в том, чтобы определить отношение между зависимой переменной и независимой переменной (переменными). Регрессионный анализ также стремится установить достоверность оценок и, следовательно, достоверность полученных прогнозов.

Регрессионный анализ позволяет, кроме того, проверить, являются ли результаты статистическими. значимым и является ли связь между переменными реальной или только кажущейся.
Регрессионный анализ имеет различные применения во всех областях науки. Фактически, с того момента, как мы определяем отношение между переменными, мы можем предсказывать будущие значения при условии, что условия остаются неизменными и всегда существует ошибка . Обычно мы проверяем достоверность регрессионного анализа с помощью остаточного анализа.

ПРИМЕРЫ

См. линейную регрессию (множественную или простую), регрессию LAD и гребень.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Эйзенхарт, К.: Боскович и комбинация наблюдений. В: Кендалл, М., Плакетт, Р.Л. (ред.) Исследования по истории статистики и вероятностей, том. II. Гриффин, Лондон (1977)

Гальтон, Ф.: Естественное наследование. Макмиллан, Лондон (1889 г.)

Гаусс, CF: Theoria Motus Corporum Coelestium. Верке, 7 (1809 г.)

Лаплас, PS de: Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pendule. Histoire de l'Académie royale des inscriptions et belles lettres, avec les Mémoires de littérature tirées des

реестры Cette Académie. Париж (1789 г.)

Лежандр, АМ: Новые методы определения орбит комет.

Курсье, Париж (1805 г.)

Плакетт, Р.Л.: Исследования по истории вероятностей и статистики. В: Кендалл М., Плакетт Р.Л. (ред.) Открытие метода наименьших квадратов. об. II. Гриффин, Лондон (1977)

Стиглер, С.: История статистики, измерение неопределенности до 1900 года. Белкнап, Лондон (1986).

Анализ остатков

Дисперсионный анализ

Коэффициент детерминации

Матрица шляпы

Наименьших квадратов

Точка рычага

Множественная линейная регрессия

Нормальные уравнения

Остаток

Простая линейная регрессия

Область отклонения

Область отклонения – это интервал, измеренный в выборочном распределении исследуемой статистики, который приводит к отклонению нулевой гипотезы H0 при проверке гипотезы . Область отклонения дополняет область приемлемости и связана с вероятностью α, называемой уровнем значимости теста или ошибкой I рода.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим проверку гипотезы о параметре θ. Мы используем статистику T , чтобы оценить

сопоставьте этот параметр.

Случай 1: односторонний тест (правый хвост)

Гипотеза следующая:

$$H0 : \theta = \theta0$$

$$H1 : \theta > \theta0 .$$

где θ0 – заданное значение параметра θ.

В одностороннем тесте (тест на правый хвост) область отбраковки соответствует интервалу, ограниченному справа от критического значения:

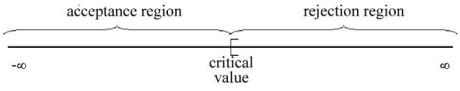
Область отклонения = [критическое значение; [, и
Область приема =] ; критическое значение[,

при условии, что критическое значение равно:

$$\mu T + z\alpha \cdot \sigma T ,$$

где μT — среднее значение выборочного распределения для статистики T, σT — стандартная ошибка для статистики T, α — уровень значимости теста.
Значение zα получают из статистических таблиц распределения T. Если значение T

вычисленная в выборке , лежит в области отклонения, то есть, если T больше или равно критическому значению, нулевая гипотеза H0 должна быть отвергнута в пользу альтернативной гипотезы H1. В противном случае нулевая гипотеза H0 не может быть отвергнута.



Cas 2: Односторонний тест (левый хвост)

Гипотеза следующая:

$$H0 : \theta = \theta0$$

$$H1 : \theta < \theta0 .$$

В одностороннем тесте (тест на правый хвост) область отторжения ограничена с левой стороны критическим значением:

область отклонения =] ; критическое значение] ,

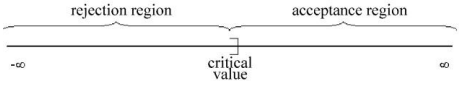
и область приема:

область приемлемости =] критическое значение; [,

при условии, что критическое значение равно:

$$\mu T \quad z\alpha \cdot \sigma T .$$

Если вычисленное по выборке значение T лежит в области отклонения, т. е. если T меньше или равно критическому значению, то нулевая гипотеза H0 должна быть отвергнута в пользу альтернативной гипотезы H1. В противном случае нулевая гипотеза H0 не может быть отвергнута.



Случай 3: двусторонний тест

Гипотеза следующая:

$$H0 : \theta = \theta0$$

$$H1 : \theta = \theta0$$

В двустороннем тесте область отторжения делится на два интервала:

область отбраковки =\ нижнее критическое значение;

верхнее критическое значение

и область приема:

область приемлемости = нижнее критическое значение;

верхнее критическое значение

где критические значения равны:

$$\frac{}{2} \cdot \sigma T \text{ и } \mu T + z\,\alpha \quad \mu T \quad z\,\alpha \quad \frac{}{2} \cdot \sigma T .$$

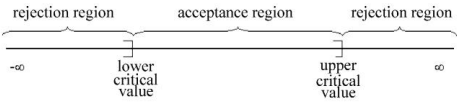
p

Machine Translated by Google

Если значение T , вычисленное в выборке , лежит в области отторжения, то есть, если T больше больше или равно верхнему критическому значению orifT меньше или равно нижнему критическому значению, нулевая гипотеза H0 должна быть отвергнута в пользу альтернативной гипотезы H1.

В противном случае нулевая гипотеза H0 не может быть отклоненный.

На следующем рисунке показаны области принятия и отклонения для упомянутых выше случаев.



ОБЛАСТИ И ОГРАНИЧЕНИЯ

В одностороннем тесте область отторжения уникален и обнаруживается:

- На правой стороне распределения выборки для одностороннего теста в правом хвосте.
- В левой части выборочного распределения для одностороннего теста в левом хвосте.

Напротив, область отторжения разделена на две части в двустороннем тесте, который находятся на крайних левой и правой сторонах выборочного распределения. Каждому из этим регионам связана вероятность $\frac{\alpha}{2}$, где α — уровень значимости теста.

ПРИМЕРЫ

Компания производит стальные тросы. На основе выборка из n = 100 единиц, он хочет проверить равен ли диаметр тросов 0,9 см.

Стандартное отклонение населения равно Известно, что он равен 0,05 см.

Проверка гипотезы в этом случае является двусторонней проверкой. Гипотеза следующая:

Нулевая гипотеза H0 : μ = 0,9

Альтернативная гипотеза H1 : μ = 0,9

Для уровня значимости , установленного на α = 5%, значение $z_{\frac{\alpha}{2}}$ нормальной таблицы равно 1,96. Тогда критические значения:

$$\mu_{\bar{x}} \pm z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{x}};$$

где $\mu_{\bar{x}}$ – среднее дискретное распределение для средних:

$$\mu_{\bar{x}} = \mu = 0{,}9$$

$\sigma_{\bar{x}}$ — стандартная ошибка среднего :

$$\sigma_{\bar{x}} = \frac{o}{\sqrt{n}} \text{ "=" } \frac{0{,}05}{\sqrt{100}} = 0{,}005$$

$$\mu_{\bar{x}} \pm z_{\frac{\alpha}{2}} \sigma_{\bar{x}} = 0{,}9 \pm 1{,}96 \cdot 0{,}005$$

$$= 0{,}9 \pm 0{,}0098 .$$

Интервал]0,8902 ; 0,9098[соответствует в область принятия нулевой гипотезы.

Область отторжения является комплементарной интервал и делится на две части:

$$]\quad ; 0{,}8902] \text{ и } [0{,}9098;\quad [.$$

Если полученное среднее значение выборки лежит в этом интервале, нулевая гипотеза H0 должна быть отвергается в пользу альтернативной гипотезы H1.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Область приема

Критическое значение

Проверка гипотезы

Уровень значимости

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бикель, Р. Д.: Выбор оптимальной области отклонения для множественного тестирования. Работающий документ, Управление биостатистики и биоинформации, Медицинский колледж Джорджии (2002)

Machine Translated by Google

Связь

Понятие отношения выражает взаимопонимание
которое существует между двумя случайными величинами.

Это одно из важнейших понятий в
статистика. Он проявляется в таких понятиях, как
классификация, корреляция, зависимость,
регрессионный анализ, временные ряды и т. д .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Классификация
- Коэффициент корреляции
- Зависимость
- Модель
- Регрессивный анализ

Временная последовательность

Относительный риск

Относительный риск определяется как отношение
между риском по отношению к отдельным лицам
кто подвергается воздействию фактора. Он измеряет
относительное действие определенной причины, т.
степень связи между заболеванием
и его причина.

ИСТОРИЯ

См. риск.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Относительный риск – это отношение вероятностей, которое
оценивается по:

$$\text{относительный риск} \frac{\text{риск разоблачения}}{\text{= риск неэкспонированных}}$$

$$\text{"="} \frac{\text{уровень заболеваемости для облученных}}{\text{уровень заболеваемости для неэкспонированных}} .$$

Обратите внимание, что относительный риск является мерой с
на единицу, так как она определяется как отношение двух
величины, измеряемые в одних и тех же единицах.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Относительный риск выражает риск
единиц, подвергающихся воздействию фактора, кратного
риск объектов, не подвергающихся воздействию этого фактора:

$$\text{риск единиц, выставленных} = \text{риск единиц, не подвергшихся воздействию} \cdot \text{относительный риск} .$$

Мы хотим изучить связь между риском
фактор и заболевание. Когда нет ассо
связи, относительный риск теоретически должен
быть равным 1. Если воздействие фактора риска
увеличивает число случаев заболевания,
тогда относительный риск становится больше 1.
Если, с другой стороны, воздействие фактора
снижается количество случаев заболевания,
тогда относительный риск находится между 0 и 1.

Легко вычислить доверительный интервал P = (1
α) % для относительного риска. Мы мошенники
Рассмотрим здесь следующий общий случай:
Доверительный интервал рассчитывается в
логарифмическая шкала, а затем преобразуется обратно
в арифметическую шкалу. Строчная пусть
см. в следующей таблице:

	Заболевание	Незаболевание	Итого
Незащищенный	a	n - a	н
Неэкспонированный	c	м - см	

Доверительный интервал P = (1 α) % для
тогда относительный риск, RR, определяется как:

$$\exp \ln (RR) \pm z\alpha \sqrt{\frac{1}{a} \frac{1}{н} + \frac{1}{c} \frac{1}{м}} ,$$

с «exp» показателем степени выражения
в скобках и zα значение в норме
стол. Предполагается, что a , n, c и m
достаточно большой. Если доверительный интервал для
относительный риск включает значение 1,0, тогда
делаем вывод, что фактор воздействия не
играть какую-либо статистически значимую роль в
вероятность наличия заболевания.

p

Machine Translated by Google

ПРИМЕРЫ

Мы используем здесь пример 1 из раздела о атрибутивный риск. Относительный риск груди рак, связанный с воздействием дыма, соответствует риску (или уровню заболеваемости) для каждый уровень воздействия, разделенный на риск (или уровень заболеваемости) для необлученных:

Группа	Уровень заболеваемости (100000 лет) ($\hat{p}_A$)	Родственник риск ($\hat{p}_A$ /570)
Неэкспонированный	57,0	1,0
Пассивные курильщики	126,2	2.2
Активные курильщики	138,1	2,4

Относительный риск составляет 2,2 для пассивных курильщиков и 2,4 для активных курильщиков. В сравнении к риску необлученных, риск для Активные курильщики и пассивные курильщики в настоящее время в 2,4 и 2,2 раза больше. Существует примерно скромная ассоциация, которая подчеркивает факт что рак молочной железы зависит также от факторов кроме курения. Относительный риск обеспечивает, таким образом, частичное описание отношения между курением и раком молочной железы. Само собой он не позволяет сделать каких-либо однозначных выводов и указывает на то, что курение, вероятно, ответственны за 50% случаев рака молочной железы среди населения.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Атрибутивный риск
- Предотвратимый риск
- Причина и следствие в эпидемиологии
- Уровень заболеваемости
- Шансы и отношение шансов
- Уровень распространенности
- Риск

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Корнфилд, Дж.: Метод оценки сравнительных показателей на основе клинических данных. Применить катионов к раку легких, молочной железы и шейка матки. Дж. Натл. Рак инст. 11, 1269–1275 гг. (1951)

Лилиенфельд, А.М., Лилиенфельд, Д.Э.: Основы эпидемиологии, 2-е изд. Кларен Дон, Оксфорд (1980)

MacMahon, B., Pugh, TF: Эпидемиология: Принципы и методы. Маленький Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Морабия, А .: L'Epidémiologie Clinique. Издания «Que sais-je?». Presses Universitaires de France, Париж (1996)

Повторная выборка

При использовании выборки для оценки определенного параметр , касающийся населения из которой взят образец, мы можем использовать один и тот же образец более одного раза, чтобы улучшить наши статистический анализ и выводы. Это происходит путем дискретизации внутри исходной выборки. В этом заключается идея методов передискретизации. Используя повторную выборку, мы можем вычислить смещение или стандартное отклонение расчетного параметр, или мы также можем сделать выводы из доверительного интервала построения и проверка гипотез для оцениваемого параметра. Среди методов, использующих повторную выборку, можно упомянуть, например, бутстрап, складной нож и перекрестную проверку. Их особый интерес представляют непараметрические методы, поскольку:

- Точное распределение данных не известный и
- Данные возникают в результате нелинейного проектирования, что очень сложно смоделировать.

С другой стороны, проблема, связанная с таким процедур возникает, когда образец не

Machine Translated by Google

представляют население, и мы рискуем сделать абсолютно ложные выводы.

ИСТОРИЯ

Метод складного ножа был введен в 1950-х годах как способ уменьшить погрешность оценки . В 1960-х и 1970-х годах повторная выборка широко использовалась как способ оценки дисперсии .

Метод начальной загрузки был изобретен Эфроном Б. в конце 1970-х гг. и с тех пор разрабатывался многочисленными исследователями в различных областях.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

См. бутстрап, складной нож.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Начальная загрузка
Метод складного ножа
Метод Монте-Карло
Рандомизация

ССЫЛКИ Edington,

ES: Рандомизационные тесты. Марсель Деккер, Нью-Йорк (1980)

Эфрон, Б.: Складной нож, бутстрап и другие планы повторной выборки, монография CBMS. № 38. СИАМ, Филадельфия (1982).

Фишер, Р.А.: План экспериментов. Оливер и Бойд, Эдинбург (1935)

Pitman, EJG: Критерии значимости, которые можно применять к образцам из любых популяций. Дополнение, Дж. Рой. Стат. соц. 4, 119–130 (1937)

Остаток

Остатки определяются как разница между наблюдаемыми значениями и оценкой сопряженные (подогнанные) значения регрессионной модели. Они представляют то, что не объясняется уравнением регрессии.

ИСТОРИЯ

См. ошибку.

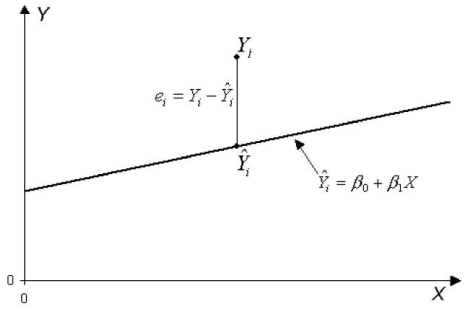
МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Остатки ei определяются как:

$$ei = Yi \quad Y\hat{}i , \qquad я = 1,...,n,$$

где Yi обозначает наблюдение , а Y^i — его подобранное значение, полученное путем подбора регрессионной модели.

Следующий график иллюстрирует остатки в простой линейной регрессии:



ПРИМЕРЫ

См. остаточный анализ.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Анализ остатков
Ошибка
Матрица шляпы
Точка рычага
Множественная линейная регрессия
Регрессивный анализ
Простая линейная регрессия

p

Machine Translated by Google

Райс, Стюарт Артур

Райс, Стюарт Артур (1889–1969) учился в Вашингтонском университете (получив степень бакалавра наук в 1912 году и степень магистра в 1915 году). Его научная работа была посвящена развитию методологии социальных наук.
В начале 1930-х Райс, ЮАР присоединилась к администрации Франклина Д. Рузвельта. Он стал заместителем директора переписи офис с 1933 до начала 1950-х годов. Его главный вклад в статистику состоит в модернизации использования статистики в правительстве. Он способствовал развитию COGSIS (Комитет по государственной статистике и информационным службам) и работал над адаптацией современных методов выборки, а также математической статистики в федеральных агентствах. В 1955 г. он создал Stuart A. Rice Associates (позднее — Surveys & Research Corporation), консультационное статистическое агентство для государственного и частного секторов.
Он вышел на пенсию в 1960-х годах.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Перепись
- Официальная статистика

Ридж-регрессия

Основная цель гребенчатого метода состоит в том, чтобы свести к минимуму среднеквадратичную ошибку оценок, то есть свести к минимуму систематическую ошибку и дисперсию. Рассматривая коллинеарность как проблему квазисингулярности W W, где W обозначает матрицу объясняющих переменных X , соответствующим образом преобразованную (отцентрированную и/или масштабированную), мы используем метод гребневой регрессии, который модифицирует матрицу W Win, чтобы удалить особенности.

ИСТОРИЯ

Гребневая регрессия была введена в 1962 г. Хёрлом А.Э., который заявил, что наличие корреляции между независимыми переменными может вызвать ошибки в оценке при применении метода наименьших квадратов. В качестве альтернативы он разработал гребенчатую регрессию, которая позволяет вычислять оценки со смещением, но с меньшей дисперсией, чем в наименьшей оценки квадратов. В то время как метод наименьших квадратов обеспечивает несмещенные оценки, гребень минимизирует среднеквадратичную ошибку оценок.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим модель множественной линейной регрессии :

$$Y = X\beta + \varepsilon \quad ,$$

где Y = (y1,...,yn) — (n × 1) вектор наблюдений, связанных с зависимая переменная (n наблюдений), β = β0,..., βp T — оцениваемый (p × 1) вектор параметров, ε = (ε1,..., εn) — (n × 1) вектор ошибки и

$$X = \begin{matrix} 1 & X11 & \dots & X1(p-1) \\ \vdots & \vdots & & \vdots \\ 1 & Xn1 & \dots & Xn(p-1) \end{matrix}$$

– матрица (n × p) с независимыми переменными.

Оценка β методом наименьших квадратов определяется по формуле :

$$\beta = \beta 0,..., \beta p-1 = XX^{-1}XY \quad .$$

Рассмотрим стандартизированную матрицу данных:

$$Вт = \begin{matrix} Xc_{11} & Xc_{12} & \cdots & Xs\ 1p-1 \\ Xc_{21} & Xc_{22} & \cdots & Xs\ 2p-1 \\ \vdots & \vdots & \dots & \vdots \\ Xc_{n1} & Xs_{n2} & \cdots & Xs\ np-1 \end{matrix} \quad .$$

Machine Translated by Google

Следовательно, модель становится:

$$Y = \mu + W\gamma + \varepsilon ,$$

где

$$\mu = (\gamma 0,\ldots, \gamma 0) = \beta s_{0,\ldots,} \beta_{0} \text{ и}$$

$$\gamma = \gamma 1,\ldots, \gamma p \quad 1 = \beta s_{1,\ldots,} \beta s_{p} \quad 1 .$$

Оценки методом наименьших квадратов задаются формулой $\gamma^{c}_{0} = \bar{Y}$ и по:

$$\gamma = \gamma 1,\ldots, \gamma p \quad 1$$

$$= \beta^{c}_{1,\ldots,} \beta^{c}_{p} \quad 1$$

$$= (Вт \cdot Вт) \quad 1 Вт \cdot Y .$$

Эти оценки являются несмещенными, то есть:

$$E \beta j = \beta j , j = 0,\ldots,p \quad 1 \; E \gamma j = \gamma j , j ,$$

$$= 0,\ldots,p \quad 1 .$$

Среди всех возможных линейных несмещенных оценок это оценки с минимальной дисперсией . Мы знаем это

$$Var \beta = \sigma 2 XX^{1} ,$$

$$Var(\gamma) = \sigma 2 W W \quad 1 ,$$

где σ2 — дисперсия ошибки. Обратите внимание, что WW равно (n 1) умножению на корреляционную матрицу для независимых переменных. Следовательно, в наиболее благоприятном случае коллинеарность не возникает, так как WW кратно I. Чтобы матрица WW была максимально близкой насколько это возможно на благоприятный случай, заменим WW на WW + kI, где k — положительная скалярная величина.

Гребневая регрессия заменяет оценки методом наименьших квадратов

$$\gamma = W W \quad 1 WY$$

по оценкам хребта

$$\gamma R = WW + kI^{1} ВГ$$

$$= \gamma R1 ,\ldots, \gamma R(p \quad 1) .$$

Всегда существует значение k , для которого общая среднеквадратическая ошибка оценок хребта меньше, чем общая среднеквадратическая ошибка оценок методом наименьших квадратов . г. (p 1)σ2 Hoerl et al. предположил, $\frac{B\ 1975}{\gamma\ \gamma}$ что k = дает более низкую общую среднеквадратичную ошибку, где (p 1) — число объясняющих переменных. Чтобы вычислить k, мы должны оценить σ и γ через s и использование гребневой регрессии с k = $\frac{\gamma . Затем}{\gamma\ \gamma}$ γ переоцен . Однако обратите внимание, что использованное k является случайной величиной, поскольку оно зависит от данных.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Свойства оценок хребта

Оценки хребта смещены, так как ожидаемые значения E γRj оценок γRj находятся в вектор :

$$E \gamma Rj = WW + kI^{1} W W\gamma = \gamma ,$$

$$\kappa > 0 .$$

Отметим, что значение k = 0 соответствует оценкам методом наименьших квадратов. Оценки горного хребта немного предвзяты, чтобы уменьшить изменчивость. оценка оценок. У нас есть:

$$Var(\gamma R) = \sigma 2 VR ,$$

с:

$$VR = WW + kI^{1} BBBB + кИ^{1} .$$

Заметим, что диагональные элементы матрицы VR меньше диагональных элементов, соответствующих W W 1 , а это означает, что ребро оценки имеют более низкую дисперсию, чем наименьший оценки квадратов.

Таким образом, оценки хребта менее изменчивы, чем оценки методом наименьших квадратов. Приемлема ли их предвзятость? Для измерения компромисса между

p

Machine Translated by Google

смещения и дисперсии оценки, мы обычно вычисляем ее среднеквадратичную ошибку, обозначаемую поMSE. Для оценки θ θ это определяется как:

$$\text{СКО } \theta = E\,\theta - \theta \qquad ^2$$

$$= \text{Var } \theta + E\,\theta - \theta \qquad ^2 .$$

Эквивалентно, мы определяем общее среднее значение квадрат ошибки оцениваемого вектора на сумма среднеквадратических ошибок его компоненты. У нас есть:

$$\text{TMSE } (\gamma R) = \sum_{j=1}^{p-1} \text{СКО } \gamma Rj$$

$$= \sum_{j=1}^{p-1} E\,\gamma Rj \quad yj \qquad ^2$$

$$= \sum_{j=1}^{p-1} \text{Вар } \gamma Rj$$

$$+ E\,\gamma Rj \quad yj \qquad ^2$$

$$= (p \quad 1)\,\sigma 2\text{Trace } (VR)$$

$$+ \sum_{j=1}^{p-1} E\,\gamma Rj \quad yj \qquad ^2 .$$

Этот метод называется адаптивной гребневой регрессией. Свойства гребневых оценок справедливы только тогда, когда k является константой, которая не случай адаптивной гребневой регрессии. Тем не менее, свойства гребневых оценок предполагают, что адаптивные гребневые оценки являются хорошими оценками. На самом деле, статистические исследования показали, что для многих наборов данных адаптивная оценка гребня лучше среднеквадратичная ошибка, чем оценка методом наименьших квадратов.

ПРИМЕРЫ

В следующем примере 13 образцов застыл цемент. Для каждого, за проценты четырех химических ингредиентов были измерены и представлены в таблице ниже. также было измерено количество выделяемого тепла.

Цель состояла в том, чтобы определить, как величины xi1, xi2, xi3 и xi4 влияют на yi, количество тепла излучаемый.

Таблица: Тепловыделение

Образцы	Ингредиент				Нагревать
	1234				
я	Си1	Си2	Си3	Си4	И
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

где Yi — тепло, выделяемое образцом i,
Xi1 – количество ингредиента 1 в образце i,
Xi2 – количество ингредиента 2 в образце i,
Xi3 – количество ингредиента 3 в образце i,
Xi4 – количество ингредиента 4 в образце i.

Сначала мы проводим множественную линейную регрессию. Модель линейной регрессии:

$$yi = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + \beta 3Xi3$$

$$+ \beta 4Xi4 + \varepsilon i .$$

Получаем следующие результаты:

Machine Translated by Google

Переменная	Коэффициент	std. разв.	TK
Постоянный	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	0,1441	0,7091	-0,20

Расчетный вектор, полученный по стандартизированным данным :

$$y = (9,12; 7,94; 0,65; -2,41) ,$$

со стандартным отклонением 4,381, 11,26, 4,834 и 11,87 соответственно. оценка для s от σ s = 2,446. Мы следуем по хребту анализ с использованием:

$$k = \frac{(p \quad 1)\,\sigma 2}{y\,y}$$

$$"=" \quad \frac{4 \cdot (2,446)^{2}}{(9.12)\,2+(7.94)\,2+(0.65)\,2+(2.41)\,2}$$

$$= 0,157 .$$

Это адаптивная гребневая регрессия, поскольку k зависит от данных. Однако мы используем термин гребневая регрессия. Мы получаем:

$$YR = (7,64; 4,67; -0,91; -5,84) ,$$

со стандартным отклонением 1,189, 1,377, 1,169 и 1,391 соответственно. У нас есть расчетные значения:

$$Yi = 95,4 + 7,64Xs_{i1} + 4,67X_{c\,i2} \quad 0,91X_{c\,i3} \quad 5,84X_{c\,я4} .$$

Расчетное значение константы не отличаются от оценки методом наименьших квадратов.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коллинеарность
- Среднеквадратическая ошибка
- Моделирование
- Стандартизированные данные

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Хёрл А.Э.: Применение Ридж-анализа к Проблемы регрессии. хим. англ. прог. 58 (3), 54–59 (1962)

Хорл А.Е., Кеннард Р.В.: Регрессия хребта: смещенная оценка для неортогональных задач. Технометрика 12, 55–67 (1970)

Хорл А.Э., Кеннард Р.В., Болдуин К.Ф.: Гребневая регрессия: некоторые симуляции. коммун. Стат., вып. 4, стр. 105–123. (1975)

Риск

В эпидемиологии риск соответствует вероятность появления заболевания в популяции в течение заданного интервала времени. Риск одинаково определена для любой совокупности или однородной подгруппы лиц. Однородность используется здесь в смысле что люди подвергаются воздействию определенного фактора, предположительно вызывающего заболевание, изучал. Понятие риска тесно связано с понятием заболеваемости (заболеваемость соответствует частоте заболевания, риск к его вероятности).

ИСТОРИЯ

В 1951 году Джером Корнфилд ясно заявил впервые появилось понятие риска. Этот результат последующего исследования, в котором риск факторы сравнивались для двух групп, одна состоящая из больных людей и других здоровых лиц. С тех пор риск широко используется в эпидемиологических исследованиях в качестве мера связи между заболеванием и сопутствующие факторы риска.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Риск – это вероятность заболевания появление в определенный период времени в группе лиц, подвергающихся воздействию фактора риска в изучать. Риск оцениваем соотношением:

$$\text{риск} = \frac{\text{инциденты (в экспонированном населении)}}{\text{население в группе риска}}$$

в течение заданного периода времени. Номинатор соответствует количеству новых случаев заболевания в течение установленного периода времени в изучаемого населения, а знаменатель соответствует числу лиц в риска в тот же период времени.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Понятие риска интуитивно представляет собой вероятностную меру, определяемую для данного населения и определенный период времени. Страховые компании используют его для прогнозирования частота событий, таких как смерть, болезнь, и т. д. в определенной популяции и в соответствии с такими факторами, как пол, возраст, вес и т. д. Риск можно интерпретировать как вероятность для человека, чтобы заразиться болезнью, учитывая, что он подвергся воздействию определенного фактор риска. Мы приписываем личности среднее значение риска для физических лиц своей группы. Страховые компании используют риск для каждой группы населения, чтобы определить вклад каждого застрахованного лица. Врачи использовать риск для оценки вероятности того, что у их пациентов развивается определенное заболевание, имеет осложнение или выздоравливает от болезни простота.

Важно отметить, что понятие риска всегда определяется для определенного периода времени: риск через 2, 3 или 10 лет.
Минимальный риск 0%, максимальный риск 100%.

ПРИМЕРЫ

Предположим, что 1 января 1992 г. группа Отобрано 1000 60-летних мужчин и зарегистрировано 10 случаев сердечного приступа во время исследования. год. Риск сердечного приступа для этой группы населения 60-летних мужчин равняется 10/1000 = 1% в году.
Мы рассматриваем здесь случай рака молочной железы в течение 2 лет в популяции 100 000 женщин, распределенных по трем группам в зависимости от степени воздействия фактор «дым». Следующее непредвиденное обстоятельство таблица иллюстрирует этот пример:

Группа	Количество случаи (инциденты)	Количество лица
Неэкспонированный	40	35100
Пассивные курильщики	140	55500
Активные курильщики	164	59400
Общий	344	150000

Риск для каждой группы рассчитывается как отношение числа случаев к делению по общему количеству особей в каждом группа. Например, риск пассивного курильщики 140/55500 = 0,002523 = 252,3/100000 за 2 года. Имеем следующую таблицу:

Группа	Количество случаи (инциденты)	Число инди видуалы	Риск в два годы (100000)
Нонекс поставленный	40	35100	114,0
Пассивный курильщики	140	55500	252,3
Активный курильщики	164	59400	276,1
Общий	344	150000	229,3

Понятие риска подразумевает понятие прогноза: среди женщин данного населения

Machine Translated by Google

ции, какая часть из них разовьется рак молочной железы в ближайшие 2 года? Из таблице выше отметим, что 229,3/100000 может заболеть этим заболеванием в ближайшие 2 года.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Атрибутивный риск
- Предотвратимый риск
- Причина и следствие в эпидемиологии
- Уровень заболеваемости
- Шансы и отношение шансов
- Уровень распространенности
- Относительный риск

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Корнфилд, Дж.: Метод оценки сравнительных показателей на основе клинических данных. Применение при раке легкого, молочной железы и шейка матки. Дж. Натл. Рак инст. 11, 1269–1275 гг. (1951)

Лилиенфельд, А.М., Лилиенфельд, Д.Э.: Основы эпидемиологии, 2-е изд. Кларен Дон, Оксфорд (1980)

MacMahon, B., Pugh, TF: Эпидемиология: Принципы и методы. Маленький Браун, Бостон, Массачусетс (1970)

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.)

Морабия, А .: L'Epidémiologie Clinique. Издания «Que sais-je?». Presses Universitaires de France, Париж (1996)

Надежная оценка

Оценка называется робастной, если оценки используемые маторы не чувствительны к выбросы и небольшие отклонения от идеализированных (модельных) предположений. Эти методы оценки включают Максимальное правдоподобие типовая оценка, известная как M-оценки, линейная комбинация порядковых статистик, известная как L-оценки и методы, основанные на статистических рангах, R-оценки. Выбросы, вероятно, являются ошибками , возникающими в результате плохого чтения, плохая запись или любая другая причина, связанная с в экспериментальную среду. Среднее арифметическое и другие оценки наименьшего квадраты очень чувствительны к таким выбросам. Поэтому в этих случаях предпочтительнее использовать оценщики, которые являются более надежными.

ИСТОРИЯ

Термин робастный, а также поле для с которым он связан, был введен в 1953 г. Box, GEP Термин был признан области статистики только в середине 1960-х гг. Тем не менее важно отметить, что это концепция не нова; в конце 19 века, несколько ученых уже имели четкое представление о это понятие. По сути, первая математическая работа над надежной оценкой, по-видимому, датируется назад к 1818 году с Лапласом, П.С. в его работе Le deuxième supplement à la théorie ana lytique des probabilités. Распределение медиану можно найти в этой работе . Это 1964 г., статья Хубера П.Дж. что эта область стала отдельной областью статистика, известная как робастная статистика.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим следующую модель:

$$Y = X \cdot \beta + \varepsilon \quad ,$$

где Y — (n × 1) вектор наблюдений относительно зависимой переменной (n наблюдений), β - (p × 1) вектор оцениваемые параметры, а ε – (n × 1) вектор ошибок. Существует несколько методов робастной оценки. , которые позволяют оценить вектор параметров β .

Machine Translated by Google

Вот некоторые из методов:

- Оценка L1 :

$$\text{Свернуть} \sum_{i=1}^{n} |\varepsilon_i| .$$

- В более общем виде концепция M оценок, основанная на идее замены квадрата ошибок (метод наименьших квадратов), определяется функцией δ(i), где δ — симметричная функция, имеющая единственный минимум в нуле :

$$\text{Минимизировать} \sum_{i=1}^{n} \delta(\varepsilon_i) .$$

В 1964 г. Huber, PJ предложил следующую функцию:

$$\delta(x) = \begin{cases} \frac{x^2}{2}, & \text{если } |x| \le k \\ k \cdot |x| - \frac{k^2}{2}, & \text{если } |x| > k \end{cases},$$

где k — положительная константа.

Заметим, что при δ(x) = |x| получаются L1 -оценки.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Оценка
- Оценщик
- оценка L1
- Максимальная вероятность
- Выброс

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Вставка, GEP: ненормальность и тесты на дисперсию. Биометрика 40, 318–335 (1953).

Хубер, П.Дж.: Надежная оценка параметра местоположения. Анна. Мат. Стат. 35, 73–101 (1964)

Хубер, П.Дж.: Надежная статистика. Уайли, Нью-Йорк (1981)

Хубер, П.Дж.: Надежность: где мы сейчас? Студент, вып. 1, нет. 2, стр. 75–86 (1995).

Лаплас, П.С.: Deuxième supplément à la theorie Analytique des Probabilités. Курсье, Париж (1818 г.). стр. 531–580 dans les Œuvres complètes de Laplace, vol. 7. Готье-Виллар, Париж, 1886 г.

Machine Translated by Google

C

Образец

Выборка – это подмножество генеральной совокупности , на которой проводятся статистические исследования с целью сделать выводы относительно населения .

ПРИМЕРЫ

Домохозяйка готовит суп. Чтобы определить, достаточно ли соли в супе, она пробует ложку. В зависимости от того, что она наблюдала на образце, представленном здесь ложкой, она может решить, нужно ли добавлять соль в суп, символизирующий изучаемую популяцию .

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Неравенство Чебышева
- Оценка
- Население
- Размер образца
- Выборка

Размер образца

Размер выборки – это количество индивидуумов или элементов, принадлежащих выборке .

ИСТОРИЯ

Нордин, Дж. А. (1944) рассматривает проблему оценки размера выборки в рамках теории решений через относительный пример потенциальных продаж на рынке, где продавец хочет присутствовать. Cornfield, J. (1951) очень точно иллюстрирует оценку размера выборки для формирования классов с известными пропорциями. Ситтиг Дж. (1951) обсуждает выбор размера выборки для решения типичной экономической задачи, принимая во внимание контроль и стоимость дефектных изделий в принятой партии и правильных изделий в бракованной партии.
Многочисленные исследования были также проведены Коксом, Д.Р. (1952), который был вдохновлен работами Штейна, К. (1945), для идеального определения размера выборки.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X1,...,Xn — случайная выборка , взятая без замены совокупности из N наблюдений , имеющих неизвестное среднее значение μ.
Мы предполагаем, что выборочное распределение статистики следует нормальному распределению.

Неравенство Чебышева зависит от числа n наблюдений и позволяет

оценить среднее значение μ по эмпирическому среднему полученному значению $\overline{Xn}$.

При достаточно большом N размер выборки n можно определить с помощью неравенства Чебышева, так что вероятность попадания $\overline{Xn}$ в интервал [μ ε , μ + ε] (ε > 0, фиксированный при требуемом точность) может быть сколь угодно велика. Если мы позвоним

Machine Translated by Google

α уровень значимости, получаем:

$$P(|\bar{x} - \mu| \varepsilon) = \alpha \text{ или}$$

$$P(|\bar{x} - \mu| < \varepsilon) = 1 \alpha ,$$

который мы также можем записать как:

$$P \quad z\alpha/2 \quad \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad z\alpha/2 = 1 \quad \alpha ,$$

где $\sigma_{\bar{x}}$, стандартная ошибка среднего, оценивается через $\frac{\sigma}{\sqrt{n}}$, где σ – стандарт σ n отклонение заселенности и zα/2 стандартное критическое значение, связанное с конфи уровень плотности 1 α.

Мы строим доверительный интервал с с помощью таблицы нормалей для доверительной вероятности 1 α:

$$z\alpha/2 \cdot \frac{\sigma}{\sqrt{n}} \quad \bar{x} \quad \mu \quad z\alpha/2 \cdot \frac{\sigma}{\sqrt{n}} .$$

Поскольку нормальное распределение симметрично, нас интересует только неравенство на верно. Это указывает на то, что $z\alpha/2 \cdot \frac{\sigma}{\sqrt{n}}$ самый большой значение, которое может принимать $\bar{x}$ μ. Этот предел также задается желаемой точностью. Откуда:

$$\varepsilon = z\alpha/2 \cdot \frac{\sigma}{\sqrt{n}} ,$$

то есть:

$$n = z\alpha/2 \cdot \frac{\sigma}{\varepsilon}^2 .$$

Если желаемая точность может быть выражена в процент относительно среднего μ, или k%, мы тогда есть:

$$\delta = \kappa \cdot \mu$$

$$\text{и } n = \frac{z\alpha}{\kappa} \cdot \frac{\sigma}{\mu}^2 ,$$

где $\frac{\sigma}{\mu}$ называется коэффициентом вариации ция. Он измеряет относительную дисперсию изучаемая переменная. В случаях, когда сам распределение нормального распределения не является нормальным, эти формулы приблизительно справедливы когда размер выборки n намного больше, чем (n 30).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Определение размера выборки важный шаг для организации статистического исследования. Это подразумевает не только сложные расчеты, но и учет с учетом различных требований само расследование. Таким образом, желаемая точность результаты расследования подлежат рассмотрению. Как правило, чем больше размер выборки, тем больше точность. Стоимость расследования является еще одним важным ограничением. меньше бюджет, тем больше ограничений размер образца. Более того, наличие другие ресурсы, такие как существование данных исходя из переписи или системы заявителей или продолжительность расследования, как а также территория, подлежащая покрытию, может иметь драматическое влияние на структуру образца.

Теория переписи придает большое значение образцы «оптимальной» размерности, особенно при оптимизации (максимизации или минимизации) целевой функции, принципы которой сформулированы математически. Более того, важно подчеркнуть, что в общем случае оптимальный размер выборки не зависят (или мало зависят) от размера исходного населения.

В области планирования эксперимента также возникает проблема оптимального размера выборки. В этом случае следует найти минимальное количество наблюдений, которые необходимо принять во внимание обнаружить существенную разницу (если есть один) между различными видами лечения, такими столько же лекарств или производственных процессов.

ПРИМЕРЫ

На фабрике работает несколько тысяч рабочих. Основываясь на предыдущем эксперименте, исследователь знает, что недельная заработная плата нормально распределены со стандартным отклонением 40 CHF. Исследователь хочет оценить

Machine Translated by Google

определите среднюю недельную зарплату с точностью до 10 швейцарских франков и уровня достоверности 99%.

Каким должен быть размер выборки?

Для уровня достоверности 99% (α = 0,01 = 1 - 0,99) значение z0,01 , которое мы находим в обычной таблице , равно 2,575.

Непосредственное применение формулы дает нас:

$$n \quad z\alpha \cdot \frac{o}{—} \quad 2 \quad ,$$

$$n \quad 2{,}575 \cdot \frac{}{40}\ 102 \ ,$$

$$n \quad 106{,}09 \ .$$

Мы будем иметь желаемую точность, если размер выборки больше или равен 107 работ.

эрс.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Население

Образец

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Cochran, WG: Sampling Techniques, 2-е изд. Уайли, Нью-Йорк (1963)

Корнфилд, Дж.: Методы отбора проб человеческого населения. Определение размера выборки. амер. Дж. Публ. Здоровье 41, 654–661 (1951)

Кокс, Д.Р.: Оценка методом двойной выборки. Биометрика 39, 217–227 (1952).

Харнетт, Д.Л., Мерфи, Дж.Л.: Вводный статистический анализ. Аддисон-Уэсли, Рединг, Массачусетс (1975)

Нордин, Дж. А.: Определение размера выборки. Варенье. Стат. доц. 39, 497–506 (1944)

Sittig, J.: Экономичный выбор системы отбора проб при приемочном отборе. Бык. Междунар. Стат. Инст. 33 (B), 51–84 (1952)

Штейн, К.: Тест с двумя выборками для линейной гипотезы, мощность которой не зависит от дисперсия. Анна. Мат. Стат. 16, 243–258 (1945).

Образец пространства

Выборочное пространство случайного эксперимента — это набор всех возможных результатов эксперимента , обычно обозначаемый .

ИСТОРИЯ

См. вероятность.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Выборочное пространство может быть: • конечным (исходы счетны); • Бесконечная счетность (каждому возможному исходу можно поставить в соответствие натуральное число так, что каждый возможный исход будет иметь разное число, однако число возможных исходов бесконечно);

• Бесконечные несчетные (невозможно перечислить все возможные исходы).

Выборочное пространство случайного эксперимента также называют надежным событием. Вероятность достоверного события равна 1 :

$$P() = 1 \quad .$$

ПРИМЕРЫ

Конечное пространство выборки:

- Подбрасывание монеты

$= \{\text{Орел, решка}\}$

- Бросание кубика

$= \{1, 2, 3, 4, 5, 6\}$

Бесконечное счетное пространство выборки:

– Подбрасывание монеты столько раз, сколько необходимо для получения «орла» (H = орел и T = решка)

$$= \{H, TH, TTH, TTTH, TTTTH, TTTTT, TTTTTT,...\} .$$

C

Machine Translated by Google

Бесконечное несчетное пространство выборки:

– Местоположение координат удара
пули на панели размером 10×10 см. В
В этом случае выборочное пространство состоит
бесконечного несчетного множества координат
динаты, потому что координаты средние
уверен в шкале действительных чисел (это
невозможно дать исчерпывающий перечень
возможные результаты).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Событие

Вероятность

Случайный эксперимент

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бейли, Д.Э.: Вероятность и статистика: модели для
исследований. Уайли, Нью-Йорк (1971)

Выборка

С общей точки зрения термин выборка означает отбор части совокупности , называемой выборкой, изучение определенных
характеристики x данного образца, а также тот факт, что делать выводы относительно населения. Выборка также представляет собой набор операций, необходимых для получения выборки из данной совокупности.

ИСТОРИЯ

Понятие репрезентативности выборки очень недавний.
Хотя первые попытки экстраполяции появились еще в XIII в.
во Франции, где количество домов было используется для оценки населения , переписи оставался более популярным, чем выборка, пока 19 век.
Принцип отбора проб с или без замена появилась впервые в произведение «De ratiociniis in Aleae Ludo», опубликовано в 1657 г. голландским ученым Гюйгенсом Христианом (1629–1695).
Это было в 1895 году, в Берне, во время конгресса Международного института статистики, что Киаер, Андерс Николаи представил мемуар под названием «Наблюдения и опыт, касающийся представителей представителей», Киаер, как директор Центрального бюро Статистика Королевства Норвегия на время сравнил в своем изложении структуру выборки со структурой населения , полученной переписью. Поэтому он был определение термина апостериорный контроль и для впервые употребил термин «представитель». Kiaer, AN столкнулся с противодействием почти все участники конгресса, которые были убежден, что полное перечисление было единственный способ продолжить.
Тем не менее, в 1897 г. Kiaer, AN подтвердил понятие репрезентативности и получила поддержку Конференции Скандинавские статистики. Ему дали возможность отстаивать свою точку зрения на нескольких конференциях: Санкт-Петербург (1897 г.), Стокгольм (1899 г.), Будапешт (1901 г.), и Берлин (1903 г.).
Новый шаг был сделан с работами Марш, Люсьен о роли случайности во время выборки; он первым разработал идея вероятностной выборки, также называемая случайная выборка.
К работе подключились и другие статистики. проблема. Фон Борткевич Л., профессор из Берлина, предложил вычисление вероятностей для проверки отклонения между распределениями выборки и общая популяция по ключевым переменным. Боули, А.Л. (1906) разработал случайный Процедура выборки называется стратификацией. Согласно Desrosières, A. (1988), он был также интересует понятие уверенности

Machine Translated by Google

интервал, первые расчеты которого он представил в 1906 г. в Королевскую статистическую службу. Общество.

1925 год стал новым этапом в истории отбора проб. Это действительно год Конгресс Международного института статистики в Риме, на котором была представлена случайная выборка (choicedonebychance) и целенаправленная выборка (разумный выбор). Отныне проблема никогда не возникнет еще раз о выборе между выборочный и полный подсчет, но это возникли в связи с различными выборками методы.

Из различных применений выборки наиболее часто используемый в Соединенных Штатах был тот, который сделан различным общественным мнением институтов и особенно предвыборных расследований. Третий день ноября 1936 года, день результаты президентских выборов были обнародовано, знаменует собой один из самых важных дней в этом отношении. В этом опросе Ридерс Дайджест предсказал победу Лэндона, в то время как опросщики, не консультируясь друг с другом, Кроссли, Роупер и Гэллап объявили победа его противника Рузвельта. Ру Севельт был избран. Ошибка «Ридерз Дайджест» состояла в том, что он основывал свой прогноз на результаты выборов по телефонному опросу, но в то время только богатые люди владели телефон, поэтому выборка была необъективной! 1938 год знаменует собой рождение во Франции IFOP (Французский институт мнений Publice, Французский институт общественного Мнение), созданный Штетцелем, Жаном. Интересно вернуться немного в прошлое подчеркнуть ту роль, которую российские статистики сыграли в эволюции выборки методы. Действительно, уже в XIX веке они были известны и использовались в своих целях. страна. По словам Тасси, доктора философии (1988), Чупров А.И. (1842–1908) был одним из предшественники современного семплирования, и из 1910 г. его сын Чупров А.А. выборки, ссылались на кластерную выборку, стратифицированную выборку с заменой и без замены, а также под влиянием Марков А. поощрял использование вероятностей в статистике. Даже если русский школа была известна своими выборочными работами очень поздно, статистики этой школы видимо знал об этих приемах давно раньше, чем в Западной Европе. Отметим, что в 1924 г. Ковалевский А.Г. уже усиленно работал над опросами стратификация и оптимальное распределение по stratum, в то время как этот результат был переоткрыт лишь 10 лет спустя Нейманом, Дж. Наконец, эта история была бы неполной. не говоря уже о вкладе Нейман, Ежи на пробу. Обдуманный один из основоположников статистической теории ри опросов, он выступал за случайные выборка на целенаправленной выборке, и он работал над выборкой с денежными переводами или без таковой, стратифицированной выборкой и кластерной выборка. Он уже использовал оценки с помощью частного, разности и регрессии, методы, которые будут разработаны позже. позже Horvitz, DG, Cochran, WG, Хансен, М. М., Гурвиц, В. Н., и Мэд Оу, В. Г.

C

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Существует несколько методов выборки. Они можно разделить на случайные методы и неслучайные методы.

Наиболее широко используемая неслучайная процедура является квотная выборка.

Случайные методы (или вероятностная выборка, иногда также называемая статистической выборкой) используют вероятностный механизм. для формирования выборки , взятой из популяции.

Machine Translated by Google

Переменные , наблюдаемые в выборке , являются случайными величинами. Из этих переменных можно вычислить соответствующие значения генеральной совокупности и ошибки, связанной с выборкой.

К основным случайным методам относятся:

• Простая случайная выборка • Стратифицированная выборка • Систематическая выборка • Кластерная выборка

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Цель выборки состоит в том, чтобы предоставить достаточно информации, чтобы сделать выводы о чаре. характеристики популяции . Поэтому необходимо выбрать образец , который воспроизведет эти характеристики как можно точнее.

Предпринимаются попытки максимально уменьшить потенциальную ошибку , связанную с выборкой. Но даже если эта ошибка выборки неизбежна, она допускается, потому что она в значительной степени компенсируется преимуществами метода выборки. Это:

• Снижение затрат
С учетом всех затрат, которые могут быть понесены при сборе данных их, очевидно, можно уменьшить путем выборки. • Экономия времени.
Скорость принятия решения может быть обеспечена только в том случае, если исследование ограничено выборкой . • Расширенные возможности В некоторых областях полный анализ всего населения невозможен; Таким образом, выбор заключается между использованием выборки или откладыванием исследования на полку. • Точные результаты
Поскольку одной из целей выборки является воссоздание репрезентативного образа характеристик совокупности, более крупные образцы не дали бы результатов, которые были бы значительно более точными.

Только методы случайной выборки могут оценить ошибку из-за выборки, поскольку она основана на расчете вероятностей.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Выборочное обследование
Сбор данных
Оценка
Оценщик
Выборка по квоте
Простая случайная выборка
Стратифицированная выборка
Систематическая выборка

ССЫЛКИ

Боули, А.Л.: Обращение президента к экономическому отделу Британской Ассоциации ссылка Дж. Рой. Стат. соц. 69, 540–558 (1906)

Деланиус, Т.: Элементы выборки обследования. CAPEK, Стокгольм (1985)

Desrosières, A. La partie pour le tout: com ment généraliser? La préhistoire delacon trainte de représentativité. Оценка и зондаж. Cinq взносы à l'histoire de la statistique. Экономика, Париж (1988)

Foody, WN, Hedagat, AS: О теории и применении конструкций BIB с повторяющимися блоками. Статистические анналы, V = S 932–945 (1977).

Хедаят, А.С., Синха, Б.К.: Дизайн и вывод в выборке конечной совокупности. Уайли, Нью-Йорк (1991)

Тасси, доктор философии: De l'exhaustif au partiel: un peu d'histoire sur le développement des sondages. Оценка и зондаж, несколько вкладов в историю статистики. Экономика, Париж (1988)

Machine Translated by Google

Выборочное распределение

Выборочное распределение – это распределение статистики T(X1,... , Xn), где T (.) функция наблюдаемых значений побега dom переменные X1,...,Xn.

ИСТОРИЯ

См. выборку.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим популяцию размера N и параметры μ и σ (μ — неизвестная среднее значение населения и σ его стандарт отклонение). Берем выборку размера n из это население. Количество возможных выборки размера n задается количеством комбинаций n объектов среди N:

$$к = Сп\ н \overset{"="}{} \frac{Н!}{(Н - п)! \cdot н!} .$$

Каждая выборка может быть охарактеризована, например, своим выборочным средним, обозначаемым как:

$$x\bar{}1, x\bar{}2,..., x\bar{}к .$$

Совокупность этих k средних формирует выборку распределение средств.

Примем следующие обозначения для охарактеризовать среднее и стандартное отклонение выборочное распределение означает:

Среднее значение: $\mu x\bar{}$ и стандартное отклонение: $\sigma x\bar{}$.

Среднее значение выборки Распределение средних

Рассмотрим X1,...,Xn независимые случайные переменные распределены согласно распределению со средним μ и дисперсией σ2. К определение среднего арифметического, среднего выборочного распределения средств равно:

$$\mu x\bar{} = \frac{\sum_{я=1}^{к} x\bar{}я}{к} ,$$

где k - количество выборок размера n которые можно сформировать из населения.

Ожидаемое значение среднего значения выборочного распределения средних равно среднее значение населения μ :

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

Потому что:

$$E[\mu x\bar{}] = E\left[\frac{\sum_{я=1}^{к} x\bar{}я}{1к}\right] \overset{"="}{} \frac{1}{к} \cdot Э \sum_{я=1}^{к} x\bar{}я$$

$$\overset{"="}{} \frac{1}{к} \cdot \sum_{я=1}^{к} E[x\bar{}i]$$

$$\overset{"="}{} \frac{1}{к} \cdot \sum_{я=1}^{к} E\left[\frac{\sum_{j=1}^{н} xj}{н}\right]$$

$$\overset{"="}{} \frac{1}{к} \cdot \sum_{я=1}^{к} \frac{1}{н} \cdot Э \sum_{j=1}^{н} хдж$$

$$\overset{"="}{} \frac{1}{к} \cdot \sum_{я=1}^{к} \frac{1}{н} \cdot \sum_{j=1}^{н} E\,[xj]$$

$$\overset{"="}{} \frac{1}{к} \cdot \sum_{я=1}^{к} \frac{1}{н} \cdot п \cdot мк$$

$$\overset{"="}{} \frac{1}{к} \cdot k \cdot \mu \text{ знак равно } \mu .$$

Стандартное отклонение выборки Распределение средств

По определению, стандартное отклонение выборочное распределение средств, а также

С

Machine Translated by Google

называется стандартной ошибкой, равна:

$$\sigma_{\bar{x}} = \frac{\sqrt{\sum_{i=1}^{k} (\bar{x}_i - \mu_{\bar{x}})^2}}{k}.$$

Точно так же, как существует связь между средним значением выборочного распределения средних значений и средним значением генеральной совокупности, существует также связь между стандартной ошибкой и стандартным отклонением.

Оценка совокупности: • В случае исчерпывающей выборки (без замены) на конечной совокупности стандартная ошибка среднего определяется как:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}},$$

где σ — стандартное отклонение Население. •

Если совокупность бесконечна или изучаемая выборка не является исчерпывающей (с замещением), стандартная ошибка среднего равна стандартному отклонению совокупности , деленному на квадратный корень размера выборки :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Характеристики выборочного распределения средств

• Если он достаточно велик (n 30), выборочное распределение средних приблизительно соответствует нормальному распределению, независимо от распределения генеральной совокупности . • Если совокупность распределена в соответствии с нормальным распределением, то выборочное распределение средних всегда следует нормальному распределению, независимо от размера выборки .

• Выборочное распределение средних значений всегда симметрично относительно их среднего значения.

Аналогичные соображения можно сделать для статистические данные , отличные от среднего арифметического.

Рассмотрим распределение дисперсий получена из всех возможных выборок размера n, взятых из нормальной совокупности с дисперсией σ2. Это распределение называется выборочным распределением дисперсий.

Одна из особенностей этого распределения заключается в том, что она может принимать только неотрицательные значения, причем дисперсия s2 определяется суммой квадратов.

Выборочное распределение дисперсий связано с распределением хи-квадрат следующим соотношением:

$$2\,c \text{ "=" } \frac{\chi^2\sigma^2}{в} \text{ "=" } \frac{\chi^2\sigma^2}{п - 1},$$

где χ2 — случайная величина , распределенная по закону хи-квадрат с v = n 1 степенью свободы.

Зная математическое ожидание и дисперсию распределения хи-квадрат:

$$Ex2 = v _ ,$$

$$Вар\ \chi^2 = 2v ,$$

мы можем определить эти же характеристики для распределения s2 в случае, когда выборки происходят из бесконечной совокупности.

Среднее выборочное распределение Отклонения

Среднее значение распределения s2 равно дисперсии населения :

$$мкс2 = E\,c^2 = E\left(\frac{\chi^2\sigma^2}{в}\right)$$

$$\text{"="} \frac{\sigma^2}{в} \cdot E\,x2$$

(поскольку σ2 /v — константа)

Machine Translated by Google

$$\text{"="} \quad \frac{\sigma 2}{\text{в}} \cdot \text{в}$$

$$= \sigma 2 \quad .$$

Дисперсия выборочного распределения отклонений

Дисперсия распределения $s2$ _ зависит не только от дисперсии населения $\sigma 2$, но и от размера образцы $n = v + 1$:

$$\sigma 2_{c2} = \text{Вар с} \quad 2 = \text{Вар} \quad \frac{\chi 2 \sigma 2}{\text{в}}$$

$$\text{"="} \quad \frac{\sigma 4}{v2} \cdot \text{Вар } \chi 2$$

$$\text{"="} \quad \frac{\sigma 4}{v2} \cdot 2\text{в}$$

$$= 2 \cdot \frac{\sigma 4}{\text{в}} \quad .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Изучение выборочного распределения статистика позволяет судить о близости статистики, измеренной на одном образце с неизвестным параметром населения . Поэтому он также позволяет указать допустимая погрешность оценок , рассчитанная из самого образца.

Понятие выборочного распределения поэтому в основе конструкции доверительные интервалы, указывающие степень точности оценки , а также осуществление проверки гипотез на неизвестные параметры населения .

ПРИМЕРЫ

На примере покажем связь, существующую между средним значением выборки распределение и среднее значение населения и связь между стандартной ошибкой и стандартное отклонение этой совокупности.

Рассмотрим совокупность пяти магазинов, чьи средняя цена на определенный продукт, который мы хотим знать. В следующей таблице указана цена продукта для каждого магазина, принадлежащего Население:

Хранить нет.	Цена
1	30.50
2	32.00
3	37.50
4	30.00
5	33.00

Среднее значение населения равно :

$$\text{м} = \frac{\sum_{\text{я}=1}^{5} xi}{5}$$

$$\text{"="} \quad \frac{30{,}50 + 32 + 37{,}50 + 30 + 33}{5}$$

$$= 32{,}60 \quad .$$

Отбирая выборки размера 3, мы можем определить различные способы выборки, которые могут результат случайного выбора выборки.
Мы получаем:

$$\text{к} = \text{С3}_5 \quad \text{"="} \quad \frac{5!}{3! \cdot 2!} = 10 \text{ возможных выборок.}$$

В следующей таблице представлены различные образцы и соответствующие им образцы. означает:

Количество магазинов в выборке		Выборка _ означает $x_{\text{я}}$
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..

C

Machine Translated by Google

Количество магазинов в выборке		Выборка означает $\bar{x}_я$
7.	2, 3, 4	33.166..
8.	2, 3, 5	34.166..
9.	2, 4, 5	31.666..
10.	3, 4, 5	33.500
	Общий	326.000

По данным этой таблицы мы можем рассчитать среднее значение выборочного распределения означает:

$$\mu_{\bar{x}} = \frac{\sum_{я=1}^{к} \bar{x}_я}{к} = \frac{326.00}{10} = 32{,}60 .$$

В этом примере мы замечаем, что среднее значение выборочное распределение равно среднему населения :

$\mu_{\bar{x}}$ знак равно μ .

Мы можем действовать таким же образом, чтобы проверить связь между стандартной ошибкой и стандартное отклонение населения . По определению, стандартное отклонение население рассчитывается следующим образом:

$$\sigma = \sqrt{\frac{\sum_{я=1}^{Н} (xi - \mu)2}{Н}} .$$

По данным первой таблицы мы получать:

$$\sigma = \sqrt{\frac{35{,}70}{5}} = 2{,}6721 .$$

По тому же определению можно вычислить стандартное отклонение выборочного распределения средних значений:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{я=1}^{к} (\bar{x}_i - \mu_{\bar{x}})^2}{к}}$$

$$= \sqrt{\frac{11.899}{10}} = 1{,}0908 .$$

В следующей таблице представлены детали расчеты:

Количество магазинов в образце		Выборка означает $\bar{x}_я$	$(\bar{x}_я - \mu_{\bar{x}})2$
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
	Общий	326.000	11.899

По полученным результатам:

$$\sigma = 2{,}6721 \text{ и } \sigma_{\bar{x}} = 1{,}0908 ,$$

можно проверить следующее равенство:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{Н - n}{Н - 1}}$$

потому что $\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}} = \frac{2.6721}{\sqrt{3}} \cdot \sqrt{\frac{5 - 3}{5 - 1}} =$ 1.0908. Эти результаты позволяют сделать вывод что, зная параметры μ и σ населения , мы можем оценить соответствующие характеристики выборки распределение средств . И наоборот, мы можно также определить параметры населения, зная характеристики выборочное распределение.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Доверительный интервал
- Оценка
- Проверка гипотезы
- Стандартная ошибка

Machine Translated by Google

диаграмма рассеяния

Диаграмма рассеяния получается путем расшифровки данные образца на графику . В двух измерений диаграмма рассеяния может представлять n пар наблюдений (xi; yi), i = 1, 2,...,н.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Диаграмма рассеяния получается путем размещения пар наблюдаемых значений в системе координат . Для двух переменных X и Y пары соответствующие точки (x1, y1), (x2, y2), ..., (xn, yn) размещены на прямоугольной оси sys тем.

Зависимая переменная Y обычно отображается по вертикальной оси (ординате) и независимая переменная X по горизонтальной оси (абсцисса).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Диаграмма рассеяния — очень полезный и мощный инструмент, часто используемый в регрессионном анализе. Каждая точка представляет собой пару наблюдаемых значений зависимой переменной и независимая переменная. Он позволяет графически определить, существует ли зависимость между двумя переменными перед выбором подходящая модель. Более того, эти графики рассеяния очень полезны при анализе остатков. поскольку они позволяют проверить, является ли модель подходит или нет.

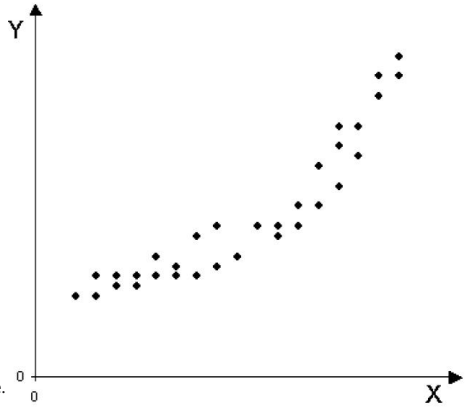
ПРИМЕРЫ

Рассмотрим следующие два примера точечных диаграмм, представляющих пары наблюдений. (xi, yi) относительно двух количественных переменных X и Y:

В этом примере распределение наблюдений, по-видимому, указывает на линейную зависимость между двумя переменными:



Распределение наблюдений в следующий рисунок, по-видимому, указывает на нелинейную зависимость между X и Y:



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Анализ остатков
- Регрессивный анализ

C

Шеффе, Генри

Шеффе Генри родился в 1907 году в Нью-Йорке. Йорк. Шеффе, докторская диссертация Генри «Асимптотические решения некоторых линейных дифференциальных уравнений, в которых Коэффициент параметра может иметь ноль», руководили Лангер, Рудольф Е. Сразу же он начал свою карьеру в качестве профессора математики в Университете Висконсина, а затем в Университете штата Орегон. университет, а также колледж Рид за пределами

Machine Translated by Google

Портленда. В 1941 году он поступил в Принстонский университет, где выросла группа специалистов по статистике. Вторая карьера началась для него как универ

городской профессор статистики. Он работал на факультете Сиракузского университета в 1944–1945 учебном году и в Калифорнийском университете в Лос-Анджелесе с 1946 по 1948 год.

Он уехал из Лос-Анджелеса, чтобы поступить в Колумбийский университет, где стал заведующим кафедрой статистики. Проработав 5 лет в Колумбийском университете Шеффе, Генри в 1953 году поступил в Беркли в качестве профессора статистики.

Шеффе, исследования Г. в основном касались дисперсионного анализа. Одна из его самых важных работ появилась в 1953 г. по S-методу одновременного доверительного взаимодействия.

vals для оцениваемых функций в подпространстве пространства параметров. Он также изучал парные сравнения и смешанные модели. В 1958 г. и снова в 1963 г. он опубликовал статьи об экспериментах со смесями, а в 1973 г.

методы калибровки.

Шеффе Х. стал научным сотрудником Института математической статистики в 1944 г., Американской статистической ассоциации в 1952 г. и Международного статистического института в 1964 г. Он

был избран президентом Международного статистического института и вице-президентом Американской статистической ассоциации. Он умер в Калифорния в 1979 году.

Избранные произведения Генри Шеффе:

1959 Дисперсионный анализ. Уайли, Нью-Йорк.

1952 г. Дисперсионный анализ для парных сравнений. Варенье. Стат. Ассоциат., 47, 381–400.

1953 Метод оценки всех контрастов в дисперсионном анализе. Биометрика, 40, 87–104.

1956 г. Альтернативные модели дисперсионного анализа. Анна. Мат. стат., 27, 251–271.

1973 г. Статистическая теория калибровки. Анна. Стат., 1:1–37.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Дисперсионный анализ

Тест на наименьшее существенное различие

Сезонный индекс

Рассмотрим временной ряд Yt , компоненты которого являются:

• Вековой тренд Tt • Циклические колебания Ct • Сезонные колебания St • Нерегулярные колебания It

Влияние сезонных колебаний должно быть нейтральной в течение всего года, а сезонные вариации St теоретически повторяются одинаковым образом от периода к периоду (что не проверяется в реальном случае).

Чтобы удовлетворить требованиям теоретической модели и иметь возможность изучать реальные временные ряды, вместо наблюдаемых St необходимо оценивать периодические вариации, одинаковые каждый год (месяц за месяцем или триместр за триместром), называемые сезонными вариациями. Они обозначаются Sj, где j изменяется следующим образом:

j = от 1 до 12 месяцев (более n лет) или

j = от 1 до 4 по триместрам (за n лет).

Обратите внимание, что за n лет существует только 12 сезонных индексов, относящихся к месяцам, и 4, относящихся к триместрам.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Существует множество методов расчета сезонных индексов:

Machine Translated by Google

1. Процент среднего

Данные, соответствующие каждому месяцу, выражается в процентах от среднего. Мы вычисляем его, используя среднее (без с учетом экстремальных значений) или медиана процентов каждого месяца за разные годы. Принимая среднее значение за каждый месяц мы получаем 12 процентов за заданные индексы. Если среднее не равно 100%, мы должны сделать корректировку на умножая индексы на удобный фактор.

Рассмотрим месячный временной ряд в течение n лет, обозначаемых через Yi,j, где j = 1, 2,..., 12 для разных месяцев и i = 1,...,n, указанный годом.

Сначала сгруппируем данные в таблицу тип:

	Месяц		
1 год	Январь ...	...	Декабрь 12
1	Y1,1	...	Y1,12
...	...	...	...
н	Ин,1	...	Ин, 12

Затем вычисляем среднегодовые значения для n лет :

$$\text{Ми} = \frac{\bar{Y}_{i,.}}{12}$$

для i = 1,...,n.

Таким образом, мы вычисляем для каждого элемента данных:

$$\text{Ци,j} = \frac{\text{й, дж}}{\text{Ми}} \quad \text{для i} = 1,\ldots,n$$

и j = 1,..., 12, о чем мы сообщаем в следующая таблица:

	Месяц		
1 год	Январь ...	...	Декабрь 12
1	Q1,1	...	Q1,12
...	...	...	...
н	Qn,1	...	Qn,12

В каждом столбце мы включаем арифметику среднее значение, обозначаемое через $\bar{Q}_{i,.}$, и мы вычисляем сумма этих 12 средних; если это не так равным 1200, то введем поправочный коэффициент:

$$\text{к} = \frac{1200}{\sum_{\text{я}=1}^{12} \bar{Q}_{\text{я},.}}.$$

Наконец, сезонные индексы даны для каждый месяц на среднее значение, умноженное на k.

2. Коэффициент тенденции

Данные каждого месяца выражаются как проценты от месячной стоимости светская тенденция. Индекс получен из соответствующего среднего процента возрастов соответствующих месяцев. Мы следует скорректировать результаты, если среднее значение не равно 100% для получения искомых сезонных индексов.

3. Коэффициент скользящего среднего

Рассчитываем скользящую среднюю по 12 месяцы. Поскольку соответствующие результаты растянутый на многие месяцы, вместо быть в середине месяца, как в исходных данных, вычисляем скользящее среднее за 2 месяца из скользящего среднего за 12 месяцев. Это то, что мы называем центрированной скользящей средней по 12 месяцев.

Затем выразим исходные данные каждый месяц в процентах от соответствующей центрированной скользящей средней за 12 месяцев. Мы берем среднее значение соответствующих процентов каждого месяца, который дает индексы. Мы должны тогда настроить на 100%, чтобы получить желаемые сезонные индексы.

4. Относительные цепочки

Данные каждого месяца выражаются в процентах к данным предыдущего месяца (отсюда выражение относительная цепочка). Мы рассчитываем подходящий

Machine Translated by Google

среднее значение цепей для соответствующих месяцы. Мы можем получить относительные проценты каждого месяца по отношению к тем, января (= 100%). Мы находим тогда что в следующем январе обычно будет процент, равный или меньший, или более 100%. Таким образом, тенденция либо увеличивается, либо уменьшается. индексы, полученные в итоге, с поправкой на их средние, равные 100%, дают искомые сезонные индексы.

ПРИМЕРЫ

Мы рассматриваем временной ряд любого явления; у нас есть данные за триместр 6 лет, суммированные в следующей таблице:

1-й год		2-й	3-й	4-й
	подрезать.	подрезать.	подрезать.	подрезать.
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

Сезонные индексы определяем с помощью процент от среднего:
Сначала вычисляем среднее арифметическое для 6 лет:

Год	Сумма	Среднее
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

Разделим строки исходной таблицы на среднее значение соответствующего года за три месяца и умножить на 100.

Тогда, если результаты показывают достаточно сильные отклонения полученных значений для одного и того же триместр, лучше брать медиану вместо среднего арифметического, чтобы получить знаковые числа.

Год	1-й	2-й	3-й	4-й
	подрезать.	подрезать.	подрезать.	подрезать.
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
Медиана	103,84	95,89	108,33	91,94

Средний процент для каждого триместра составляет указана в последней строке предыдущей таблицы. сумма этих процентов ровно 400%, корректировка не требуется, и числа, которые мы имеем в последней строке, представляют собой желаемые сезонные показатели.
Мы видели, что есть три других метамфетамина. методы определения сезонных индексов и они дают следующие значения для одного и того же проблема:

	Триместр			
Метод	1	2	3	4
Процент среднего	103,84	95,89	108,33	91,94
Тенденция соотношение	112,62	98,08	104,85	84,45
Соотношение движущийся средний	111,56	98,87	106,14	83,44
Цепи	112,49	98,24	104,99	84,28

Как видим, результаты согласуются друг с другом, несмотря на разнообразие методов, используемых для определить их.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Циклическое колебание
- Неправильная вариация
- Сезонная вариация
- Светская тенденция

Временная последовательность

Сезонная вариация

Сезонные колебания составляют одну из основных компоненты временного ряда. Они есть вариации периодического характера, которые регулярно повторяются. В экономическом контексте сезонные колебания могут быть вызваны следующими событиями:

- Климатические условия
- Таможня, соответствующая населению
- Религиозные праздники

Это периодические колебания более или менее регулярные, которые накладываются на внесезонное движение.

ИСТОРИЯ

Смотрите временной ряд.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Оценка сезонных колебаний производится путем определения индекса сезонной изменчивости, называемого сезонным индексом.

ПРИМЕРЫ

Мы рассматриваем временной ряд любого явления; у нас есть данные за триместр в течение 6 лет, суммированные в следующем стол:

1-й год		2-й	3-й	4-й
	подрезать.	подрезать.	подрезать.	подрезать.
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

Сезонные индексы определяем с помощью процент от среднего:

Во-первых, мы вычисляем трехместную арифметику значит на 6 лет:

Год	Сумма	Среднее
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

Разделим строки исходной таблицы на среднее значение соответствующего года за три месяца и умножить на 100.

Затем, если результаты показывают достаточно большие отклонения для полученных значений за тот же триместр, вместо этого лучше взять медиану получить более значащие числа.

Год	1-й	2-й	3-й	4-й
	подрезать.	подрезать.	подрезать.	подрезать.
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
Медиана	103,84	95,89	108,33	91,94

Средний процент для каждого триместра составляет дается последней строкой предыдущей таблицы. Поскольку эти проценты в сумме составляют ровно 400%, корректировка не требуется, и

Machine Translated by Google

числа в последней строке представляют желаемые сезонные индексы.

Мы видели, что существуют и другие методы определить сезонные индексы; они дают следующие значения для одной и той же задачи:

	Триместр			
Метод	1	2	3	4
Процент от среднего	103,84	95,89	108,33	91,94
Коэффициент тенденции	112,62	98,08	104,85	84,45
Соотношение перемещение ср.	111,56	98,87	106,14	83,44
Цепи	112,49	98,24	104,99	84,28

Мы видим, что результаты согласуются друг с другом, несмотря на разнообразие методов их определения.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Среднее арифметическое

Циклическое колебание

Прогнозирование

Индекс

Скользящее среднее

Сезонный индекс

Светская тенденция

Временная последовательность

ССЫЛКИ Фуллер, А.В.: Введение в статистические временные ряды. Джон Вили'н'Сас (1976)

Светская тенденция

Светское течение образует одно из четырех основные компоненты временного ряда. Он описывает долгосрочное движение временного ряда, который в глобальном масштабе может быть возрастающим, убывающим или стабильным.

Светская тенденция может быть линейной или нет. в В последнем случае мы должны выбрать тип функции (экспоненциальная, квадратичная, логистическая и т. д.), которая лучше всего адаптируется к наблюдаемому распределению на графическом представлении временного ряда, прежде чем оценивать вековой тренд.

ИСТОРИЯ

Смотрите временной ряд.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть Yt будет временным рядом. Его можно разложить с помощью четырех компонентов: • Вековой тренд Tt • Циклическое колебание Ct • Сезонные вариации St • Нерегулярные вариации It При изучении временного ряда мы начинаем с анализа векового тренда, сначала решая, может ли этот тренд быть предполагается линейным или нет.

Если предположить, что это линейно, то оценку вековой тенденции можно произвести разными способами: 1. Графический метод или метод поднятой руки.

После графического представления временного ряда оно состоит в корректировке линии или кривой тренда при поднятой руке на графике. Это приближение к светскому тренду.

2. Метод наименьших квадратов. Он позволяет математически определить линию, представляющую наилучшие данные, то есть:

$$Yт = а + б \cdot т \quad ,$$

где Yt — значение временного ряда в данный момент времени, a — значение линии, когда t находится в начале координат, b — наклон линии, а t соответствует другому значению за весь выбранный период.

Линия наименьших квадратов Y^t соответствует оценке векового тренда Tt.

Machine Translated by Google

Вычисление векового тренда.

Предположим, что время временного ряда — это год, и начнем с идеи о перемещении начала координат в соответствующий год в середине тех лет, в которые проводились измерения.

а. Четное число лет n Пусть

j = n 1/2, и положим xi = ti tj для i = 1, 2,...,n; середина года tj является новым началом, а единицей измерения новой переменной x является год.

Таким образом

$$\hat{} = \frac{\text{и}^{\text{н}}_{\text{я=1}}}{\text{н}} \quad \text{и } b\hat{} = \frac{\text{xiYi}^{\text{н}}_{\text{я=1}}}{x2_{\text{я}}}$$

дает нам Y^t = ^a+bx^ или Y^t = (a^ bt^ j)+ bt^, где t — год без изменения начала координат.

б. Нечетное число лет n

Пусть j = n/2 + 1, где tj — новое начало координат, положим xi = 2 · (ti tj) + 1 для i = 1, 2,...,n.

Таким образом

$$\hat{} = \frac{\text{и}^{\text{н}}_{\text{я=1}}}{\text{н}} \quad \text{и } b\hat{} = \frac{\text{xiYi}^{\text{н}}_{\text{я=1}}}{x2_{\text{я}}}$$

дайте нам Y^t = ^a + bx^ . Единица x – это полугодие, отсчитываемое с 1 января года tj.

3. Метод скользящего среднего

Пусть Y1, Y2, Y3,... заданный временной ряд. Построить новый ряд:

$$\frac{\text{У1} + \ldots + \text{Ун}}{\text{н}}; \quad \frac{\text{Y2} + \ldots + \text{Yn+1}}{\text{н}}, \ldots;$$

$$\frac{\text{Yj} + \ldots + \text{Yn+j} \quad 1}{\text{н}}, \ldots$$

чья оценка светского тренда

сделанный одним из двух предыдущих методов.

4. Метод полусредних

Пусть Y1, Y2,...,Yn — заданный временной ряд. Обозначим через k целую часть $\frac{-}{\text{п}\,2}$. числа. Вычислим:

$$\text{y1} = \frac{\text{У1} + \ldots + \text{Ук}}{\text{к}} \quad \text{и}$$

$$= \text{n} \quad \frac{\text{Yk+1} + \ldots + \text{Yn y2}}{\text{kk} + 1\ \text{x1} =}$$

$$\text{x2} = \frac{2\ \text{п} + \text{к} + 1}{2} \quad \text{и}$$

и определить линию векового тренда проходит через две точки:

(x1; y1) и (x2; y2).

Когда вековой тренд нельзя считать линейным, мы уточняем метод наименьших квадратов. Мы должны применить его, заменив Yt = a + bt другой кривой тренда, например: • Экспоненциальная: Yt = a · bt • Модифицированная экспоненциальная: Yt = c + a · bt • Логистическая кривая: Yt = (c + a · bt) 1 • Кривая Гомперца: Yt = c · abx

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Анализ вековой тенденции преследует следующие цели: •

Создать описательную модель ситуации прошлого. ация.

• Делайте проекции постоянной структуры. • Устранить вековые тенденции для изучения других компонентов временного ряда.

Четыре метода оценки Secu

лар тенденции имеют определенные недостатки. Назовем основные для каждого: 1.

Графический метод

Слишком субъективен.

С

Machine Translated by Google

2. Метод наименьших квадратов

Этот метод можно использовать только в период, когда движение полностью то же направление. Когда первая часть соответствует восходящему движению и другая часть к нисходящему движению, должны быть созданы две длинные линии, чтобы приспособиться к этим данным, каждая из которых относится к к периоду с уникальным направлением.

У этих линий есть несколько интересных свойств:

- На этом же графе находим множество значений Yi временного ряда и тренда линия. Сумма отклонений между наблюдаемые значения Yi и оценка светского теченияYˆi всегда будет равным нуль:

$$\sum_{i=1}^{n} (Yi - \hat{Y}i) = 0 .$$

- Линия тренда минимизирует сумму отклонений:

$$\sum_{i=1}^{n} (Yi - \hat{Y}i)^2 = \text{минимальное значение} .$$

3. Метод скользящего среднего

Данные о начале и конце исходные временные ряды «теряются». Скользящие средние могут создавать циклы и другие движения, отсутствующие в исходных данных; кроме того, они сильно подвержены влиянию случайные выбросы. Этот метод позволяет выявить изменения направления.

4. Метод полусредних

Даже если этот метод прост в применении, он может давать результаты без ценности. Мысль действителен в случае, когда данные могут быть классифицированы на две группы, где тенденции линейный метод применим только тогда, когда тенденция глобально линейна или приблизительно линейна. Если процедура легкая приложение, это не очень строго.

ПРИМЕРЫ

Оценим вековой тренд годового продажи фабрики печенья Surfin с использованием следующие данные:

Годовой объем продаж фабрики печенья Surfin (в млн. евро)

Год продаж (млн. $)
1975 7,6
1976 6,8
1977 8,4
1978 9,3
1979 12,1
1980 11,9
1981 12,3

Воспользуемся методом наименьших квадратов : Давайте переместим начало координат в середину 7 лет, то есть 1978. Заполним следующую таблицу:

Год продаж (млн евро) Xt	Yt	Код год t	tYt	t^2
1975	7,6	3	22,2	9
1976	6,8	2	13,6	4
1977	8,4	1	8,4	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
Всего	68,4	0	28,0	28

Положим ti = Xi 1978 для i = 1, 2,..., 7.

середина 1978 года-это новое происхождение, а устройство t - год.

Таким образом:

$$\hat{} = \frac{\sum_{t=-3}^{3} Yt}{7} "=" \frac{68,4}{7}$$

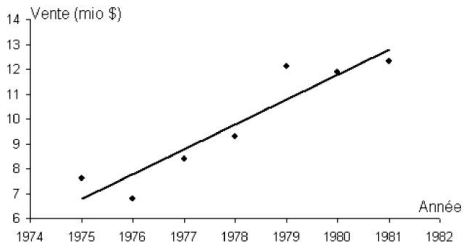
= 9,7714 (млн евро) и

Machine Translated by Google

$$\hat{б} = \frac{\sum_{т=-3}^{3} tYt}{\sum_{т=-3}^{3} т^2} \text{ "=" } \frac{28}{} = 1{,}0 \text{ (миллион евро) , 28}$$

что дает нам Tt = Yˆt = 9,7714 + t для уравнения линии векового тренда.

Это дает графически:



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Циклическое колебание
Графическое представление
Наименьших квадратов
Скользящее среднее
Временная последовательность

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Чатфилд, К.: Анализ временных рядов. Введение. Чепмен и Холл (2003)

Полулогарифмический Сюжет

Полулогарифмический график — это графическое представление временного ряда. Он определяется арифметической шкалой времени t , отложенной по оси абсцисс, и логарифмической шкалой, отложенной по оси ординат, что означает, что логарифм наблюдаемого значения Yt будет записан по ординате арифметической шкалы.

Обычно используются печатные полулогарифмические листы бумаги.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В системе координат начало координат и единица длины выбираются произвольно; абсцисса арифметически отградуирована по миллиметру бумага.

По оси ординат можно изменить единицу длины, выбранную изготовителем полулогарифмической бумаги, которая соответствует расстоянию между двумя последовательными степенями 10 на логарифмической шкале. Это расстояние, являющееся единицей длины соответствующей арифметической шкалы, называется модулем.

Значения , напечатанные на полях логарифмической шкалы, можно умножать на одно и то же число; эта мультипликативная константа выбрана как можно более простой, чтобы полученные значения было легко использовать для записи точек на графике.

Имеющиеся в продаже полулогарифмические листы бумаги обычно имеют один, два, три или четыре модуля.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Полулогарифмический график обладает следующими свойствами.

- Экспоненциальные кривые Y представлены линией. Если y(x) является экспонентой:

$$y(x) = c \cdot a^x ,$$

логарифмируя:

$$\log y(x) = \log c + x \cdot \log a \text{ , где } \log y(x) = m \cdot x + h$$

$$m = \log a \text{ и } h = \log c$$

где ,

находим линейное выражение по x.

- Если a = 1+i, где i — годовая процентная ставка Y , и с начальной суммой денег

C

Machine Translated by Google

Y0, то :

$$Y = Y0 \cdot (1 + i)^{\text{Икс}}$$

дает сумму денег через x лет.

Также можно работать с базой Напьера (e) , поскольку ax = ex·ln a и, следовательно, y(x) = c · ex·ln a или y(x) = c · er·x, где r = ln a равно называется мгновенной скоростью роста y. В данном случае это отношение мгновенного изменения $\frac{dy}{dx}$ к значению y .

Использование полулогарифмического графика целесообразно в следующих ситуациях:

- Когда существуют большие различия в значениях переменной , чтобы избежать выхода за пределы графика.
- Когда кто-то хочет сделать относительную вариацию появляются. •

Когда кумуляции темпов роста, приводящие к появлению экспоненциальных величин, должны быть представлены в линейной форме.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Графическое представление

Временная последовательность

Серийная корреляция

Мы называем последовательной корреляцией или автокорреляцией.

зависимость между наблюдениями временных рядов.

ИСТОРИЯ

Смотрите временной ряд.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x1,...,xn — n наблюдений случайной величины X , зависящей от времени (мы говорим здесь о временных рядах). Чтобы измерить зависимость между наблюдениями и временем, используем следующий коэффициент:

$$д = \frac{\sum_{т=2}^{н} (хт - хт\text{-}1)^2}{\sum_{т=1}^{н} x2_т},$$

что указывает на связь между двумя последовательные наблюдения. Этот коэффициент принимает значения от 0 до 4. Если он близок к 0, так что разница между двумя последовательными наблюдениями минимальна, говорят о положительной автокорреляции. Если оно близко к 4, то мы имеем отрицательную автокорреляцию: за наблюдением с большим значением может следовать низкое значение, и наоборот. В случае, когда наблюдения независимы

в зависимости от времени коэффициент d имеет значение, близкое к 2.

Мы также можем измерить последовательную корреляцию

наблюдений x1,...,xn вычисляя коэффициент корреляции между

ряды xt и ряды xt «перешли на k»:

$$рк = \frac{\sum_{t=k+1}^{н} (xt \quad \bar{x})(xt \quad k \quad x)}{\sum_{т=1}^{н} (хт - \bar{x})^2}$$

и, анализируя таким же образом, измерить последовательную корреляцию для коэффициента корреляции

слияние двух разных серий.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Авторегрессионные модели

Когда наблюдается серийная корреляция между остатками модели sim

ри линейной регрессии можно использовать удобные модели области временных рядов с

успеха, например, модели типа

$$Yt = \beta0 + \beta1Xt + \varepsilon t ,$$

Machine Translated by Google

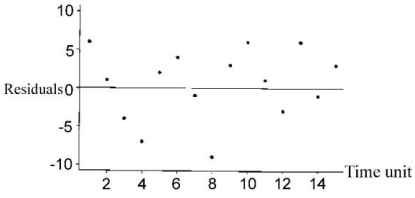
где члены ошибки εt не являются независимыми, а описываются соотношением типа

$$\varepsilon t = \rho s \varepsilon t \quad 1 + \delta t ,$$

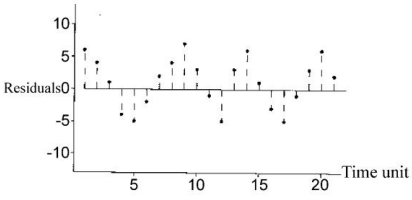
где ρs можно оценить по последовательной корреляции εt , а δt — независимые ошибки.

ПРИМЕРЫ

График остатков как функции времени Этот тип графика особенно используется при анализе временных рядов, то есть, когда время является объясняющей переменной модели.
Остатки по времени.



Графическое представление остатков во времени всегда влечет за собой появление последовательной корреляции, положительной или отрицательной, как в случае с рисунком выше. Положительная или отрицательная последовательная корреляция означает, что ошибки не являются независимыми.
Остатки, представляющие серийную корреляцию.



ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Автокорреляция
- Проверка гипотезы

Временная последовательность

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Бартлетт, М.С.: О теоретической спецификации выборочных свойств временных рядов, связанных с автокоррекцией. Дж. Рой. Стат. соц. сер. Б 8, 27–41 (1946)

Box, GEP, Jenkins, GM: Анализ временных рядов: прогнозирование и контроль (серии inTimeSeriesAnalysis). HoldenDay, Сан-Франциско (1970).

Дурбин, Дж., Уотсон, Г.С.: Тестирование последовательной корреляции в регрессии по методу наименьших квадратов, II. Биометрика 38, 159–177 (1951).

Подписать тест

Критерий знаков является непараметрическим критерием и может использоваться для проверки гипотезы об отсутствии различий между распределением две случайные величины. Своим названием мы обязаны тот факт, что он использует знаки «+» и «-» вместо количественных значений. Таким образом, он применим даже тогда, когда ряд имеет порядковую меру.

ИСТОРИЯ

В своей первой публикации о непараметрических критериях Арбатнотт Дж. (1710 г.) изучил список рождений, зарегистрированных в Лондоне за период в 82 года. Для каждого года он сравнивал число детей, рожденных каждым полом. Он обозначил знаком «+» событие «рождается больше мальчиков, чем девочек», а знаком «-» противоположное событие. (равенства не было).
К его величайшему удивлению, он 82 раза употребил знак «+» и ни разу — знак «-»; это заставило его отвергнуть нулевую гипотезу о равенстве рождений по отношению к полу ребенка.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим n пар наблюдений (x1, y1), (x2, y2),...,(xn, yn). Для каждой пары (xi, yi)

Machine Translated by Google

мы делаем следующее сравнение в соответствии с размерностью xi относительно yi :

«+ если xi < yi

« если xi > yi

“ = если xi = yi

Вычитаем из размерности сам ple количество отображаемых знаков «=», что то есть учитываем только те пары, которые представляют собой положительную или отрицательную разницу. Обозначим через m это количество пар. Считаем затем количество знаков «+», которую мы обозначаем T.

Гипотезы

По тесту односторонняя или двусторонняя нулевая и альтернативная гипотезы соответствующие тесту:

A: Двусторонний чехол:

H0 : P (X < Y) = P (X > Y)

H1 : P (X < Y) = P (X > Y)

B: Односторонний чехол:

H0 : P (X < Y) P (X > Y)

H1 : P (X < Y) > P (X > Y)

C: Односторонний случай:

H0 : P (X < Y) P (X > Y)

H1 : P (X < Y) < P (X > Y)

В случае A мы делаем нулевую гипотезу (H0) что вероятность того, что X меньше, чем Y равна вероятности того, что X больше чем Ю.

В случае B мы априори предполагаем, что вероятность того, что X меньше, чем Y, меньше, чем или равно вероятности того, что X больше чем Ю.

Наконец, в случае C мы априори предполагаем, что вероятность того, что X меньше Y, равна больше или равна вероятности того, что X больше Y.

В двустороннем случае вероятность того, что X меньше Y (так как X > Y) равно $\frac{1}{2}$.

Знаковый тест является частным случаем бино миальный тест с p = $\frac{1}{2}$, то есть то, что под нулевая гипотеза H0 мы хотим получить как можно больше «+» как «-».

Правила принятия решений

Случай A

Мы используем биномиальную таблицу и ищем число X с соответствующим двучленом вероятность равна $\frac{1}{2}$.

Нам нужно найти в таблице значение, наиболее близкое к $\frac{\alpha}{2}$, где α – уровень значимости.

Мы отвергаем H0 на уровне α, если:

T tα/2 или T m tα/2 ,

где T = количество «+».

Когда m > 20, мы можем использовать обычную таблицу как приближение биномиального распределения.

Преобразуем статистику T в случайную переменная Z после стандартного нормального распределение:

$$Z = \frac{T - \mu}{\sigma},$$

где

$$\mu = m \cdot p = m \frac{1}{2}$$

$$\text{и } \sigma = m \cdot p \cdot q = 2 \quad \frac{1}{2} \quad \frac{1}{m}$$

среднее значение и стандартное отклонение биномиальное распределение.

Таким образом получаем:

$$Z = \frac{T - \frac{1}{2} m}{\frac{1}{2} \frac{1}{m}}.$$

Machine Translated by Google

Аппроксимация значения биномиальной таблицы тогда:

$$t\alpha/2 = \frac{1}{2} m + z\alpha/2 \sqrt{m},$$

где zα/2 находится в обычной таблице на уровне 2 . Тогда $\frac{\alpha}{}$ решающее правило та же.

Случай Б

Мы отвергаем H0 на уровне α, если:

$$T \quad м - t\alpha ,$$

где tα – значение соответствующей биномиальной таблицы и уровень значимости α (или ближайшее значение).

При m > 20 делаем приближение

$$1\ t\alpha = \frac{}{2} м + z\alpha \sqrt{m},$$

где zα находится в нормальной таблице на уровне α.

Случай С

Мы отвергаем H0 на уровне α, если:

$$T \quad t\alpha ,$$

где значение α такое же, как и для случая B.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Требования использования знакового теста следующие: 1.

Пары случайных величин (Xi, Yi), i = 1, 2,...,n должны быть взаимно независимыми.

2. Шкала измерения должна быть не ниже ordi nal, то есть с каждой парой может быть связан знак «+», «-» или «=».

3. Пары (Xi, Yi), i = 1, 2,...,n, должны представлять логику между собой, т. е. если P "+ > P " для одной пары (Xi, Yi), то P «+ > P » для всех остальных пары.

Тот же принцип применяется, если P "+ < P "- или P "+ = P "- .

ПРИМЕРЫ

Производитель шоколада изучает новый тип упаковки для одного из своих продуктов. С этой целью он предлагает 10 потребителям оценить старую и новую упаковку по шкале от 1 до 6 (где 1 означает «категорически не нравится», а 6 означает «сильно нравится»).

Мы выполняем односторонний тест (соответствующий случаю С), где записана нулевая гипотеза, которую мы хотим проверить:

$$H0 : P "+ \quad P "-$$

относительно альтернативной гипотезы:

$$H1 : P "+ < P "- \quad .$$

Знак «+» означает, что новая упаковка предпочтительнее старой.

Получены следующие результаты:

Против-шумер	старый упаковка	Классифицировать новый упаковка	Знак отличия источник
14	5		+
23	5		+
31	1		"="
45	4		
54	4		"="
62	6		+
74	6		+
84	5		+
92	3		+
10	1	4	+

В таблице показано, что два потребителя сказали разницы между новой и новой не было. старая упаковка. Таким образом, имеется восемь различий (m = 8), среди которых новая упаковка в семи случаях предпочтительнее старой. Таким образом, мы имеем T = 7 (число «+»).

Поэтому мы отвергаем H0 на уровне α, если

$$T \quad t\alpha ,$$

Machine Translated by Google

где tα – значение биномиальной таблицы.

Выбираем α = 0,05; значение tα равно 1 (при m = 8 и α = 0,0352). Таким образом, мы имеем T > tα , поскольку T = 7 и tα = 1. Мы не отвергаем H0-гипотезу: P "+ P "-, и производитель шоколада может сделать вывод, что потребители в целом предпочитают новую упаковку.

Возьмем еще раз тот же пример, но на этот раз исследование проводится на 100 потребителях. Предполагаемые результаты следующие инг:

83 человека предпочитают новую упаковку.

7 человек предпочитают старую упаковку.

10 человек равнодушны.

Таким образом, мы имеем:

$$м = 100 - 10 = 90 \quad ,$$

$$T = 83 .$$

Значение tα можно аппроксимировать нормальным распределением:

$$1\ t\alpha \underset{=2}{—} м + z\alpha \quad \overline{m} ,$$

где zα — значение нормальной таблицы при α = 0,05. Это дает:

$$1\ t\alpha \underset{=2}{—} 90 - 1{,}64 \quad \overline{90}$$

$$= 37{,}22 .$$

Если T больше tα (83 > 37,22), то нулевая гипотеза H0: P "+ P " не отвергается и производитель шоколада делает тот же вывод, что и раньше.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Биномиальная таблица
- Биномиальный тест
- Проверка гипотезы
- Непараметрический тест
- Нормальное распределение

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Арбетнотт, Дж.: Аргумент в пользу Божественного Провидения, основанный на постоянной регулярности, наблюдаемой в рождении обоих полов. Фи лос. Транс. 27, 186–190 (1710) (3.4).

Уровень значимости

Уровень значимости является параметром проверки гипотезы, и его значение заранее фиксируется пользователем.

Уровень значимости проверки гипотезы, обозначаемый α, представляет собой вероятность отклонения нулевой гипотезы H0 , когда она верна:

$$\alpha = P\{\text{отклонить } H0 \mid H0 \text{ верно}\} .$$

Уровень значимости также называют вероятностью ошибки первого рода.

ИСТОРИЯ

В 1928 г. Ежи Нейман и Эгон Пирсон обсудили проблемы, связанные с тем, можно ли считать, что выборка вероятно взята из определенной популяции. Они выявили два типа ошибок, связанных с возможными решениями. Вероятность ошибки I рода (отвержение данной гипотезы верно) соответствует уровню значимости теста.

См. проверку гипотез.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Обеспокоенные возможностью ошибки из-за отклонения нулевой гипотезы, когда она верна, статистики строят проверку гипотезы таким образом, чтобы вероятность ошибки этого типа не превышала фиксированного значения, заданного заранее. Это значение соответствует уровню значимости теста.

Machine Translated by Google

Обратите внимание, что статистики, проводящие проверку гипотез, должны столкнуться со вторым типом ошибок. ошибка, называемая ошибкой типа II. Это имеет место когданулевая гипотеза не отвергаетсякогда на самом деле это неверно.

Чтобы уменьшить глобальный риск ошибки при принятии решений, недостаточно уменьшить уровне значимости, но компромисс должен находиться между уровнем значимости и вероятность ошибки второго рода.

Один метод основан на изучении мощности теста.

ПРИМЕРЫ

В следующем примере мы иллюстрируем влияние выбора значимости уровне при проверке гипотезы.

Рассмотрим правосторонний тест на среднее μ совокупности , дисперсия которой σ2 равняется 36.

Сформулируем следующую гипотезу:

Нулевая гипотеза H0 : м = 30

Альтернативная гипотеза H1 : μ > 30 .

Предположим, что среднее $\bar{x}$ выборки размерности n = 100, взятый из этой совокупности, равен 31.

Дело 1

Выберите уровень значимостиα = 5%. Мы находим критическое значение zα = 1,65 в нормальной таблице . Верхний предел приема регион определяется:

$$\mu_{\bar{x}} + z\alpha \cdot \sigma_{\bar{x}},$$

где $\mu_{\bar{x}}$ — среднее выборочного распределения средних значений, а $\sigma_{\bar{x}}$ — стандарт ошибка (или стандартное отклонение выборочного распределения средних):

$$\sigma_{\bar{x}} = \frac{o}{\sqrt{n}} \cdot$$

Этот верхний предел равен:

$$30 + 1{,}65 \cdot = 30{,}99 \frac{6}{\sqrt{100}}$$

Область отклонения нулевой гипотезы соответствует интервалу [30,99; [а область приемлемости на интервал] - ; 30,99[.

Случай 2

Если зафиксировать меньший уровень значимости, для Например, α = 1%, значение zα в нормальной таблице равно 2,33, а верхний предел область приемлемости теперь равна:

$$\mu_{\bar{x}} + z\alpha \cdot \sigma_{\bar{x}} = 30 + 2{,}33 \cdot \frac{6}{\sqrt{100}}$$

$$= 31{,}4 .$$

Область отклонения нулевой гипотезы соответствует интервалу [31,4; [и область приема на интервал] ; 31,4[.

Из наблюдаемых средних $\bar{x} = 31$ мы можем видим, что в первом случае (α = 5%) мы должны отклонить нулевую гипотезу, поскольку выборочное среднее (31) находится в области отклонения. В втором случае (α = 1%) необходимо сделать обратное решение: нулевая гипотеза не может быть отвергнута, поскольку среднее значение выборки находится в зоне приема.

Мы видим, что выбор уровня значимости оказывает прямое влияние на решение о нулевой гипотезы и, как следствие, на процесс принятия решений. Меньший уровень значимости (и, следовательно, меньшая вероятность ошибки первого рода) ограничивает области отторжения и увеличивает прием область нулевой гипотезы.

Меньший риск ошибки связан с меньшая точность.

C

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Ошибка I типа

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Нейман Дж., Пирсон Э.С.: Об использовании и интерпретация определенных тестовых критериев для цели Статистического вывода Часть I (1928). Перепечатано издательством Cambridge University Press в 1967 г.

Простой номер индекса

Простое порядковое число — это отношение двух значения, представляющие одну и ту же переменную, измеренные в двух разных ситуациях или в двух разных разные периоды. Например, простой указатель число цены даст относительную вариацию цены между текущим периодом и отчетный период. Наиболее часто используемые простые порядковые номера цена, количество и ценность.

ИСТОРИЯ

См. индекс.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Номер индекса In/0, который представляет переменную G в ситуации n относительно той же переменной в ситуации 0 (ссылка ситуации) определяется:

$$B/0 = \frac{Гн}{G0},$$

где Gn — значение переменной G в ситуации n , а G0 — значение переменной G в ситуации ситуация 0.

Как правило, простой порядковый номер выраженный в основе 100 в исходной ситуации 0:

$$B/0 = \frac{Гн}{G0} \cdot 100 .$$

Свойства простых индексов

- Идентичность: если две сравниваемые ситуации (или два периода) идентичны, значение номер индекса равен 1 (или 100):

$$B/н = I0/0 = 1 .$$

- Простой порядковый номер обратим:

$$B/0 = \frac{1}{I0/n} .$$

- Простой порядковый номер может передаваться. Рассмотрим два индекса I1/0 и I2/0. представляющие две ситуации, 1 и 2, с относительно той же базовой ситуации 0. Если мы рассмотрим номер индекса I2/1, мы можем скажи это:

$$I2/0 = I2/1 \cdot I1/0 .$$

Обратите внимание, что переносимость (также называемая цикличность) подразумевает обратимость:

Если $I0/1 \cdot I1/0 = I0/0 = 1$, то $I0/1 = \frac{1}{I1/0}$.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В экономике очень редко встречается простой индекс цифры можно использовать сами по себе; в информация, которую они содержат, относительно ограничена. Тем не менее они имеют принципиальную значение при построении составных индексов.

ПРИМЕРЫ

Рассмотрим следующую таблицу, представляющую цены на определенный товар и количества, проданные в разные периоды:

Цена		Количество	
Период 1	Период 2	Период 1	Период 2
P0 = 50	Pn = 70	Q0 = 25	Qn = 20

Machine Translated by Google

Если Pn — цена интересующего периода и P0 - цена базисного периода, мы будем иметь:

$$B/0 = \frac{\text{номер 70}}{P0\ 50} \cdot 100 = 140$$

это означает, что в базе 100 в период 1 индекс цен товара равен 140 в период 2. Цена товара увеличилась на 40% (140-100) между базисным периодом и фактическим периодом.

Мы также можем рассчитать простой индекс количества:

$$= \frac{Qn\ 20\ In/0}{Q0\ 25} \cdot 100 = 80 \cdot 100$$

Таким образом, проданное количество уменьшилось на 20% (100–80) между отчетным периодом и фактическим периодом.
Используя эти числа, мы также можем рассчитать индекс стоимости, при этом стоимость определяется как цена, умноженная на количество:

$$B/0 = \frac{Пн \cdot Qн}{P0 \cdot Q0} \text{ "=" } \frac{70 \cdot 20}{25} \cdot 100 = 112 \cdot 50 \cdot .$$

Стоимость рассматриваемого предмета имеет там fore увеличился на 12% (112 100) между двумя рассматриваемыми периодами.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Составной порядковый номер
- Индекс Фишера
- Индекс
- Индекс Ласпейреса
- Индекс Пааше

Простой линейный Регрессия

Простая линейная регрессия представляет собой анализ регрессии , в котором зависимая переменная Y линейно зависит от одной независимой переменной X.

Целью простой линейной регрессии является не только оценка функции регрессии относительно выбранной модели , но и проверка надежности полученных оценок.

ИСТОРИЯ

См. анализ регрессии.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Модель простой линейной регрессии имеет вид:

$$Y = \beta 0 + \beta 1X + \varepsilon ,$$

где Y — зависимая переменная (или объяснимая переменная), X — независимая переменная (или объясняющая переменная), ε — член случайной ненаблюдаемой ошибки, а β0 и β1 — оцениваемые параметры.
Если у нас есть набор из n наблюдений (X1, Y1), ..., (Xn, Yn), где Yi линейно зависит от соответствующего Xi, мы можем написать

$$Yi = \beta 0 + \beta 1Xi + \varepsilon i , \quad я = 1,...,n.$$

Задача регрессионного анализа состоит в том, чтобы оценить параметры β0 и β1, выбрав значения βˆ 0 и βˆ 1 так, чтобы расстояние между Yi и (β0+ β1 · Xi) было минимальным. Мы должны иметь:

$$\varepsilon i = Yi \quad \beta 0 \quad \beta 1 \cdot Xi$$

мало для всех i = 1,...,n. Для этого мы можем выбрать один из многих критериев: 1.

$$\min_{\beta 0,\beta 1} \max_{я} |\varepsilon i|$$

2. $$\min_{\beta 0,\beta 1} \sum_{i=1}^{n} |\varepsilon i|$$

3. $$\min_{\beta 0,\beta 1} \sum_{i=1}^{я} \varepsilon 2$$

C

Machine Translated by Google

Самый распространенный метод оценки параметрами β0 и β1 является третья, называемая методом наименьших квадратов. Он направлен на минимизацию суммы квадратов ошибок.

Оценка параметров методом наименьших квадратов дает следующие оценки β^ 0 и β^ 1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Xi \quad \bar{X}\hat{Y}i \quad \bar{Y}}{\sum_{i=1}^{n} Kси - \bar{X}^{2}}$$

β^ 0 знак равно Y¯ β^ 1X .

Затем мы можем написать:

$$\hat{Y}i = \hat{\beta}0 + \hat{\beta}1Xi ,$$

где Y^i — оценочное значение Yi для данного Xi , когда известны β^ 0 и β^ 1 .

Мера достоверности оценки Д

Мы рассчитали оценку стоимости Y , или Y^, с помощью метода наименьших квадратов, основанного на наших расчетах на линейной модели, чтобы преобразовать отношение, которое связывает Y до X. Но насколько мы можем доверять этой модели?

Для ответа на этот вопрос полезно провести дисперсионный анализ и проверить гипотезу о параметрах β0 и β1 линии регрессии. Чтобы провести эти тесты, мы должны сделать следующие предположения: • Для каждого значения X Y является случайной величиной, распределенной согласно нормальному распределению.

• Дисперсия Y одинакова для всех X ; оно равно σ2 (неизвестно). • Различные наблюдения над Y не независящие друг от друга, но обусловленные значениями X.

Дисперсионный анализ

Таблица дисперсионного анализа, которую мы должны построить, выглядит следующим образом:

Дисперсионный анализ

Источник вариация	Степень свободы дом	Сумма квадраты	Среднее значение квадраты
Регрессия	1	$\sum_{i=1}^{n} \hat{Y}i \quad \bar{Y}^{2}$	$\sum_{i=1}^{n} \hat{Y}i \quad \bar{Y}^{2}$
Рези двойной	п - 2	$\sum_{i=1}^{n} \hat{Y}i \quad \bar{Y}^{2}$	$\frac{\sum_{i=1}^{n} \hat{Y}i \quad \bar{Y}^{2}}{n \quad 2}$
Всего п - 1		$\sum_{i=1}^{n} Йи - \bar{Й}^{2}$	

Если модель правильная, то

$$S2 = \frac{\sum_{i=1}^{n} \hat{Y}i \quad \bar{Y}^{2}}{n - 2}$$

является несмещенной оценкой σ2.

Дисперсионный анализ позволяет проверить нулевую гипотезу:

$$H0 : \beta1 = 0$$

против альтернативной гипотезы:

$$H1 : \beta1 = 0$$

расчет статистики:

$$Ф = \frac{ЭМСЭ}{СКО} \text{ "=" } \frac{ЭМСЭ}{S2} .$$

Эту статистику F необходимо сравнить со значением Fα,1,n 2 таблицы Фишера, где α — значимость теста.

Если F Fα,1,n 2 , то принимаем H0.

Если F > Fα,1,n 2 , то мы отказываемся от H0 в пользу H1.

Коэффициент детерминации R2 рассчитывается следующим образом: β^ XY nY^ 2

$$P2 = \frac{ECC}{TCC} \text{ "=" } \frac{}{YY \quad nY\hat{} \ 2} ,$$

где ESS — сумма квадратов регрессии, а TSS — общая сумма квадратов.

Machine Translated by Google

Гипотеза о наклоне β1 В случае простой линейной регрессии статистика F позволяет проверить гипотезу , используя параметры уравнения регрессии или β1, наклон линии. Другой способ провести тот же тест заключается в следующем:

Если H0 истинно (β1 = 0), то статистика t:

$$\tau = \frac{\hat{\beta}}{1}$$

Sβˆ 1 следует распределению Стьюдента с (n 2) степенями свободы. –

стандартное отклонение βˆ1, оценка Sβˆ 1 ed по выборке:

$$S_{\hat{\beta}\,1} \text{ "=" } \frac{S2}{\sum_{\text{я}=1}^{\text{н}} \text{Кси} - \bar{X}^{2}}.$$

Статистику t необходимо сравнить со значением _ таблицы Стьюдента, где α – t α ,n 2 2 уровень значимости и n 2 количество степеней свободы. Правило принятия решения следующее:

|τ| $t_{\alpha \frac{1}{2}}$ n 2

принимаем H0 (β1 = 0).

|τ| > $t_{\alpha, \frac{n}{2}}$ 2

мы отказываемся от H0 в пользу H1 (β1 = 0).

Можно показать, что в случае простой линейной регрессии t 2 равняется Ф. Статистика t позволяет рассчитать достоверность интервал для β1.

Гипотеза об ординате в начале координат β0

Аналогичным образом можно построить доверительный интервал для β0 и проверить гипотезу:

H0 : β0 = 0

H1 : β0 = 0

Статистика τ:

$$\tau = \frac{\hat{\beta}\,0}{S\hat{\beta}\,0}$$

также следует распределению Стьюдента с (n 2) степенями свободы.

Расчетное стандартное отклонение βˆ 0, обозначенное как Sβˆ 0 , определяется:

$$S_{\hat{\beta}\,0} \text{ "=" } \frac{S2 \sum_{\text{я}=1}^{\text{н}} X2_{\text{я}}}{\text{н} \sum_{\text{я}=1}^{\text{н}} \text{Кси} - \bar{X}^{2}}.$$

Правило принятия решения следующее:

|τ| $t_{\alpha \frac{1}{2}}$ n 2

принимаем H0 (β0 = 0).

|τ| > $t_{\alpha, \frac{n}{2}}$ 2

мы отказываемся от H0 в пользу H1 (β0 = 0).

$\alpha_{\frac{2}{2}}$n 2 получается из таблицы Стьюдента для t (n 2) степеней свободы и уровень значимости α.

Коэффициент детерминации R2 рассчитывается следующим образом:

$$P2 = \frac{ECC}{TCC} \text{ "=" } \frac{\hat{\beta}\,XY \quad n\bar{Y}^{2}}{YY \quad n\bar{Y}^{2}},$$

где ESS — сумма квадратов регрессии, а TSS — общая сумма квадратов.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Простая линейная регрессия является частным случаем множественной линейной регрессии. Матричный подход, обнаруженный в множественной линейной регрессии (модель , содержащая много независимых переменных), справедлив и для частного случая, когда имеется только одна независимая переменная.

См. анализ регрессии.

ПРИМЕРЫ

В следующей таблице представлены валовой национальный продукт (ВНП) и спрос на

C

Machine Translated by Google

основных продуктов за период с 1969 по 1980 год в некоторых странах.

Год	ВНП	Спрос на основные продукты
	Икс	Д
1969 г.	50	6
1970 г.	52	8
1971 г.	55	9
1972 г.	59	10
1973 г.	57	8
1974 г.	58	10
1975 г.	62	12
1976 г.	65	9
1977 г.	68	11
1978 г.	69	10
1979 г.	70	11
1980 г.	72	14

Мы хотим оценить спрос на скобы в зависимости от ВНП по моду эль:

$$Yi = \beta 0 + \beta 1Xi + \varepsilon i, \qquad я = 1,...,n.$$

Оценка параметров β0 и β1 по _ метод наименьших квадратов дает нам следующие оценки:

$$\hat{\beta} 1 = \frac{\sum_{я=1}^{12} Xi - \bar{X}\ Yi - \bar{Y}}{\sum_{я=1}^{12} Кси - \bar{X}^{2}} = 0,226$$

$$\hat{\beta} 0 = \bar{Y} \quad \hat{\beta} 1\bar{X} = \quad 4,04 .$$

Расчетная линия может быть записана как:

$$\hat{Y} = -4,04 + 0,226X .$$

Дисперсионный анализ

Будем вычислять степени свободы как а также сумма квадратов и среднее значение квадраты, чтобы установить таблицу дисперсионный анализ:

$$dlreg = 1 \ ,$$

$$dlres = n \quad 2 = 10 \ ,$$

$$dltot = n - 1 \text{ знак равно } 11 ;$$

$$ЭСС = \sum_{я=1}^{12} Y\hat{}i \quad \bar{Y}^{2}$$

$$= (7,20 - 9,833) \quad 2 + \cdots$$

$$+ (12,23 - 9,833)^{2}$$

$$= 30,457;$$

$$RSS = \sum_{я=1}^{12} Йи - Йи^{2}$$

$$= (6 - 7,20) \quad 2 + \cdots$$

$$+ (14 - 12,23)^{2}$$

$$= 17,21 ;$$

$$TCC = \sum_{я=1}^{12} Йи - \bar{Й}^{2}$$

$$= (6 - 9,833) \quad 2 + \cdots$$

$$+ (14 - 9,833)^{2}$$

$$= 47,667;$$

$$EMSE = = \frac{ECC}{dlreg} = 30,457$$

$$S2 = СКО = = \frac{RSS}{длрес} = 1,721 .$$

Таким образом, мы имеем следующую таблицу:

Дисперсионный анализ

Источник вар иация	Степени бесплатно дом	Сумма квадраты	Среднее значение квадраты	Ф
Ре гресия	1	30,457	30,457	17,687
Рези двойной	10	17.211	1,721	
Общий	11	47,667		

Machine Translated by Google

При уровне значимости α = 5% находим в таблице Фишера значение

$$F_{\alpha,1,n-2} = F_{0,05,1,10} = 4,96 .$$

В результате $F > F_{0,05,1,10}$ мы отвергаем нулевую гипотезу $H_0: \beta_1 = 0$, что означает, что β_1 значимо отличается от нуля.

Для расчета коэффициента детерминации

Для этого примера достаточно взглянуть на таблицу дисперсионного анализа, так как она содержит все необходимые элементы:

$$R^2 = \frac{ECC}{TCC} = \frac{30.457}{47,667}$$

$$= 0,6390 = 63,90 \%$$

Можно сделать вывод, что согласно выбранной модели вариация спроса на основные продукты на 63,90% объясняется вариацией ВНП.

Очевидно, что значение R^2 не может превышать 100; значение 63,90 достаточно велико, но недостаточно близко к 100, чтобы отговорить от попыток улучшить модель.

Это может означать:

- Что кроме ВНП, другие переменные должны быть приняты во внимание для плавника. er определение функции спрос на основные продукты, поэтому ВНП объясняет лишь часть вариации.
- Мы должны протестировать другую модель (без постоянного члена, нелинейную модель и т.д.).

Проверка гипотез по параметрам позволит определить, отличаются ли они достоверно от нуля.

Гипотеза о наклоне β_1

В таблице дисперсионного анализа имеем:

$$S^2 = CKO = 1,721 ,$$

что позволяет рассчитать стандартное отклонение ация $S_{\hat{\beta}_1}$ $\hat{\beta}_1$:

$$S_{\hat{\beta}_1} = \sqrt{\frac{S^2}{\sum_{i=1}^{12} (Kси_i - \bar{X})^2}}$$

$$= \sqrt{\frac{1,721}{(50 - 61,416)^2 + \cdots + (72 - 61,416)^2}}$$

$$= \sqrt{\frac{1,721}{596,92}} = 0,054 .$$

Мы можем рассчитать статистику:

$$\tau = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{0,226}{0,054} = 4,19 .$$

Выбрав уровень значимости α = 5%, получим:

$$t_{\alpha/2,n-2} = t_{0,025,10} = 2,228 .$$

Сравнивая это значение t со значением в таблице, получаем :

$$|\tau| > t_{\alpha/2,n-2} ,$$

что указывает на то, что нулевая гипотеза:

$$H_0 : \beta_1 = 0$$

следует отклонить для альтернативной гипотезы:

$$H_1 : \beta_1 = 0 .$$

Делаем вывод, что существует линейное соотношение между ВНП и спросом на основные продукты. Если бы H_0 было принято, это означало бы, что нет линейной зависимости между этими две переменные.

Примечание: мы видим, что

$$F = \tau^2 ,$$

$$17,696 = 4,192$$

(без учета ошибок округлости). Следовательно, мы можем безразлично использовать тест t или тест F для проверки гипотезы $H_0: \beta_1 = 0$.

C

Machine Translated by Google

Гипотеза об ординате в происхождении

Стандартное отклонение $S_{\hat{\beta}_0}$ β_0 от $\hat{\beta}_0$:

$$S_{\hat{\beta}_0} = \frac{\sqrt{S^2 \frac{\sum_{i=1}^{12} X_i^2}{12}}}{\sqrt{\sum_{i=1}^{12} (X_i - \bar{X})^2}} = \frac{\sqrt{1{,}721 \frac{502^2 + \cdots + 722^2}{12}}}{\sqrt{(12 \cdot 596{,}92)}} = \frac{\sqrt{1{,}721 \cdot 45861}}{\sqrt{7163.04}} = 3{,}31 .$$

Мы можем рассчитать статистику:

$$t = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} = \frac{4{,}04}{3.31} = 1{,}22 .$$

Сравнивая это значение t со значением таблицы Стьюдента $t_{0,025,10} = 2{,}228$, получаем :

$$|t| < t_{\alpha/2, n} \quad 2 ,$$

что указывает на нулевую гипотезу:

$$H0 : \beta 0 = 0$$

следует принимать на уровне значимости $\alpha = 5\%$. Делаем вывод, что значение $\beta 0$ значимо не отличается от нуля, а линия проходит через начало координат.

Новая модель: прохождение регрессии

Источник

Новая модель следующая:

$$Yi = \beta 1 Xi + \varepsilon i , \qquad i = 1,\ldots, 12 .$$

Делая оценку параметра $\beta 1$ по методу наименьших квадратов, имеем:

$$\sum_{i=1}^{12} \varepsilon_i^2 = \sum_{i=1}^{12} (Yi - \beta 1 Xi)^2 .$$

Приравняв к нулю производную по $\beta 1$, получим:

$$\frac{\sum_{i=1}^{12} \varepsilon_i^2}{\beta 1} = 2 \sum_{i=1}^{12} Xi\,(Yi - \beta 1 Xi) = 0$$

или

$$\sum_{i=1}^{12} (XiYi) - \beta 1 \sum_{i=1}^{12} X2_i = 0 .$$

Значение $\beta 1$, которое удовлетворяет уравнению оценщика $\hat{\beta} 1$ для $\beta 1$, равно:

$$\hat{\beta} 1 = \frac{\sum_{i=1}^{12} (СиИ)}{\sum_{i=1}^{12} X2_i} = \frac{7382}{45861} = 0{,}161 .$$

Как следствие, новая линия регрессии описывается следующим образом:

$$\hat{Y}i = \hat{\beta} 1 Xi$$

$$\hat{Y}i = 0{,}161 Xi .$$

Таблица анализа дисперсии относительно этой новой модели выглядит следующим образом:

Дисперсионный анализ

Источник вариация	Степени свобода	Сумма квадраты	Среднее значение квадраты	Ф
Регрессия	1	1188,2	1188,2	660,11
Резидуальный	11	19,8	1,8	
Всего	12	1208.0		

Обратите внимание, что в модели без константы количество степеней свободы для полной вариации равно количеству наблюдений n. Число степеней свободы остаточной вариации равно n 1.

Machine Translated by Google

Таким образом, мы можем определить R2:

$$P2 = \frac{ECC}{TCC} \text{ "=" } \frac{1188,2}{1208.0} = 0,9836$$

$$= 98,36\% .$$

Мы нашли с первой моделью val

уе R2 = 63,90%. Можно сказать, что это новое модель без константынамного лучше

чем предыдущий.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Анализ остатков
- Коэффициент детерминации
- Коэффициент корреляции

Наименьших квадратов

- Множественная линейная регрессия
- Нормальные уравнения
- Регрессивный анализ
- Остаток

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Себер, GAF (1977) Линейная регрессия

Анализ. Уайли, Нью-Йорк

Простая случайная выборка

Простая случайная выборка – это выборка метод, при котором выбираются n единиц среди N единиц населения в таких таким образом, что каждый из C_n^N возможные образцы имеет одинаковую вероятность быть выбранным.

ИСТОРИЯ

См. выборку.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Чтобы получить простую случайную выборку размером n, сначала особи популяции пронумерованы от 1 до N; тогда n чисел нарисовано между 1 и N. Рисунки сделано в принципе без замены.

Также можно использовать таблицы случайных чисел . для получения простой случайной выборки. Цель простой случайной выборки состоит в том, чтобы обеспечить оценка без смещения среднего и _ дисперсии населения . _ Действительно, если обозначить через y1, y2,...,yN характеристики популяции, а через y1,..., yn соответствующие значения в простом случайном образца, получаем следующие результаты:

	Население:	Образец:
Всего T =	$\sum_{i=1}^{N} y_i$	$t = \sum_{i=1}^{n} y_i$
Среднее $\bar{Y}$ =	$\frac{T}{N}$	$\bar{y} = \frac{t}{n}$
Вариант	$\sigma^2 \text{ "=" } \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N - 1}$	$s2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Простая случайная выборка может быть выполнена с помощью или без замены. Если розыгрыш выполняется с замещением, то заселение всегда остается одним и тем же. Образец _ поэтому характеризуется серией независимых случайных величин , распределенных одинаково. Если генеральная совокупность достаточно велика, а размер выборки относительно невелика по отношению к населению, можно считать, что случайный переменные независимы , даже если рисунки выполняются без замены. Это тогда можно оценить некоторые характеристики населения , определяя их из этот образец. Простая случайная выборка – это основы теории выборки.

ПРИМЕРЫ

Рассмотрим популяцию из 658 особей. Для формирования простой случайной выборки мы

Machine Translated by Google

каждому человеку приписать трехзначное число: 001, 002,..., 658. Затем с помощью таблицы случайных чисел мы получим выборку из десять особей. Мы случайным образом выбираем первый цифра и направление чтения. Затем мы читаем первые три цифры, пока мы не получим десять различных чисел от 001 до 658 для определения наш образец.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Выборочное обследование
Оценка
Оценщик
Выборка
Выборочное распределение
Стратифицированная выборка
Систематическая выборка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кокран, рабочая группа: Методы отбора проб, 2-е место изд. Уайли, Нью-Йорк (1963)

Моделирование

Моделирование – это метод анализа, проектирование и эксплуатация сложных систем. Моделирование включает в себя разработку модели систему и проведение экспериментов над ней по мере продвижения.

Фундаментальная проблема моделирования состоит в том, в строительстве искусственного самца пропорционально статистически известному распределению. Эти дистрибутивы обычно известны эмпирически: они приводят к статистическому исследованию, из которого мы можем определить распределения вероятностей случайных величин, характеризующих явление.

Когда распределения известны, выборка строится путем генерации случайных чисел.

ИСТОРИЯ

Метод моделирования имеет долгую историю. Впервые он был введен Стьюдентом (1908 г.), кто открыл выборочное распределение t - статистика и коэффициент корреляции лация.

В последнее время благодаря компьютерам, методам моделирования и методам Монте-Карло быстро прогрессировали.

ПРИМЕРЫ

Рассмотрим феномен ожидания в касса большого магазина. Предполагать что мы хотим знать лучшую «систему кассиры», то есть там, где сумма издержки бездействия кассиров и издержки ожидающих клиентов являются самыми маленькими. Эволюция явления существенно зависит от распределения прихода клиентов на кассах и в торговом зале на время службы. Моделирование состоит в построении выборки прибытий клиентов и выборки времени обслуживания. Таким образом, мы можем рассчитать стоимость системы. Эксперимент можно повторить для разных систем и выбрать наилучшую систему.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Начальная загрузка
Генерация случайных чисел
Метод складного ножа
Метод Монте-Карло
Случайное число

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Госсет, С.В. «Студент»: вероятное Ошибка среднего. Биометрика 6, 1–25 (1908) стр. 302–310.

Кляйнен, JPC: Статистические методы в Моделирование (в двух частях). Том. 9 Стати-

Machine Translated by Google

тики: Учебники и монографии. Марсель Деккер, Нью-Йорк (1974)

Kleijnen, JPC: Статистические инструменты для специалистов по моделированию. Том. 76 Статис тики: Учебники и монографии. Марсель Деккер, Нью-Йорк (1987)

Нейлор, Т. Х., Балинтфи, Дж. Л., Бердик, Д. С., Чу, К.: Методы компьютерного моделирования. Уайли, Нью-Йорк (1967)

Снедекор, Джордж Уоддел

Снедекор, Джордж Уоддел родился в 1881 году. в Мемфисе и умер в 1974 году в Амхерсте. В 1899 году он поступил в Алабамский политехнический институт в Оберне и пробыл там некоторое время. 2 года. Хетен потратил 2 года на получение сертификата учителя. Когда его семья переехала в Таскалуса в 1903 году, Снедекор, Джордж Вад дель был переведен в Алабамский университет, где он получил в 1905 г. степень бакалавра математики и физики. Он принял свою первую академическую должность в Военной академии Сельмы, где преподавал с 1905 по 1907 год.

С 1907 по 1910 год он преподавал математику и греческий язык в Остин-колледже в Шер. мужчина.

В 1910 году Снедекор Джордж Ваддел переехал в Анн-Арбор, штат Мичиган, где он закончил свое образование в Мичиганском государственном университете в 1913 году. получив степень магистра. В то же самое года он был принят на работу в Университет Айовы в Эймсе, где он оставался до 1958 года.

В 1927 году Математический статистический Служба была открыта в Эймсе с Снедекор, Г.В. и Брандт, А.Е. у руля. Эта служба была предшественником Статистической лаборатории, открытой в 1933 г., где Снедекор, GW был директором, а где Кокс, Гертруда работала.

В 1931 году Снедекор, Г.В., был назначен профессором математического факультета Университета. Колледж штата Айова. В течение этого периода, Snedecor, GW подружились с Фишером, Рональдом Эйлмером. В 1947 г. Снедекор, Г.В. покинул Статистическую лабораторию. он стал профессором статистики. в Государственном колледже Айовы и остался в этом должность до выхода на пенсию в 1958 году. Он умер в феврале 1974 года.

Основная работа Снедекора, Джорджа Вада дел:

1937 г. Статистические методы, применяемые к экспериментам в области сельского хозяйства и биологии. Коллегиальный, Эймс, США.

Пространственные данные

Пространственные данные — это особый вид данных (часто называемые картографическими данными, географическими данными), которые относятся пространственной локализации, такой как, например, концентрация загрязняющих веществ в данной местности, облачность и т. д., и вся информация которую можно представить в виде карты. В пространственные данные включены данные с счетный набор пространственных объектов, для например, распределение детской смертности в разных городах и странах.

Мы также можем считать, что пространственные данные реализация региональной переменной. Теория региональной переменной предполагает, что любая мера может быть смоделирована (визуализирована) как будучи реализацией случайной функции (или случайный процесс). Например, пробы, взятые на земле, можно рассматривать как реализацию региональной переменной (см. геостатистику).

ИСТОРИЯ

С развитием территориальных информационных систем (ТИС) или географических информа-

Machine Translated by Google

систем (ГИС), понятие пространственных данных стали широко использоваться в статистике.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

ГИС — это набор информационных программ, которые позволяют на основе географических привязок интеграция пространственных, непространственных, качественных и количественных данных в базы данных, которые может быть организована как уникальная система. Пространственные данные поступают из самых разных областей и часто требуют подборки и анализа по определенным критериям для каждого типа исследовать. Спутники собирают огромное количество данных, а анализируется лишь малая их часть. Это богатство данных должно быть проверено по процессу отбора. Выбор данных может быть выполнено в соответствии с типом проблемы решить и модели, используемые для создания анализ.

ПРИМЕРЫ

Как правило, пространственные данные можно рассматривать как реализации случайных величин, таких как в обычный статистический анализ. Ниже мы представить, как пример пространственных данных, данные собраны в швейцарской Юре. Это набор из 259 Данные о загрязнении тяжелыми металлами в районе г. 14,5 км2, 10 из которых представлены здесь:

Икс	Д	Граунд Рок	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

Кр	Cu	ни	Pb	цинк
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

«Земля», которая относится к использованию земля и «камень», который описывает тип подстилающей породы в месте отбора проб, – категориальные данные (качественные категориальные переменные). Различные типы рок это:

1. Арговианский,
2. Кимеридж,
3. Секваниан,
4. портландский,
5. Четвертичный период.

Использование земли соответствует:

1. лес,
2. пастбища,
3. прерия,
4. пахал.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Классификация
- Геостатистика
- Гистограмма
- Выборка
- Пространственная статистика

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Аттея, О., Дюбуа, Ж.-П., Вебстер, Р.: Geo статистический анализ загрязнения почвы в швейцарский Юра. Окружающая среда. Загрязн. 86, 315–327 (1994)

Гувертс, П.: Геостатистика природных Оценка ресурсов. Оксфорд Универ Сити Пресс, Оксфорд (1997)

Machine Translated by Google

Матерон, Г.: Теория переменных
регионалы и другие приложения. Масссон, Париж
(1965)

Пространственная статистика

Пространственная статистика касается статистического анализа пространственных данных. Пространственная статистика охватывает все методы, используемые для изучения и доказать наличие пространственной зависимости между наблюдениями, распределенными в пространстве. Пространственные данные могут поступать из геологии, земли науки, обработка изображений, эпидемиология, сельское хозяйство, экология, астрономия или лес наук. Предполагается, что данные случайны, а иногда и их место предполагается быть случайным.

ИСТОРИЯ

Статистическая обработка данных пространственных данных вернемся к Галлею, Эдмонду (1686 г.). Галлей, Э. на одну карту наложен рельеф ветра направления и муссон между и вокруг тропиках и попытался объяснить причины. Модель пространственной статистики появилась намного позже. Фишер, Рональд Эйлмер (1920–1930), во время своих исследований экспериментального станции Ротамстед в Англии, сформулировал на основе принципов случайного выбора, анализа по блокам и репликации.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Статистический анализ форм пространственных данных основной предмет пространственной статистики. основными инструментами статистического анализа являются гистограмма, кривая накопления частот, график QQ, медиана , среднее распределения, коэффициенты вариации и мера асимметрии. в

В случае многих переменных наиболее часто используемые статистические подходы заключаются в представлении их в виде диаграммы рассеяния и изучить их корреляцию. Пространственный анализ имеет отличительные характеристика, учитывающая пространственную информацию и связь, существующую между два пространственных данных. Два данных, близких друг к другу, скорее будут похожи друг на друга, чем если бы они были дальше друг от друга. Таким образом, анализируя карту концентраций загрязняющих веществ на земле, мы увидим не случайные значения, а, наоборот, малые значения. сгруппированных загрязняющих веществ, а также больших значений, и переход между классами будет быть непрерывным.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Пространственное представление в виде карт а чертежи и пространственная статистика могут быть объединены, чтобы дать лучшее визуальное понимание влияния того, что соседние значения иметь друг на друге. Наложение статистических данных информация на карте позволяет лучше анализировать информацию и сделать соответствующие выводы.

ПРИМЕРЫ

См. пространственные данные.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Автокорреляция

Геостатистика

Пространственные данные

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кресси, NAC: Статистика пространственных данных.
Уайли, Нью-Йорк (1991)

Гувертс, П.: Геостатистика природных
Оценка ресурсов. Оксфорд Универ
Сити Пресс, Оксфорд (1997)

Исаакс, Э. Х., Шривастава, Р. М.: Введение в прикладную
геостатистику. Оксфорд
Университетское издательство, Оксфорд (1989)

Machine Translated by Google

Journel, A., Huijbregts, CJ: Статистика Mining Geo. Академик, Нью-Йорк (1981)

Матерон, Г.: Теория региональных переменных и ее приложений. Масссон, Париж (1965)

Оливер М.А., Вебстер Р.:Кригинг: метод интерполяции для геоинформационной системы. Междунар. Дж. Геогр. Инф. Сист. 4 (3), 313–332 (1990)

Ранг копейщика Коэффициент корреляции

Коэффициент ранговой корреляции Спирмена (Spearman ρ) является непараметрической корреляцией измерений. Он используется для определения отношение , существующее между двумя наборами данных.

ИСТОРИЯ

Спирмен, Чарльз был психологом. В 1904 году он впервые ввел звание коэффициент корреляции. Часто называемый ρ Спирмена, это один из старейших ранговых статистических показателей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть (X1, X2,...,Xn) и (Y1, Y2,...,Yn) — две выборки размера n. RXi обозначает ранг Xi по сравнению с другими значениями выборки X для i = 1, 2,...,n. RXi = 1, если Xi — наименьшее значение X, RXi = 2, если Xi — второе наименьшее значение, и т. д., пока RXi = n , если Xi — наибольшее значение X. Таким же образом RYi обозначает ранг Yi , fori = 1, 2,...,n. Коэффициент ранговой корреляции Спирмена, обычно обозначаемый ρ, определяется следующим образом:

$$p = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

где di = RXi RYi .

Если несколько наблюдений имеют точно такое же значение, этим наблюдениям будет присвоен средний ранг. Если средних рангов много, то лучше сделать поправку и рассчитать:

$$p = \frac{\text{Секс} + \text{Си} - \sum_{i=1}^{n} d_i^2}{2\,\text{скс} \cdot \text{сы}},$$

где

$$Sx = \frac{n(n^2 - 1) - \sum_{i=1}^{g} (t_i^3 - t_i)}{12},$$

где g количество групп со средними рангами и ti размер группы i для Xsample и

$$Си = \frac{n(n^2 - 1) - \sum_{j=1}^{h} (t_j^3 - t_j)}{12},$$

где h — количество групп со средними рангами, tj — размер группы j для выборки Y.

(Если нет средних рангов, наблюдения рассматриваются как группы размера 1, что означает, что g = h = n и ti = tj = 1 для i, j = 1, 2,...,n и Sx = Sy = n(n2 - 1)/12.)

Проверка гипотезы

Коэффициент ранговой корреляции Спирмена часто используется в качестве статистического теста для определения того, существует связь между двумя случайными переменные. Тест может быть двусторонним или односторонний тест. Гипотезы:

A: Двусторонний случай

H0: X и Y взаимно независимы.

H1: существует либо положительная, либо отрицательная корреляция между X и Y.

Machine Translated by Google

Существует положительная корреляция, когда большое значения X имеют тенденцию ассоциироваться с большими значениями Y и малыми значениями X при малых значениях Y. Имеется отрицательное корреляция, когда большие значения X имеют тенденцию быть связанными с малыми значениями Y и наоборот.

B: Односторонний случай

H0: X и Y взаимно независимы.

H1: существует положительная корреляция между X и Ю.

C: Односторонний случай

H0: X и Y взаимно независимы.

H1: существует отрицательная корреляция между X и Ю.

Правила принятия решений

Правила принятия решений различаются в зависимости от гипотезы . Вот почему существуют решающие правила A, B и C относительно предыдущего случаи.

Правило принятия решения A

Отклонить H0 на уровне значимости α, если

$$\rho > tn,1 \quad \alpha \quad \overline{2} \qquad \text{или } \rho < tn, a \quad \overline{2} \ ,$$

где t - критическое значение данного теста по таблице Спирмена.

Правило принятия решения Б

Отклонить H0 на уровне значимости α, если

$$\rho > tn,1 \quad \alpha \ .$$

Правило принятия решения C

Отклонить H0 на уровне значимости α, если

$$\rho < tn, a \ .$$

Замечание: Обозначение t не означает, что коэффициенты Спирмена связаны с теми Студент.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Коэффициент ранговой корреляции Спирмена равен используется в качестве проверки гипотезы для изучения зависимости между двумя случайными величинами. Это можно рассматривать как тест на независимость.

В качестве непараметрической корреляционной меры его также можно использовать с номинальным или порядковые данные.

Измерение корреляции между двумя переменные должны удовлетворять следующим требованиям:

1. Принимает значения от 1 до +1.
2. Существует положительная корреляция между X и Y , если значение коэффициента корреляции положительный; совершенная положительная корреляция соответствует значению +1.
3. Существует отрицательная корреляция между X и Y , если значение коэффициента корреляции коэффициент отрицательный; совершенная отрицательная корреляция соответствует значению -1.
4. Между X и Y отсутствует корреляция. когда коэффициент корреляции близок к нуль; можно также сказать, что X и Y не коррелированный.

Коэффициент ранговой корреляции Спирмена представляет следующие преимущества:

- Данные могут быть нечисловыми до тех пор, пока они могут быть классифицированы по определенным критериям.
- Легко рассчитать.
- Соответствующий статистический тест не сформулировать основную гипотезу на основе форма распределения населения, из которого берутся выборки ан.

В таблице Спирмена приведены теоретические значения коэффициента ранговой корреляции Спирмена при гипотезе о независимости двух случайных величин.

Вот пример таблицы Спирмена для n = 6, 7 и 8 и α = 0,05 и 0,025:

Machine Translated by Google

н	a = 005	a = 0025
6	0,7714	0,8286
7	0,6786	0,7450
8	0,6190	0,7143

Полную таблицу Спирмена можно найти в Глассер, Г.Дж. и Винтер, РФ (1961).

ПРИМЕРЫ

В этом примере восемь пар реальных близнецов берут тесты интеллекта. Цель состоит в том, чтобы увидеть, есть ли независимость между тестом первородного близнеца и тестом рожденного вторым.

Данные приведены в таблице ниже, самые высокие баллы, соответствующие лучшим Результаты.

Пара близнецов родилась 1-й	Си	Родился вторым	И
1	90	88	
2	75	79	
3	99	98	
4	60	66	
5	72	64	
6	83	83	
7	86	86	
8	92	95	

TheX классифицируются между собой и Y между собой, и вычисляется di . Это дает:

Пара двойняшки	Рожденный 1-й Си	RXi	Рожденный 2-й И	RYi	ди
1	90	6	88	6	0
2	75	3	79	3	0
3	99	8	98	8	0
4	60	1	66	2	1
5	72	2	64	1	1
6	83	4	83	4	0
7	86	5	86	5	0
8	92	7	95	7	0

Коэффициент ранговой корреляции Спирмена равен затем вычислено:

$$p = 1 - \frac{6 \cdot \sum_{i=1}^{8} d_i^2}{n(n^2 - 1)} = 1 - 8\,\frac{6 \cdot 2}{(8^2 - 1)} = 0{,}9762 .$$

Это показывает, что существует (почти идеальный) положительная корреляция между интеллектом тесты.

Предположим, что результаты для 7 и 8 равны изменилась. Тогда мы получим следующую таблицу:

Пара двойняшки	Рожденный 1-й Си	RXi	Рожденный 2-й И	RYi	ди
1	90	6,5	88	5,5	1
2	75	3	79	3	0
3	99	8	98	7,5	0,5
4	60	1	66	2	1
5	72	2	64	1	1
6	83	4,5	83	4	0,5
7	83	4,5	88	5,5	1
8	90	6,5	98	7,5	1

Поскольку имеются средние ранги, воспользуемся формулой

$$p = \frac{Cx + Cи - \sum_{i=1}^{8} d_i^2}{2\,\text{скс} \cdot \text{сы}}$$

для расчета ранговой корреляции Спирмена коэффициент. Сначала вычисляем Sx:

$$Sx = \frac{n\,n^2 - 1 - \sum_{i=1}^{r} t_i^3 \text{ти}}{12} = \frac{8\,82 - 1 - 23 - 2 + 23 - 2}{12} = 41 .$$

Machine Translated by Google

Затем Сай:

$$Си = \frac{п\,п2 - 1 - \sum_{j=1}^{час} \frac{Дж^3 - тж}{12}}{12} = \frac{8\,82 - 1 - 23 - 2 + 23 - 2}{12} = 41 .$$

Значение ρ становится следующим:

$$\frac{41 + 41 - (1 + 0 + \ldots + 1)}{41 \cdot 41}\,p = 2 \quad = \frac{76,5}{82} = 0,9329 .$$

Мы видим, что и в этом случае имеется почти совершенная положительная корреляция.

Теперь проведем проверку гипотезы:

H0: Существует независимость между тестами интеллекта пары близнецов.

H1: Существует положительная корреляция между тестами интеллекта.

Мы выбираем значимый уровень α = 0,05.

Поскольку мы находимся в случае B, H0 отбрасывается, если

$$p > t8,0,95 ,$$

где t8,0,95 — значение таблицы Спирмена, т. е. если

$$p > 0,6190 .$$

В обоих случаях (ρ = 0,9762 и ρ = 0,9329) H0 отбрасывается. Можно сделать вывод о наличии положительной корреляции между результатами тестов интеллекта пары близнецов.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Непараметрический тест
- Тест на независимость

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Глассер, Г.Дж., Винтер, РФ: Критические значения коэффициента ранговой корреляции для проверка гипотезы о независимости. Биометрика 48, 444–448 (1961).

Спирмен, К.: Доказательство и измерение связи между двумя вещами. Являюсь. Дж. Психол. 15, 72–101 (1904)

Стол Копейщика

Таблица Спирмена дает теоретические значения коэффициента ранговой корреляции Спирмена при гипотезе о независимости двух случайных величин.

ИСТОРИЯ

См. коэффициент ранговой корреляции Спирмена.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент ранговой корреляции Спирмена

Среднеквадратичное отклонение

C

Стандартное отклонение является мерой дис человек. Он соответствует положительному квадратному корню из дисперсии, где дисперсия представляет собой среднее квадратов отклонений каждого наблюдения по отношению к среднему значению набора наблюдений.

Обычно он обозначается σ, когда он относится к генеральной совокупности , и S , когда он относится к выборке.

На практике стандартное отклонение σ популяции будет оцениваться стандартным отклонением S выборки этой популяции .

Machine Translated by Google

ИСТОРИЯ

Термин стандартное отклонение тесно связан с работами двух английских математиков, Пирсона, Карла и Госсета, В. С. Действительно, во время конференции, которую он провел перед Лондонским королевским обществом в 1893 г., Пирсон К. впервые употребил этот термин. Он снова использовал его в своей статье, озаглавленной «Рассечение асимметричных частотных кривых» в 1894 году, и именно Пирсон К. ввел символ σ для обозначения стандартного отклонения. Госсет В.С. (Стьюдент) также работал над этими проблемами. Он объяснил, почему важно отличать S (стандартное отклонение относительно выборки) от σ (стандартное отклонение относительно генеральной совокупности).

В своей статье от марта 1908 г. Госсет В.С. определил стандартное отклонение выборки следующим образом:

$$C = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}.$$

Встал вопрос о том, следует ли делить это выражение на n или на n 1. Пирсон К. спросил мнение Госсета В.С., который ответил в письме от 13 марта 1927 г., что обе формулы могут быть найти в литературе. Тот, у кого n 1

Знаменатель дает среднее значение , которое не зависит от размера выборки и , следовательно, от генеральной совокупности. С другой стороны, при больших выборках разница между обеими формулами незначительна и расчеты проще, если использовать формулу со знаменателем n .

В целом, Госсет, WS указывает, что использование n 1, вероятно, более уместно. для небольших выборок, тогда как использование n предпочтительнее для больших выборок.

Обратите внимание, что открытие стандартного отклонения должно быть помещено в контекст теории оценки и проверки гипотез.

Кроме того, сначала изучение изменчивости привлекло внимание астрономов, потому что они интересовались открытиями, связанными с распространение ошибок.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим случайную величину X с ожидаемым значением E[X]. Дисперсия σ2 совокупности определяется формулой :

$$\sigma^2 = E(X \quad E[X])^2$$

или

$$\sigma^2 = E\,X2 \quad E[X]^2 .$$

Стандартное отклонение σ представляет собой положительный квадратный корень из σ2:

$$\sigma = \sqrt{E(X - E[X])^2}.$$

Для оценки стандартного отклонения σ населения , распределение населения должно быть известно, но на практике удобно оценивать σ из случайного выборка x1,...,xn по формуле:

$$C = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

где S — оценка σ, а $\bar{x}$ — среднее арифметическое выборки :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Machine Translated by Google

Стандартное отклонение также может быть вычислено используя следующую альтернативную формулу:

$$C = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Стандартное отклонение используется в качестве шкалы измерения в тестах гипотез и доверительные интервалы. Другие меры дисперсия , такая как диапазон или среднее отклонения также используются в описательной статистике, но в тестах они играют менее важную роль. гипотез. Можно убедиться, что:

$$\text{среднее отклонение} \quad S \frac{\text{диапазон}}{2} \cdot \sqrt{\frac{n}{n-1}}.$$

Если мы рассмотрим случайную величину , которая дает нормальное распределение среднего значения μ и стандартное отклонение σ, мы можем убедиться, что 68,26% наблюдений расположены в интервал μ ± σ. Кроме того, 95,44% наблюдения расположены в интервале μ ± 2σ, а 99,74% наблюдений приходится на интервал μ ± 3σ.

ПРИМЕРЫ

Пять студентов последовательно сдали два экзамены, на которых они получили следующие оценки:

Экзамен 1:

$$3{,}5\,4{,}5\,3{,}5\,4{,}5\,4 \quad \bar{x} = \frac{20}{5} = 4$$

Экзамен 2:

$$2{,}5\,5{,}5\,3{,}5\,4{,}5\,4 \quad \bar{x} = \frac{20}{5} = 4$$

Среднее арифметическое $\bar{x}$ этих двух наборов наблюдений идентичны . Тем не менее, рассеивание наблюдений вокруг значит не то же самое.

Чтобы рассчитать стандартное отклонение, мы сначала необходимо вычислить отклонения каждого наблюдение в отношении арифметики среднее , а затем возвести в квадрат эти отклонения:

Экзамен 1:

Оценка	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3,5	0,5	0,25
4	0,0	0,00
4,5	0,5	0,25
3,5	0,5	0,25
4,5	0,5	0,25
$\sum_{i=1}^{5}(x_i - \bar{x})^2$		1,00

Экзамен 2:

Оценка	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2,5	1,5	2,25
5,5	1,5	2,25
3,5	0,5	0,25
4,5	0,5	0,25
4	0,0	0,00
$\sum_{i=1}^{5}(x_i - \bar{x})^2$		5.00

Стандартное отклонение для каждого экзамена равно:

Экзамен 1:

$$C = \sqrt{\frac{\sum_{i=1}^{5}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{4}} = 0{,}5$$

Экзамен 2:

$$C = \sqrt{\frac{\sum_{i=1}^{5}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{5}{4}} = 1{,}118$$

Поскольку стандартное отклонение оценок в второй экзамен больше, оценки больше

C

Machine Translated by Google

разбросаны вокруг среднего арифметического , чем на первый экзамен. Таким образом, вариативность второго экзамена больше, чем у первого.

первый.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент вариации
- Среднее абсолютное отклонение
- Мера рассеивания
- Дисперсия

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Барнард, Г.А., Плакетт, Р.Л., Пирсон, С.: «Студент»: статистическая биография Уильям Сили Госсет: по произведениям Э. С. Пирсона. Оксфордский университет Пресса, Оксфорд (1990)

Пирсон, К.: Вклад в математическую теорию эволюции. И. В: Карл Ранние статистические статьи Пирсона. Издательство Кэмбриджского университета, Кембридж, стр. 1–40 (1948). Впервые опубликовано в 1894 г. как: О рассечении асимметричных частотных кривых. Филос. Транс. Рой. соц. Лонд. сер. А 185, 71–110

Госсет С.В. «Студент»: вероятная ошибка среднего. Биометрика 6, 1–25 (1908).

Стандартная ошибка

Стандартная ошибка равна квадратному корню из предполагаемая дисперсия статистики , что означает стандартное отклонение выборки _ распределение этой статистики.

ИСТОРИЯ

Понятие стандартной ошибки относится к Гаусс, К. Ф. (1816 г.), который, по-видимому, не знать концепцию выборочного распределения статистики , используемой для оценки стоимости параметр населения . _

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Стандартная ошибка среднего, или стандартная отклонение выборочного распределения _ среднее значение, обозначаемое $\sigma_{\bar{x}}$, вычисляется как функция стандартного отклонения населения и соответствующего размера конечная совокупность выборки :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}},$$

где σ — стандартное отклонение популяции, N — размер популяции,
n — размер выборки.
Если размер N популяции бесконечен, поправочный коэффициент:

$$\sqrt{\frac{N-n}{N}}$$

можно опустить, так как оно стремится к 1 когда N стремится к бесконечности.
Если популяция бесконечна или выборка неполная (с замещением), тогда стандартная ошибка среднего равна к стандартному отклонению населения разделить на корень квадратный из размера образец:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Если стандартное отклонение σ совокупности неизвестно , его можно оценить по стандартное отклонение S выборки , заданное формулой квадратный корень из дисперсии S2:

$$S2 = \frac{\sum_{j=1}^{N} (x_j - \bar{x})^2}{n-1}.$$

Те же формулы справедливы для стандартных ошибка доли , или стандартное отклонение выборочного распределения долей, обозначаемое σp:

$$\sigma_p = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}}$$

Machine Translated by Google

в случае конечной популяции и

$$\sigma_{\bar{p}} = \frac{\sigma}{\sqrt{n}}$$

в случае бесконечной популяции, где σ - стандартное отклонение популяции:

$$\sigma = \sqrt{p \cdot (1 - p)} = \sqrt{p \cdot q},$$

где p — вероятность того, что элемент обладает изучаемым признаком, а q = 1 — p — вероятность того, что он им не обладает.

Стандартная ошибка разницы между двух независимых величин есть квадратный корень из суммы квадратов стандартных ошибок обеих величин. Например, стандартная ошибка разницы между двумя средними равно:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

где σ_1^2 и σ_2^2 — соответствующие дисперсии двух сравниваемых бесконечных совокупностей , а n1 и n2 — соответствующие размеры двух выборок.

Если дисперсии двух совокупностей неизвестны, мы можем оценить их с помощью дисперсий, рассчитанных для двух выборок, что дает:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

где S_1^2 и S_2^2 — соответствующие дисперсии из двух образцов.

Если учесть, что различия двух популяции равны, мы можем оценить значение этих дисперсий с помощью объединенной дисперсии $S_п^2$, вычисленной как функция S_1^2 и S_2^2:

$$S_п^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2},$$

который дает:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = S_п \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Стандартная ошибка разницы между две пропорции вычисляются так же, как стандартная ошибка разницы между двумя средними значениями. Поэтому:

$$\sigma_{p1 - p2} = \sqrt{\sigma^2_{p1} + \sigma^2_{p2}} = \sqrt{\frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2}},$$

где p1 и p2 — пропорции для бесконечных совокупностей , рассчитанные на двух выборках размера n1 и n2.

ПРИМЕРЫ

(1) Средняя стандартная ошибка

Производитель хочет проверить точность нового станка для изготовления болтов диаметром 8 мм. эфир. На партии N = 10000 шт. исследуется выборка из n = 100 шт., стандартное отклонение S равно 1,2 мм.

Стандартная ошибка среднего равна :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N}}.$$

Поскольку стандартное отклонение σ совокупности неизвестно , мы можем оценить его, используя стандартное отклонение выборки S, что дает:

$$\sigma_{\bar{x}} = \frac{C}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N}}$$

$$\approx \frac{1,2}{10} \cdot \sqrt{\frac{10000 - 100}{10000}}$$

$$= 0,12 \cdot 0,995 = 0,119 .$$

Обратите внимание, что фактор

$$\sqrt{\frac{N - n}{N}} = 0,995$$

C

Machine Translated by Google

мало влияет на результат, так как размер населения достаточно велико .

(2) Стандартная ошибка пропорции

Кандидат проводит опрос по выборке 200 человек, чтобы узнать, будет ли у него больше, чем 50% (= π0) голосов, где π0 — предполагаемое значение параметра π (доля населения).

Стандартная ошибка пропорции σp равна равно:

$$\sigma_p = \sqrt{\frac{\pi_0 \cdot (1 - \pi_0)}{n}},$$

где n — объем выборки (в нашем примере 200). Можно считать, что население бесконечно, поэтому мы не учитываем учитывать поправочный коэффициент:

$$\sigma_p = \sqrt{\frac{0{,}5 \cdot (1 - 0{,}5)}{200}} = 0{,}0354 .$$

(3) Стандартная ошибка разницы между две величины

Страховая компания решила оборудовать свою кабинеты с микрокомпьютерами. Он хочет купить эти микрокомпьютеры из двух разных поставщиков, если нет существенной разницы в долговечности между двумя брендами. Он тестирует образец из 35 микрокомпьютеров марки 1 и 32 марки 2, отметив время что прошло до первой поломки. наблюдаемые данные следующие:

Марка 1: стандартное отклонение S1 = 57,916.

Марка 2: стандартное отклонение S2 = 57,247.

Дело 1:

Мы предполагаем, что дисперсии двух популяции равны. Объединенный стандарт отклонение равно:

$$S_p = \sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}}$$

$$\overset{"="}{=} \sqrt{\frac{34 \cdot 57{,}9162 + 31 \cdot 57{,}2472}{35 + 32 - 2}}$$

$$= 57{,}5979 .$$

Зная взвешенное стандартное отклонение, можно рассчитать значение стандартного отклонения выборочного распределения используя следующую формулу:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 57{,}5979 \cdot \sqrt{\frac{1}{35} + \frac{1}{32}} = 14{,}0875 .$$

Случай 2:

Если мы предполагаем, что дисперсии двух совокупностей неодинаковы, мы должны начать с вычисления стандартной ошибки среднего для каждый образец:

$$\sigma_{\bar{x}_1} = \frac{S_1}{\sqrt{n_1}} \overset{"="}{=} \frac{57{,}916}{\sqrt{35}} = 9{,}79 ,$$

$$\sigma_{\bar{x}_2} = \frac{S_2}{\sqrt{n_2}} \overset{"="}{=} \frac{57{,}247}{\sqrt{32}} = 10{,}12 ,$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}$$

$$= 9{,}792 + 10{,}122$$

$$= 14.08 .$$

Мы замечаем в этом примере, что два случая дают значения стандартной ошибки, которые почти идентичны. Это объясняется очень небольшая разница между стандартом отклонения двух образцов S1 и S2.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Выборочное распределение
- Среднеквадратичное отклонение

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Box, GEP, Hunter, WG, Hunter, JS:
Статистика для экспериментаторов. Введение в проектирование, анализ данных и Построение модели. Уайли, Нью-Йорк (1978)

Кокран, рабочая группа: Методы отбора проб, 2-е место изд. Уайли, Нью-Йорк (1963)

Gauss, CF: Bestimmung der Genauigkeit der Beobachtungen, том. 4, стр. 109–117. (1816 г.). В: Gauss, CF Werke (опубликовано в 1880 г.). Dieterichsche Universitäts Druckerei, Геттинген

Гаусс, CF: Работа Гаусса (1803–1826) по Теория наименьших квадратов. Троттер, ВЧ, транс. Статистические методы исследования Group, Технический отчет, № 5, Princeton University Press, Принстон, Нью-Джерси (1957)

Стандартизированные данные

Стандартизированные данные — это данные, из которых мы вычесть среднее значение наблюдений relat ed к случайной величине (связанной с данных), а затем разделить на стандартное отклонение наблюдений. Таким образом стандартизированы данные имеют среднее значение 0 и дисперсию 1.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Стандартизация ряда данных X с x_i , i = 1,...n, — это то же самое, что заменить член xi ряда на x_{s_i}:

$$x_{s_i} = \frac{x_i - \bar{x}_i}{s},$$

где xi и $\bar{s}$ — соответственно среднее значение и стандартное отклонение xi.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

У некоторых статистиков есть привычка dardizing объясняющие переменные в линейной регрессионной модели, чтобы упростить числовые трудности, возникающие при матричных вычислениях, и облегчить интерпретацию и сравнение коэффициентов регрессии. Применение метода наименьших квадратов использование стандартизированных данных дает результаты, эквивалентные результатам, полученным с использованием исходных данных.

Регрессия обычно моделируется

$$Y_i = \sum_{j=1}^{n} \beta_j X_{ij}, \quad i = 1,...,n,$$

где Xj — нестандартизированные данные. Если Xij стандартизировано (среднее 0 и дисперсия 1), βj называется стандартизированным коэффициентом регрессии и эквивалентен корреляции, если Y также стандартизирован. Та же терминология можно использовать, если Xj сводится к дисперсии, равной 1, но не обязательно к нулевому среднему значению. Коэффициент β0 поглощает разницу.

ПРИМЕРЫ

В следующем примере наблюдения относятся к 13 смесям цемента.Каждая смесь состоит из четырех ингредиентов, указанных в стол. Цель эксперимента – определить, как величины X1, X2, X3 , X4 из этих четырех ингредиентов влияет на количество тепла Y , выделяемого при затвердевании цемент.

Таблица: Тепло, выделяемое цементом

Смешивание- тур я	Ингредиент Икс1	Икс2	Икс3	Икс4	Нагревать Д
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

C

Machine Translated by Google

Смешивание-	Ингредиент				Нагревать
тур я	Ик1	Ик2	Ик3	Ик4	Д
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

Источник: Биркес и Додж (1993 г.).

yi количество тепла, полученное при затвердевании i-я смесь (в джоулях);

xi1 количество ингредиента 1 (алюмината трикальций) в i-й смеси;

xi2 количество ингредиента 2 (силиката трикальция) в i-й смеси;

xi3 количество ингредиента 3 (алюмоферрит тетракальция) в i-й смеси;

xi4 количество ингредиента 4 (силикат дикальций) в i-й смеси.

Полученная модель простой линейной регрессии:

$$Y = 62,4 + 1,55X1 + 0,51X2 + 0,102X3 - 0,144X4 + \varepsilon .$$

В следующей таблице представлены данные о цемента с данными относительно стандартизированного объясняющие переменные, которые мы обозначаем через Xs_1, Xs_2, Xs_3, Xs_4 и крестики

Данные по цементу со стандартизированными объясняющими переменными

Смешивание-	Стандартизированный ингредиент				Нагревать
тур я	Xs_1	Xs_2	Xs_3	Xs_4	Д
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
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

Оценка методом наименьших квадратов полной модели с использованием стандартизированного объясняющие переменные дают:

$$Y = 95,4 + 9,12Xs_1 + 7,94Xs_2 + 0,65Xs_3 - 2,41Xs_4 .$$

Отметим, что коэффициент $\beta s_1 = 9,12$ иксов $_1$ может вычислить путем умножения коэффициента $\beta 1 = 1,55$ от X1 в исходной модели по стандартное отклонение $s1 = 5,882$ X1. Также заметим, что среднее значение $\overline{X1} = 7,46$ для X1 рассматривается не в соотношении между βs_1 и $\beta 1$ но в соотношении между β_0^c и $\beta 0$. Взяв только первые два стандартизированных пояснительных переменных, получаем решение

$$Y = 95,4 + 8,64Xs_1 + 10,3Xs_2 .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Данные

Регрессия хребта

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Биркес Д., Додж Ю.: Альтернативные методы регрессии. Уайли, Нью-Йорк (1993)

Статистическое программное обеспечение представляет собой набор компьютерных программ и процедур для обработки статистические данные.

Machine Translated by Google

Библиотеки

Библиотека — это набор программ или подпрограмм, установленных под одним и тем же операционная система (DOS, Windows или Unix) и может использоваться путем ввода отдельных команд.

IMSL (Международная математика и Библиотека статистики): собрание около 540 подпрограмм, написанных на FORTRAN конкретно по математике и статистика.

NAG (группа численных алгоритмов): библиотека алгоритмов, написанная на помощь трех языков (АЛГОЛ 60, АЛГОЛ 68 и ФОРТРАН ANSI). Мы находим здесь более 600 продвинутых подпрограммы, касающиеся моделирования, регрессионный анализ, методы оптимизации, GLIM (общие линейные модели), GenStat, временные ряды и графические представления. Кроме того, мы можем лечить простым способом получены ответы по переписи.

Программное обеспечение

Программное обеспечение представляет собой набор сложных алгоритмов представление общей структуры данных и требует минимального предварительного программирования со стороны пользователя.

BMDP: набор подпрограмм, написанных ДЛЯ TRAN, которые позволяют обрабатывать запросы с помощью опроса, для создания графических представлений, для применения многомерных методов или непараметрических тестов, или выполнить линейную или нелинейную регрессионный анализ, не забывая изучение временных рядов.

IMSL (Международная математика и Библиотека статистики): набор подпрограмм написанные на FORTRAN специально для областей математики и статистика.

MINITAB: самая простая в использовании система. Поскольку его можно использовать в интерактивном режиме, позволяет проводить анализы различных типов: описательная статистика, дисперсионный анализ, регрессионный анализ, непараметрические тесты, генерация случайных чисел или изучение временной ряд, в том числе.

NAG: библиотека алгоритмов , написанная на помощью трех языков (АЛГОЛ 60, АЛГОЛ 68, ФОРТРАН ANSI). Расширенные подпрограммы можно найти относительно моделирования, регрессии анализ, методы оптимизации , временные ряды и графические представления. Кроме того, полученные ответы путем обследования легко лечится, и многофакторный анализ, а также непараметрические тесты могут быть применены к их.

P-STAT: интерактивная система, которая обрабатывает планы экспериментов, графические представления, непараметрические тесты, и анализ данных.

SAS (система статистического анализа): набор программное обеспечение, включая базовый язык SAS, а также специальные программы. Конечно, самый полный пакет, это позволяет анализировать данные из самых разных поля.

SPSS (Статистический пакет для социальных наук): один из наиболее часто используемые системы. Он предлагает интерактивные способы обработки больших баз данных . Кроме того, некоторые в программу включены алгоритмы графического представления .

Machine Translated by Google

BMDP: в основном для биомедицинских приложений, BMDP — это полное программное обеспечение. комплект, состоящий из подпрограмм, написанных на ФОРТРАН, позволяющий обрабатывать запросы сделанные переписью, для создания графического представления, применять многомерные методы и непараметрические тесты, или проводить линейный или нелинейный регрессионный анализ, не пренебрегая исследованиями временных рядов. Каждая подпрограмма на основе высококонкурентных численных алгоритмы.

DataDesk: интерактивное программное обеспечение для анализ статистической обработки данных. Это изображает инструменты графически, позволяя видеть отношения между данными, тенденциями, подгруппы и выбросы.

EViews: программное обеспечение, предназначенное главным образом для экономистов, EViews — предпочтительный инструмент для анализа, прогнозирования и эконометрического моделирования. EViews можно использовать для общего статистического анализа, оценки временных рядов, крупномасштабного моделирования, графики или простого управления данными.

Excel: позволяет пользователю представлять и анализировать данные с помощью основных статистических методы: описательная статистика, анализ дисперсии, регрессии и корреляции, и проверка гипотез.

Gauss: статистическая программа с ориентированным матрицы, Gaussis — мощный эконометрический и графический инструмент.

JMP: статистическое программное обеспечение, предназначенное, в частности, для экспериментальных планов, JMP также содержит исчерпывающий набор статистических кал инструменты.

Lisrel: используется в социальных науках и не только в частности, для факторного анализа и моделирование, Lisrel особенно полезен для моделирования представления скрытых переменные.

Maple: полная математическая среда, предназначенная для манипулирования алгебраические выражения и решение уравнений и интегралов; это также имеет очень мощные графические инструменты в два и три измерения. Пользователь также может программа в клен.

Mathematica: статистический пакет, который позволяет выполнять описательные одномерные и многофакторный статистический анализ, данные сглаживание, классическая проверка гипотез, оценка доверительного интервала и линейная и нелинейная регрессия.

Matlab: интерактивная среда, которая можно использовать для анализа научных и статистические данные. Основные объекты Матлаб - это матрицы. Пользователь может выполнять численный анализ, обработку знаки и обработка изображений. Пользователь может программировать функции, используя C и ФОРТРАН.

MIM: программа для графического моделирования дискретные и непрерывные данные. Среди семейства доступных графических программ в этом программном пакете, мы упоминаем логлинейные модели, графические модели Гаусса, стандартные модели Manova, и многие другие модели, полезные для многомерного анализа.

P-STAT: создан в 1960 году для лечения фармакологических данных, P-STAT в настоящее время используется исследователями в области социальных наук. Он позволяет проводить анализ переписей населения. Сила P-CTAT заключается в его макросах, и он очень полезен для статистических приложений, связанных с к бизнесу. Это позволяет обрабатывать большие данные устанавливает и обрабатывает экспериментальные проекты,

Machine Translated by Google

графические представления, непараметрические тесты и анализ данных.

Язык программирования R похож на S, Ris также бесплатное программное обеспечение, широко используемое в академических кругах. Он представляет собой реализацию, отличную от S, но многие S-команды работают в R. R содержит большое разнообразие статистических команд (линейных и нелинейных моделирование, выполнение классических тестов, анализ временных рядов, классификация, сегментация, . . .), а также графическое методы.

SAS (система статистического анализа): комплект программ, включая базовые язык SAS, а также специальные программы. Без сомнения, самый полный пакет, позволяет анализировать данные разного происхождения: научные, инженерные, медицины и бизнес-приложений. Это включает значительное количество статистических инструментов: дисперсионный анализ, регрессионный анализ, категориальный анализ данных, многофакторный анализ, анализ выживаемости, деревья решений, непараметрический анализ, и «интеллектуальный анализ данных».

S-plus: программное обеспечение, написанное на мощном и гибкий объектно-ориентированный язык программирования под названием S, он позволяет матричные, математические и статистические расчеты. Он включает в себя инструмент для исследовательский визуальный анализ, который также позволяет проводить расширенный анализ данных, поскольку а также их моделирование. S-плюс позволяет геостатистики, чтобы собрать блок «S+SpatialStats», содержащий различные инструменты пространственной статистики.

Spad-T: основан исключительно на статистических методах многомерного анализа. Таблицы непредвиденных обстоятельств обрабатываемых переменных созданный с помощью многофакторного методы уступают место графическим представлениям.

SPSS (Статистический пакет для социальных наук): впервые выпущен в 1968 году, SPSS на самом деле является одним из наиболее широко используемых программы статистиков в различных домены. Он предлагает интерактивные средства для обработки больших баз данных. Это содержит алгоритмы графического представления.

Stata: статистическое программное обеспечение, позволяющее проводить статистический анализ из линейных моделирование по обобщенной линейной модели, обработка бинарных данных, использование непараметрических методы, проводить многофакторный анализ и анализ временных рядов, и использовать методы исследования графических данных. Кроме того, программное обеспечение позволяет программировать пользовательские команды. Стата содержит, в том числе, статистические инструменты для эпидемиологов.

Statistica: впервые выпущен в 1993 г., Statisti са — чрезвычайно сложное программное обеспечение набор, который содержит мощное исследование инструменты для работы с большими базами данных. Базовый программное обеспечение включает базовую статистику, дисперсионный анализ, непараметрическую статистику, линейные и нелинейные модели, методы исследовательского многомерного анализа, нейронные сети, расчет выборки размеры, средства анализа экспериментальных планов, средства контроля качества и т. д.

StatView: статистическое программное обеспечение для меди кал ученых и исследователей. Статистический вид позволяет производить описательную статистику, выполнять параметрические и непараметрические проверку гипотез, расчет корреляций и ковариаций, проведение линейных, нелинейной и логистической регрессии, составлять таблицы непредвиденных обстоятельств и выполнять факторный анализ, анализ выживания,

Machine Translated by Google

контроль качества и анализ Парето.
Имеется хороший графический интерфейс.

Systat: мощное графическое программное обеспечение. Позволяет для визуализации данных, а также
использование базовой статистики для поиска адекватного модель.

Солас: программное обеспечение для обработки недостающих данных. Солас интегрирует алгоритмы множественного вменения.

Vista: предназначена для тех, кто начинает статистики, а также для профессоров. Виста может использоваться в качестве учебного программного обеспечения в область одномерных, многомерных, графическая и вычислительная статистика.

WinBugs: программа байесовских данных моделирование на основе методов моделирования методом Монте-Карло. Он включает представление байесовских моделей с команды, позволяющие выполнять моделирование. Кроме того, доступен интерфейс, позволяющий управлять анализом в любой момент. Графические инструменты позволяют один для контроля сходимости моделирование.

XploRe: обширная статистическая среда для исследовательского анализа данных. Он запрограммирован на языке, ориентированном на матрицу, поэтому содержит большую группу статистических операций и имеет мощный интерактивный интерфейс.
интерфейс. Программа поддерживает пользовательские макросы.

ИСТОРИЯ

Первые попытки статистиков
упростить свои статистические расчеты было построение статистических таблиц. С
появление компьютеров в середине 1940-х гг., кодирование статистических подпрограмм заняло место. Языки программирования в этом времени не были широко известны исследователям, и им пришлось ждать до начала
1960 год, когда язык программирования
Был выпущен FORTRAN для разработки статистических программ.
Среди первых подпрограмм, представленных IBM, наиболее известным является псевдослучайное число генератор РАНДУ. Систематическая разработка статистического программного обеспечения началась в 1960 г. с созданием программ БДМ и
SPSS, затем SAS.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Азимов Д.: Большой тур: инструмент для просмотра многомерных данных. SIAM J. Sci.
Стат. вычисл. 6 (1), 128–143 (1985)

Беккер, Р.А., Чемберс, Дж.М., Уилкс, А.Р.:
Новый язык S. Уолдсворт и
Брукс/Коул, Пасифик-Гроув, Калифорния (1988)

Диксон, главный редактор WJ: BMDP Statistical Программные коммуникации. Издательство
Калифорнийского университета. Беркли, Калифорния (1981)

Чемберс, Дж. М. и Хасти, Т. К. (ред.): Sta тистическими моделями в S. Waldsworth и
Брукс/Коул, Пасифик-Гроув, Калифорния (1992)

Додж Ю., Хэнд Д.: Что должно быть в будущем? Статистическое программное обеспечение вычисл. Стат. Анализ данных. (1991)

Додж, Ю., Уиттакер, Дж. (ред.): Вычислительная статистика, тома. 1 и 2. COMP STAT, Невшатель, Швейцария, 24–28.
Август 1992 г. (английский). Physica-Verlag, Гейдельберг (1992)

Эдвардс, Д.: Введение в графику Моделирование. Springer, Берлин Гейдельберг Нью-Йорк (1995)

Фрэнсис И.: Статистическое программное обеспечение: сравнительный обзор. Северная Голландия, Нью-Йорк (1981)

Machine Translated by Google

Хейс, А.Р.: Статистическое программное обеспечение: обзор и критика его разработки.
Управление военно-морских исследований, Арлингтон, Вирджиния (1982 г.)

Международный математический и статистический Библиотеки .: Информация о библиотеке IMSL. IMSLK, Хьюстон (1981)

Рипли, Б.Д., Венейблс, В.Н. (2000) S-программирование. Спрингер, Берлин Гейдельберг Нью-Йорк

Статистические науки: статистический анализ в S-Plus, версия 3.1. Сиэтл: StatSci, подразделение MathSoft (1993).

Венейблс, В. Н., Рипли, Б. Д.: Современная прикладная статистика с S-PLUS. Спрингер, Берлин Гейдельберг, Нью-Йорк (1994)

Вегман, Э.Дж., Хейс, А.Р.: Статистическое программное обеспечение. Энцикл. Стат. науч. 8, 667–674 (1988)

Статистическая таблица

Статистическая таблица дает значения функции распределения или индивидуальной вероятности случайной величины при определенном распределении вероятностей.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Необходимость статистической таблицы обусловлена тем, что определенная функция распределения не могут быть выражены в удобной для использования математической форме. В других случаях невозможно (или слишком сложно) найти первообраз функции плотности для непосредственного вычисления соответствующей функции распределения.

Значения функции распределения рассчитываются заранее и представляются в статистической таблице для упрощения использования.

ПРИМЕРЫ

Обычная таблица, таблица Стьюдента , таблицы Фишера и хи-квадрат являются примерами статистических таблиц.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Биномиальная таблица
- таблица хи-квадрат
- Функция распределения
- Таблица Фишера
- Таблица Крускала-Уоллиса
- Обычный стол
- Ученический стол
- Подписанная таблица Уилкоксона
- Таблица Уилкоксона

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кокоска, С., Цвиллингер, Д.: Стандартные таблицы вероятностей и статистики CRC и формулы. КПР Пресс (2000)

Статистика

Статистика — это результат применения функции к выборке данных, где сама функция не зависит от распределения выборки.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Статистика — это наблюдаемая случайная переменная способный, это отличает его от параметра. Полнота, достаточность и беспристрастность — важные желательные свойства статистики.

ПРИМЕРЫ

среднее арифметическое, дисперсия, стандартное отклонение.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Критерий независимости хи-квадрат
- Процентиль

Machine Translated by Google

p-значение
квантиль
Студенческий тест

Статистика

Слово статистика, происходящее от латыни, относится к понятию состояния (статуса): «Что относится к государству». правительства имеют острая потребность в подсчете и измерении многочисленных событий и действий, таких как изменение демография, рождаемость, тенденции иммиграции и эмиграции, изменения уровня занятости, предприятия и т. д.

С этой точки зрения термин «статистика» используется для обозначения набора доступных данных о данное явление (например, безработица статистика).

В более современном и более точном смысле от слова «статистика» считается дисциплиной, занимающейся числовыми данными.

Он состоит из набора технических приемов. получение знаний из неполных данных, от строгой научной системы управления сбором данных, их организацией, анализом и интерпретацией, когда это возможно представить их в числовом виде.

В популяции особей мы хотели бы знать, с точки зрения статистической теории, есть ли у данного человека машина и курит ли он. На с другой стороны, мы хотели бы знать, как многие люди имеют автомобиль и курят, и если существует связь между обладанием автомобиль и курение в исследуемой популяции.

Мы хотели бы знать черты населения в глобальном масштабе, не касаясь себя каждым человеком или каждым объектом в мире.

Население.

Мы различаем два подмножества методов:

(1) связанные с описательной статистикой и (2) связанные с выводной статистикой. Основной целью описательной статистики является представление информации в понятном и удобном для использования формате. логический статистика, с другой стороны, направлена на облегчение обобщения этой информации или, более конкретно, делать выводы (относительно популяций) на основе выборок эти популяции.

ИСТОРИЯ

Термин «статистика» происходит от латинского «статус» (состояние) был использован впервые, согласно Кендаллу, М.Г. (1960), историком по имени Гилини Джироламо в 1589 году.

Согласно Kendall, MG (1960), истинный Возникновение современной статистики восходит к 1660.

Статистические методы и статистическое распределение новации развились в основном в 19 в. века благодаря важной роли статистика в экспериментальных и гуманитарных науках энц.

В 20 веке статистика стала отдельной дисциплиной из-за богатства и многообразие используемых им методов. Это также объясняет рождение специализированных статистических энциклопедий, таких как Международная Опубликована Статистическая энциклопедия (1978 г.) в двух томах и Энциклопедии статистических наук (1982–1988 гг.) в девяти томах. умес.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Статистика используется во многих областях, все они очень отличаются друг от друга. Таким образом мы находим, что он используется в промышленном производстве в области окружающей среды и медицины, как так и в официальных агентствах.

Рассмотрим некоторые области, которые могут заинтересовать статистика:

Machine Translated by Google

- Фундаментальные исследования в статистике: исследования в области вероятностей, статистических методов и теории.
- Биология: фундаментальные исследования и эксперименты, связанные с основными явлениями живых организмов и биометрией.
- Торговля: управление данными, объем продаж, управление, методы инвентаризации, промышленное планирование, связь и теоретический контроль и учетные процедуры.
- Демография: изучение роста населения (рождаемость и смертность, миграционное движение), изучение структуры населения (личной, социальной и экономические характеристики).
- Экономика: мера объема производства, торговли, ресурсов, занятости и уровня жизни; анализ поведение потребителей и производителей, реакции рынка на изменение цен, Влияние рекламы и правительства политики.
- Образование: меры и тесты, связанные с процесс обучения.
- Инженерия: исследования и эксперименты включая тестирование дизайна и эффективности, улучшения в методах тестирования, вопросы связанные с выводом и тестами точности, усовершенствование методов контроля качества.
- Здоровье: стоимость несчастных случаев, медицинского обслуживания, госпитализация.
- Страхование: определение смертности и аварийность застрахованных лиц и население в целом; определение базовая ставка по каждой категории риска; применение статистических методов к управление страховыми полисами.
- Маркетинг и исследования потребления: проблемы, связанные с рынками, система дистрибуции, выявление трендов; изучение потребительских предпочтений и поведения.
- Медицина: эпидемиология, фундаментальная исследования и эксперименты над причинами, диагностика, лечение и профилактика болезни.
- Менеджмент и администрирование: проблемы, связанные с управлением персоналом, оборудования, методов производства и Условия труда.
- Психология и психометрия: проблемы связанные с оценкой исследования способности, интеллект, личные качества, нормальное и ненормальное поведение физического лица, а также установить шкалы измерений и инструментов меры.
- Социальные науки: создание технических Особенности отбора проб и теоретических испытаний о социальных системах и социальных благополучие. Анализ стоимости социальные науки; анализ культурных различий на ценностно-социальном уровне поведение.
- Космические исследования: интерпретация экспериментов и анализ данных, полученных в ходе космических полетов.
- Науки (в целом): фундаментальные исследования в области естественных и социальных наук энц.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Байесовская статистика

Описательная статистика

Генетическая статистика

Выведенный статистика

Официальная статистика

Пространственная статистика

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кендалл, MG: Где история
Статистика начинается? Биометрика 47, 447–449
(1960)

Коц, С., Джонсон, Н.Л.: Энциклопедия
StatisticalSciences, тома 1–9. Wiley, New
Йорк (1982–1988)

Крускал, WH, Танур, JM: Международный
Энциклопедия статистики, тт. 1–2.
Свободная пресса, Нью-Йорк (1978)

Стиглер, С.: История статистики,
Измерение неопределенности до
1900. Белкнап, Лондон (1986).

Стебель и лист

См. схему «стебель-лист».

Диаграмма стебля и листа

Схема «стебель-листья» представляет собой графическую
представление частот. Базовый
Идея состоит в том, чтобы предоставить информацию
о частотном распределении и сохранить значения
данных одновременно. Действительно, ствол
соответствует интервалам классов и
лист к количеству наблюдений в
класс, представленный разными данными. Это
тогда можно напрямую прочитать значения
данные.

ИСТОРИЯ

Происхождение диаграммы «стебель-листья»
часто ассоциируется с Tukey, JW (1977). Его
Концепция основана на гистограмме, которая
восходит к 18 веку.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Чтобы построить диаграмму «стебель-лист», каждый
номер должен быть вырезан на основную часть (стержень)
и второстепенная часть (лист). Например, 4,2,
42 или 420 можно разделить на 4 штока
и 2 для листа. Затем стебли перечисляются
ed по вертикали (одна строка на стержень) в порядке
величина и листья написаны рядом с
их тоже по порядку. Если набор
данные включают значения 4.2, 4.4, 4.8, 4.4,
и 4.1, то стебель и лист такие:

$$4\,|12448\ .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Диаграмма «стебель-лист» предоставляет
информацию о распределении, симметрии,
концентрации, пустых множествах и выбросах.
Иногда слишком много номеров с
тот самый ствол. В этом случае нельзя
поставить их всех на одну линию. Большинство
обычное решение заключается в разделении
десять листьев на две группы по пять листьев, например,
4 |0122444 и 4 |567889.
Можно сравнить два ряда данных
положив два листа слева и справа
того же ствола:

$$012244\,|4\,|567889\ .$$

ПРИМЕРЫ

ВНП на душу населения - 1970 (в долларах)

АФРИКА			
Алжир	324	Маврикий	233
Ангола	294	Марокко	224
Бенин	81	Мозамбик	228
Ботсвана	143	Нигер	90
Бурунди	67	Нигерия	145
Центральный Африканская республика	127	Воссоединение	769
Чад	74	Руанда	60
Коморы	102	Сенегал	236

Machine Translated by Google

АФРИКА			
Конго	238	Сьерра Леоне	167
Египет	217	Сомали	89
Экв. Гвинея	267	Южная Африка	758
Эфиопия	72	Родезия (Юг)	283
Габон	670	Судан	117
Гамбия	101	Свазиленд	270
Гана	257	Того	135
Гвинея	82	Тунис	281
Гвинея Бисау	259	Уганда	135
Кот-д'Ивуар	347	Объединенная респ.	
Кения	143	Камеруна	185
Лесото	74	Объединенная респ.	
Либерия	268	Танзании	100
Мадагаскар	133	Высоковольтный	59
Малави	74	Заир	88
Мали	54	Замбия	421
Мавритания	165		

Источник: Ежегодная статистика ООН, 1976 г.

В приведенной выше таблице указан ВНП на душу населения в долларов для большинства африканских стран в 1970 г.

Чтобы построить диаграмму ствола и листа из эти данные, они сначала классифицируются в порядке величины. Затем выбирается стебель: в этот пример мы будем использовать сотни. Десятки Затем номер помещается справа от соответствующая сотня, которая дает лист часть схемы. Единицы не берутся в учетную запись.

Стебель и лист выглядят следующим образом:

Раздел 1	1 = 110
0	5566777788889
1	00012333444668
2	12233355667889

Раздел 1	1 = 110
3	24
4	2
5	
6	7
7	56

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение частоты
- Графическое представление
- Гистограмма

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Tukey, JW: Объяснительный анализ данных, ограниченное предварительное издание. Эддисон Уэсли, Рединг, Массачусетс (1970–1971)

Стохастический процесс

Стохастический процесс — это семейство случайных величин. На практике он служит для моделирования большого количества временных явлений. где в игру вступает случай.

Мы различаем многие типы случайных процессов, используя определенные математические свойства. Наиболее известны:

- Марковский процесс (или цепь Маркова)
- Мартингейл
- Стационарный процесс
- Процесс независимых приращений

ИСТОРИЯ

Первоначально стохастические процессы благодаря достижениям начала 20 века в некоторые прикладные отрасли статистики, такие как статистическая механика (Гиббс, Дж. В., Больцман, Л., Пуанкаре, Х., Смолуховский, М., и Ланжевен, П.). Теоретические основы были сформулированы позже Дубом, Дж.Л. Колмогоров А.Н. и др. (1930–1940). В этот период слово «стохастический»,

Machine Translated by Google

от греческого слова «угадывать, верить, представить», начали использоваться. Другие достижения были сделаны, исходя из броуновского движения в физике (Эйнштейн А., Леви П., и Винер, Н.).

Марковский процесс был изобретен Марков, Андрей Андреевич (1856–1922), который заложил основы теории марковских процессов за конечное время, обобщенные русским математиком А. Колмогоровым (1936).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Случайный процесс — это семейство (Xt)t I (где I — дискретный или непрерывный набор) случайные величины, определенные в одном и том же вероятностном пространстве. Процессы в дискретном времени представлены следующими семействами (Xt)t (т.е. при I =) и процессы в непрерывное время семействами (Xt)t + (т.е. при I = = [0,)). +

Термин «случайный процесс» используется, если множество I бесконечно. В физике элементы I представляют время.

Для события w , принадлежащего фундаментальному множеству, определим траекторию исследуемого процесса как следующее семейство:

$$(Xt(w))t \quad I,$$

так что «ряд» (для процесса в дискретном времени) значений случайных величин принимается за выбранный и фиксированный элемент.

Основные объекты изучения случайных процессом являются свойства случайных величин и траекторий.

Мы характеризуем различные типы процессов относительно вероятностных условий. к которым они относятся.

1. Марковский процесс

   Марковский процесс — это случайный процесс (Xt)t в дискретном времени, для которого вероятность события (за время t + 1) зависит только от непосредственного прошлого (время т).

   Математически это выражается следующее условие:

   Для каждого t и для каждого i0, i1,..., it+1 + имеем:

$$P(Xt+1 = оно+1 \mid Xt = оно, \quad Xt \quad 1 = it \quad 1,...,X0 = i0) = P(Xt+1 = оно+1 \mid Xt = оно).$$

2. Мартингейл

   Мартингейл — это случайный процесс (Xt)t в дискретное время, что подтверждает следующее состояние:

   Для каждого t :

$$E(Xt+1 \mid Xt, Xt \quad 1,...,X0) = Xt.$$

3. Стационарный процесс

   Стационарный процесс (в строгом смысле термин) является случайным процессом(Xt)t для в которых совместные распределения процесса зависят только от разницы во времени (между рассматриваемыми переменными в совместное распределение), а не на «абсолютное» время. Математически:

   Для каждого t1, t2,...,tk I(k 2) и для каждого a I Xt1 , Xt2 ,...,Xtk имеют то же совместное распределение, что и Xt1+a, Xt2+a,...,Xtk+a.

4. Процесс с независимыми приращениями

   Процесс с независимыми приращениями случайный процесс (Xt)t I, для которого:

   Для каждого t1 < t2 < < tk ⋯ I: Xt2 Xt1 , Xt3 Xt2 ,...,Xtk Xtk 1 случайные переменные взаимно независимы.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Теоретически нет разницы между изучение случайных процессов и самой статистики. На практике при изучении стохастических процессов больше внимания уделяется структуре

Machine Translated by Google

и свойства моделей, чем на выводах из реальных данных, что, собственно, является областью статистики.

ПРИМЕРЫ 1.

Здесь мы приводим пример, показывающий тесную математическую связь между статистика и случайные процессы. Мы рассматриваем случайную величину Xi = возраст индивидуума i в популяции (потенциально бесконечной). Типичной задачей является оценка среднего значения E (X) при следующей гипотезе:

- Случайная величина Xi независима.
- Все Xi одинаково распределены.

2. Случайное блуждание. Пусть Zn, n 2, — независимые и одинаково распределенные случайные величины, принимающие значение +1 с вероятностью и значение 1 с вероятностью q = 1 p.

Определим стохастический процесс случайного блуждания следующим образом:

$$X0 = 0,\ Xt = \sum_{n=1}^{t} Zn \text{, если } t > 1 \cdot$$

Выборка случайного блуждания обычно представляется следующим образом:

(0, 1, 2, 1, 0, 1, 0, 1, 2, 3, 2,...).

Легко видеть, что заданный случайный процесс является марковским процессом (этот факт является прямым следствием независимости Xt) со следующими вероятностями:

$$P(Xt+1 = i + 1 \mid Xt = i) = p,\ P(Xt+1 = i \quad 1 \mid Xt = i) = q.$$

У нас есть это:

$$E(Xt+1 \mid Xt) = Xt + (+1) \cdot p + (\quad 1) \cdot q = Xt + (p \quad q).$$

Случайное блуждание является мартингалом тогда и только тогда, когда p = q = 0,5.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Конвергенция
- Случайная переменная
- Статистика
- Временная последовательность

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Барлетт, MS: Введение в стохастические процессы. Издательство Кембриджского университета, Кембридж (1960)

Биллингсли, П.: Статистический вывод для марковских процессов. Издательство Чикагского университета, Чикаго (1961)

Кокс, Д.Р., Миллер, HD: Theory of Stochastic Processes. Wiley, New York; Methuen, London (1965).

Дуб, Дж. Л.: Что такое стохастический процесс? Являюсь. Мат. Ежемесячно 49, 648–653 (1942)

Дуб, Дж. Л.: Стохастические процессы. Уайли, Нью-Йорк (1953)

Колмогоров А.Н.: Матем. Сборник. НС 1, 607–610 (1936)

Стратифицированная выборка

При стратифицированной выборке совокупность сначала делится на подсовокупности , называемые стратами. Они не должны проникать друг в друга, и совокупность этих слоев должна составлять всего населения.

Как только страты определены, из каждой страты берется случайная выборка (не обязательно одинакового размера), причем эта выборка делается независимо в разных слои.

ИСТОРИЯ

См. выборку.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Что касается определения числа индивидуумов, которые должны быть отобраны из каждой страты, то в основном два подхода: репрезентативная стратифицированная выборка и оптимальная стратифицированная выборка.

Репрезентативная стратифицированная выборка

В каждой страте число индивидуумов, пропорциональное величине страты, равно взятый.

Пусть Ni — величина страты i, N количество особей населения, и n размер глобальной выборки. Размер подвыборка ni определяется следующим образом:

$$ни = \frac{ни}{Н} \cdot n,\ i = 1, 2,...,k,$$

где k – количество слоев населения .

Оптимальная стратифицированная выборка

Цель состоит в том, чтобы определить размеры подвыборок, которые необходимо взять для получения наилучшей возможной оценки. Это делается путем минимизации дисперсия оценки m при _ _ ограничение

$$\sum_{я=1}^{к} ни = п\ ,$$

где m - среднее значение выборки , определенное к:

$$м = \frac{1}{Н} \sum_{я=1}^{к} Никси\bar{}i\ ,$$

где x¯i — среднее значение подвыборки i. Это классическая задача оптимизации и дает следующий результат:

$$ни = \frac{н}{\sum_{я=1}^{к} Ni \cdot \sigma i} Ni \cdot \sigma i\ ,$$

где σi — стандартное отклонение страты i.

При постоянном множителе $n/\sum_{я=1}^{к} Ni \cdot \sigma i$, размер подмножества пропорционален величине страты и стандартному отклонение.

Чем больше дисперсия изучаемой переменной в страте и чем больше размер Ni, тем больше размер подвыборки ni.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Стратифицированная выборка основана на идее получение в результате контроля определенных переменные, образец , воссоздающий изображение что максимально приближено к населению. Разброс населения может быть сильный; в этом случае для поддержания выборки погрешность на приемлемом уровне, очень большая должна быть создана выборка , которая может увеличить стоимость значительно. Чтобы избежать этой проблемы, хорошей альтернативой может быть стратифицированная выборка. Население можно разделить на страты таким образом, что вариация внутри слои относительно низкие, что уменьшит требуемый размер выборки для данной выборки ошибка.

Стратифицированная выборка может быть оправдана для следующие причины:

1. Если население слишком неоднородно, оно желательно иметь более однородный группы, чтобы получить выборку , более репрезентативную для населения.
2. Если кто-то хочет получить информацию о конкретных аспектах населения (например, с по отношению к разным государствам), для этого лучше подходит стратифицированная выборка.

Управляющий символ, используемый для определения слоев должны соблюдать следующие правила:

- Он должен находиться в тесной взаимосвязи с изучаемые переменные.

Machine Translated by Google

• Он должен иметь известное значение для каждой единицы населения .

Некоторые из наиболее часто используемых элементов управления переменными являются пол, возраст, регион, социально-профессиональный класс и др.

ПРИМЕРЫ

Рассмотрим совокупность 600 студентов, разделенных на 4 слоя. Мы заинтересованы в результаты теста с баллами от 0 до 100.

Предположим, у нас есть следующая информация:

Размер ни	Стандартное отклонение σя
Слой 1 285	8,6
Пласт 2 150	5.2
Слой 3 90	2.2
Слой 4 75	1,4

Если мы хотим взять стратифицированную выборку из 30 лиц, получаем следующие подвыборки:

1. Репрезентативная стратифицированная выборка

$$ни = \frac{ни}{Н} \cdot н$$

Затем мы получаем следующие значения, округляется до ближайшего целого числа:

$$n1 = \frac{N1}{№} \cdot н$$

$$"=" \frac{285}{600} \cdot 30$$

$$= 14{,}25$$

14

n2 = 7,5 7

n3 = 4,5 5

n4 = 3,75 4

2. Оптимальная стратифицированная выборка

$$ни = \frac{н}{к} \cdot Ni \cdot σi$$

$$\sum_{я=1} Ni \cdot σi$$

У нас есть:

N1 · σ1 = 2451

N2 · σ2 = 780

N3 · σ3 = 198

N4 · σ4 = 105

4

Постоянный фактор n/ $\sum_{я=1} Ni \cdot σi$ равно к:

$$\frac{н}{к} "=" \frac{30}{3534} = 0{,}0085 .$$

$$\sum_{я=1} Ni \cdot σi$$

Получим следующие значения, округлив до ближайшего целого числа:

n1 = 0,0085 · 2451 = 20,81 21

n2 = 0,0085 · 780 = 6,62 6

n3 = 0,0085 · 198 = 1,68 2

n4 = 0,0085 · 105 = 0,89 1 .

Если мы сравним оба метода, то увидим, что при более однородной страте второй метод берет большее количество особей из первой страты (наиболее рассеянной) и меньшее количество лиц в второй пласт.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Оценка
- Оценщик
- Выборка
- Простая случайная выборка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Йейтс, Ф.: Методы выборки для переписей и опросы. Гриффин, Лондон (1949)

Студент

См . Сили, Госсет Уильям.

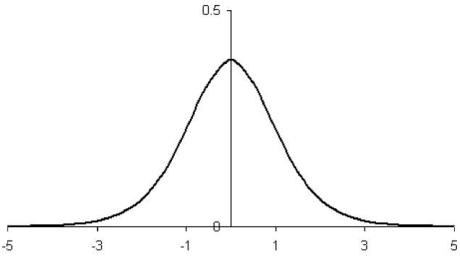
Machine Translated by Google

Распределение студентов

Случайная величина T следует за студентом распределения, если его функция плотности имеет вид:

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\cdot\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}},$$

где – гамма-функция (см. гамма-распределение) и v – число степеней свободы.



Распределение Стьюдента, ν = 7

Распределение Стьюдента является непрерывным распределение вероятностей.

ИСТОРИЯ

Распределение Стьюдента было создано Госсетом У. С., известным как «Студент», который в 1908 г. опубликовал статью, в которой он описал функцию плотности разности между средним значением выборки и средним значением генеральной совокупности , из которой была взята выборка, деленной на стандартную отклонение образца. В той же статье он привел и первую таблицу соответствующей функции распределения. Студент продолжил свои исследования и в 1925 г. опубликовал статью, в которой предложил новые таблицы, более обширные и точнее.

Фишер, Р.А. интересовался работами Госсета. Он написал ему в 1912 году, чтобы предложить геометрическую демонстрацию распределения Стьюдента и ввести понятие степени свободы. В 1925 г. он опубликовал статью, в которой определяет отношение t:

$$t = \frac{Z}{\sqrt{\frac{\text{Икс}}{v}}},$$

где X и Z — две независимые случайные величины, X распределено согласно распределению хи-квадрат с v степенями свободы, а Z распределено согласно центрированному и редуцированному нормальному распред Это отношение называется t Стьюдента.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если X является случайной величиной , которая следует распределение хи-квадрат с v степенями свободы, а Z — случайная величина, распределенная согласно стандартному нормальному распределению, то случайная величина

$$T = \frac{Z}{\sqrt{\frac{\text{Икс}}{v}}}$$

следует распределению Стьюдента с v степенями свободы, если X и Z независимы.
Рассмотрим случайную величину T:

$$T = \frac{\sqrt{n}(\bar{x} - \mu)}{C},$$

где $\bar{x}$ — среднее значение выборки , μ — среднее значение генеральной совокупности , из которой была взята выборка, а S — стандартное отклонение выборки, определяемое формулой:

$$C = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Тогда T следует распределению Стьюдента с n 1 степень свободы.

Machine Translated by Google

Ожидаемое значение случайной величины T
следующее распределение Стьюдента определяется как:

$$\text{Э}[T] = 0 \quad , \quad v > 1 \quad ,$$

а дисперсия равна:

$$\text{Вар}(T) = v - \frac{в}{2} \quad , \quad v > 2 \quad ,$$

где v — число степеней свободного
дом.

Распределение Стьюдента связано с другими
непрерывные распределения вероятностей:

- Когда число степеней свободы достаточно велико,
  распределение приближается к нормальному
  распределению.
- Когда число степеней свободы
  v равно 1, распределение Стьюдента равно
  идентично распределению Коши с
  α = 0 и θ = 1.
- Если случайная величина X следует за студентом
  распределение с v степенями свободы,
  то случайная величина X2 следует
  распределению Фишера с 1 и v степенями
  свобода.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Распределение Стьюдента является симметричным
распределения со значением, равным 0. Чем
больше число степеней свободы, тем
меньше становится его стандартное отклонение .
Распределение Стьюдента используется в
логической статистике в отношении t- отношения Stu
вмятину в проверке гипотез и в построении
доверительных интервалов для
среднее значение населения .
В тестах дисперсионного анализа можно
использовать распределение Стьюдента, когда сумма
квадратов между группами (или сумма
квадраты факторов) для сравнения только
имеет одну степень свободы.

Нормальное распределение можно использовать как
аппроксимация распределения Стьюдента
когда количество наблюдений n велико
(n > 30 по данным большинства авторов).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Распределение Коши
- Непрерывное распределение вероятностей
- Распределение Фишера
- Нормальное распределение
- Ученический стол
- Студенческий тест

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Приложения «Студенческих»
распределение. Метрон 5, 90–104 (1925)

Госсет С.В. «Студент»: вероятная ошибка
среднего. Биометрика 6, 1–25 (1908).

Госсет, Ю.В. «Студент»: Новые таблицы для
проверка значимости наблюдений.
Метрон 5, 105–120 (1925)

Студенческий стол

В таблице Student приведены значения
функция распределения случайной величины
в состоянии после распределения Стьюдента.

ИСТОРИЯ

Госсет, Уильям Сили, известный как «Студент»,
разработал в 1908 году тест Стьюдента.
и стол с его именем.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть случайная величина T следует за Стьюдентом
распределение с v степенями свободы.
функция плотности случайной величины T есть
предоставлено:

$$\phi(c) = \frac{\frac{v+1}{2}}{\sqrt{v\pi} \cdot \frac{в}{2}} \left(1 + \frac{c2}{в}\right)^{(v+1)/2} ,$$

C

Machine Translated by Google

где представляет собой гамма-функцию (см. гамма-распределение).

Функция распределения случайных величин способный T определяется:

$$F(t) = P(T \quad t) = \int_{\quad}^{t} f(s)\,ds.$$

Для каждого значения v таблица Стьюдента дает значение функции распределения F (t).

Поскольку распределение Стьюдента симметрично относительно начала координат, мы имеем:

$$F(t) = 1 \quad F(\quad t),$$

что позволяет определить F (t) для каждого отрицательного значения t.

Таблица Студента используется в обратном смысле: найти значения, часто соответствующие заданной вероятности. В этом случае таблица Стьюдента дает критическое значение для числа степеней свободы v и уровня значимости α. Мы обычно обозначаем через tv,α значение случайной величины T , для которого:

$$PT \quad t_{v,\alpha} = 1 \quad \alpha .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

См. Студенческий тест.

ПРИМЕРЫ

См. Приложение E.

Рассмотрим односторонний критерий Стьюдента на выборка размерности n . При v = n 1 степени свободы таблица Стьюдента позволяет определить критическое значение tv,α теста для заданного уровня значимости α.

Пример использования таблицы Student см. в разделе Тест Student.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Статистическая таблица
- Студенческое распределение
- Студенческий тест

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Применение «студенческого» распределения. Метрон 5, 90–104 (1925)

Госсет, С. В. «Студент»: вероятная ошибка среднего. Биометрика 6, 1–25 (1908).

Госсет, С.В. «Студент»: Новые таблицы для проверки значимости наблюдений. Метрон 5, 105–120 (1925)

Студенческий тест

Критерий Стьюдента представляет собой параметрическую проверку гипотезы о среднем значении выборки или о сравнении средних значений двух выборок.

ИСТОРИЯ

См. в истории студенческого стола.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X — случайная величина , распределенная по нормальному распределению. Случайная величина T , определенная выше, следует за Stu распределение вмятин с n 1 степенями свободы:

$$T = \frac{\bar{X} \quad \mu}{\bar{C}^{\bar{H}}},$$

где $\bar{X}$ — среднее арифметическое выборки , μ — среднее значение генеральной совокупности, S — стандартное отклонение выборки, n — размер выборки.

Тест Стьюдента состоит в вычислении этого отношения для выборочных данных и сравнении его с теоретическим значением Стьюдента . стол.

A. Тест Стьюдента для одной или одной выборки

1. Гипотезы, которые мы хотим проверить: – Нулевая гипотеза:

$$H0 : \mu = \mu 0 .$$

Machine Translated by Google

Мы должны проверить, происходит ли выборка из популяции с заданным средним значением μ0, или существует ли статистически значимая разница между наблюдаемым средним значением в выборке и гипотетическим средним значением популяции (μ0) при H0.

– Альтернативная гипотеза:

Альтернативная гипотеза может принимать три формы:

H1 : μ>μ0 (односторонний критерий справа)

H1 : μ<μ0 (односторонний критерий слева)

H1 : μ = μ0 (двусторонний критерий) .

2. Расчет t-статистики как:

$$t = \frac{\bar{x} \quad \mu 0}{\frac{C}{=}}\ ,$$

где μ0 — среднее значение генеральной совокупности, заданной H0, $\bar{x}$ — среднее значение выборки, S — стандартное отклонение выборки, а n — размер выборки.

3. Выбрать уровень значимости α теста. 4. Сравнить рассчитанное значение t с соответствующим критическим значением tn 1 (значение из таблицы Стьюдента с n 1 степенями свободы). RejectH0 , если абсолютное значение t больше критического значения tn 1.

Критические значения для разных степеней свободы и разных уровней значимости приведены в таблице Стьюдента.

Для одностороннего теста возьмем значение tn 1,1 α в таблице, а для двустороннего теста возьмем значение tn 1,1 α $\frac{}{2}$.

B. Тест Стьюдента для двух выборок

1. Гипотезы, которые мы хотим проверить:

– Нулевая гипотеза:

H0 : μ1 = μ2 .

В тесте, предназначенном для сравнения двух выборок, мы хотим знать, существует ли значительная разница между средним значением двух совокупностей, из которых были взяты выборки.

– Альтернативная гипотеза: она может принимать следующие три формы:

H1 : μ1 > μ2 (односторонний критерий справа)

H1 : μ1 < μ2 (односторонний критерий слева)

H1 : μ1 = μ2 (двусторонний критерий) .

2. Вычислить t-статистику по формуле:

$$\text{т} = \frac{\bar{x}1 \quad \bar{x}2}{\text{Сп} \cdot \frac{}{1\ \text{п1}}\ 1 + \frac{}{\text{п2}}}\ ,$$

где n1 и n2 — соответствующие размеры выборки, а Sp — сводное стандартное отклонение:

$$\text{Сп} = \frac{(\text{п1} \quad 1)\ \text{S2} + (\text{n2} \quad 1)\ \text{S2}}{\text{п1} + \text{п2} - 2}\ \frac{}{2}$$

$$\frac{2}{\text{ни}}$$

$$\text{"="} \quad \frac{\text{я=1}\ \text{Дж=1}\ \text{xij} \quad \bar{x}\text{i}}{\text{п1} + \text{п2} - 2}\ ^{2}\ ,$$

где S2 1 представляет собой стандартное отклонение выборки 1, S2 — выборку 2, а xij — наблюдение j выборки i.

3. Выбрать уровень значимости α теста 4. Сравнить рассчитанное значение t с критическим значением tn1+n2 2,1 α (значение из таблицы Стьюдента с (n1 + n2 2) степенями свободы) . Отклонить H0 , если абсолютное значение t больше критического значения.

Если тест односторонний, то взять значение tn1+n2 2,1 α из таблицы Стьюдента, а если двусторонний, то взять значение tn1+n2 2,1 α $\frac{}{2}$.

C

Machine Translated by Google

Тест значимости коэффициента регрессии.

Тест Стьюдента также можно применять к задачам регрессии. Рассмотрим простую модель линейной регрессии , для которой данные состоят из n пар наблюдений (xi, yi), i = 1,...,n, смоделированных с помощью

$$yi = \beta 0 + \beta 1 \cdot xi + \varepsilon i ,$$

где εi — случайные независимые и нормально распределенные ошибки со средним значением 0 и дисперсией σ2.

Оценки β0 и β1 задаются следующим образом:

$$\hat{\beta} 0 = \bar{y} \quad \hat{\beta} 1 \cdot \bar{x} ,$$

$$\hat{\beta} 1 = \frac{\sum_{i=1}^{n} (xi - \bar{x})(yi - \bar{y})}{\sum_{i=1}^{n} (xi - \bar{x})^2} .$$

Оценка $\hat{\beta} 1$ нормально распределена со средним значением β1 и дисперсией

$$Var(\hat{\beta} 1) = \frac{\sigma 2}{\sum_{i=1}^{n} (xi - \bar{x})^2} .$$

Если σ2 оценить средним квадратом невязок S2 и заменить в выражении дисперсией $\hat{\beta} 1$, то отношение $\hat{\beta} 1 \quad \beta 1$

$$T = \frac{}{\text{Var } \hat{\beta} 1}$$

распределяется по распределению Стьюдента с n 2 степенями свободы.

При определенном уровне значимости α мы можем провести тест Стьюдента для коэффициента β1. Гипотезы следующие:

Нулевая гипотеза H0 : β1 = 0

Альтернативная гипотеза H1 : β1 = 0 .

Если абсолютное значение рассчитанного отношения T больше значения t в таблице Стьюдента $t_{n \quad 2,1 \quad \alpha/2}$, тогда мы можем сделать вывод коэффициент β1 значимо отличен от нуля. В противном случае мы не можем отвергнуть нулевую гипотезу (β1 = 0).

В этом случае мы заключаем, что наклон линии регрессии достоверно не отличается от нуля.

Аналогичный тест можно провести для константы β0:

Оценка $\hat{\beta} 0$ нормально распределена со средним значением β0 и дисперсией

$$Var(\hat{\beta} 0) = \sigma 2 \cdot \frac{\sum_{i=1}^{n} x2_i}{\sum_{i=1}^{n} (xi - \bar{x})^2} .$$

Если σ2 оценить средним квадратом невязок S2 и заменить выражением Var $\hat{\beta} 0$, тогда отношение

$$T = \frac{\hat{\beta} 0 \quad \beta 0}{\text{Var } \hat{\beta} 0}$$

распределяется по распределению Стьюдента с n 2 степенями свободы.

Тест Стьюдента касается следующих гипотез:

Нулевая гипотеза H0 : β0 = 0

Альтернативная гипотеза H1 : β0 = 0 .

Если абсолютное значение рассчитанного отношения T больше значения t в таблице Стьюдента $t_{n \quad 2,1 \quad \alpha/2}$, тогда мы можем сделать вывод коэффициент β0 значимо отличен от нуля. В противном случае мы не можем отвергнуть нулевую гипотезу (β0 = 0).

Мы заключаем, что регрессия достоверно не отличается от нуля.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Как правило, мы должны быть очень внимательны к выбор теста для выполнения. Тест Стьюдента часто применяется ошибочно. На самом деле, это что-то значит, только если наблюдения происходят из нормально распределенной популяции или близки к нормально распределенной один. Если это требование не выполняется, то критическое значение критерия Стьюдента не дают стопроцентную гарантию.

Если применимо нормальное распределение, Тест Стьюдента на образце используется, когда стандартное отклонение населения неизвестно, и когда размерность выборка невелика (n < 30).

Аналогично, при наличии двух выборок Тест Стьюдента используется, когда стандартный двух популяций неизвестны и предполагается равным и когда сумма размеров выборок относительно невелика (n1 + n2 < 30).

ПРИМЕРЫ

Студенческий тест на образце

Рассмотрим выборку размерности n = 10 где наблюдения нормально распределены и таковы:

47 51 48 49 48 52 47 49 46 47 .

Мы хотим проверить следующие гипотезы:

Нулевая гипотеза H0 : м = 50

Альтернативная гипотеза H1 : μ = 50 .

Среднее арифметическое наблюдений равно:

$$\bar{x} = \frac{\sum_{i=1}^{10} xi}{n} = \frac{484}{10} = 48{,}4 .$$

Подведем итоги расчета таблица:

xi	$(xi - \bar{x})^2$
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
Всего 32,4	

Стандартное отклонение равно:

$$C = \sqrt{\frac{\sum_{i=1}^{10} (xi - \bar{x})^2}{n - 1}} = \sqrt{\frac{32{,}4}{9}} = 1{,}90 .$$

Рассчитываем соотношение:

$$t = \frac{\bar{x} - \mu}{\frac{C}{\sqrt{n}}} = \frac{48{,}4 - 50}{\frac{1{,}90}{\sqrt{10}}} = -2{,}67 .$$

Количество степеней свободы, связанных с тестом, равно n 1 = 9.

Если мы выберем уровень значимости α, равный 5%, то значение tn 1,1 α 2 ученического стола равно:

$$t_{9;0{,}975} = 2{,}26 .$$

Как |т| = 2,67 > 2,26, мы отбрасываем нуль гипотезу и делаем вывод, что при уровне значимости 5% выборка не приходит из населения со средним значением μ = 50.

Студенческий тест на двух образцах

Проводится исследование для сравнения среднего срок службы двух марок шин. Образец был берется за каждую марку. Результаты:

- Для марки 1:
  - Средний срок службы: $\bar{x}_1$ = 72000 км
  - Стандартное отклонение: S1 = 3200 км
  - Размер выборки: n1 = 50

C

Machine Translated by Google

• Для марки 2:

Средний срок службы: $\bar{x}_2 = 74400$ км

Стандартное отклонение: $S_2 = 2400$ км

Объем выборки: $n_2 = 40$

Мы хотим знать, существует ли значительная разница в продолжительности жизни двух брендов.

шин при уровне значимости α, равном 1%.

Отсюда вытекают следующие гипотезы:

Нулевая гипотеза H_0 : $\mu_1 =$

μ_2 или $\mu_1 \quad \mu_2 = 0$

Альтернативная гипотеза H_1 :

$\mu_1 = \mu_2$ или $\mu_1 \quad \mu_1 = 0$.

Взвешенное стандартное отклонение Sp равно:

$$S_p = \sqrt{\frac{(n_1 \quad 1) \cdot S_2 + (n_2 \quad 1) \cdot S_2}{n_1 + n_2 - 2}}^2$$

$$= \sqrt{\frac{49 \cdot 32002 + 39 \cdot 24002}{50 + 40 \quad 2}}$$

$= 2873{,}07$.

Мы можем рассчитать отношение t:

$$t = \frac{\bar{x}_1 \quad \bar{x}_2}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{72000 \quad 74400}{2873.07 \cdot \sqrt{\frac{1}{50} + \frac{1}{40}}}$$

$= -3{,}938$.

Количество степеней свободы, связанных с тестом, равно:

$$\nu = n_1 + n_2 - 2 = 88 .$$

Значение таблицы $t_{\nu,1 \quad \frac{\alpha}{2}}$ нашел в Студенте для α = 1% составляет:

$$t_{88,0{,}995} = 2{,}634 .$$

Как абсолютное значение рассчитанного отношения больше этого значения, $|t| = |\ 3{,}938| >$ 2.634, мы отвергаем нулевую гипотезу для альтернативной гипотезы и заключаем, что при уровне значимости 1% шины двух марок не имеют одинакового среднего срока службы.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Параметрический тест
- Студенческое распределение
- Ученический стол

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Применение «студенческого» распределения. Метрон 5, 90–104 (1925)

Госсет, С. В. «Студент»: вероятная ошибка среднего. Биометрика 6, 1–25 (1908).

Госсет, С.В. «Студент»: Новые таблицы для проверки значимости наблюдений. Метрон 5, 105–120 (1925)

Верное событие

См. образец пространства.

Опрос

Опрос посредством опроса или просто опрос — это опрос ограниченной части населения. Эта часть совокупности составляет выборку , и методы, позволяющие построить эту выборку, называются методами выборки.

Приступая к обследованию, мы получаем информацию о выборочных единицах. Затем с помощью логической статистики мы обобщаем эту информацию на всю популяцию. Это обобщение вносит определенную ошибку. Важность этой ошибки будет зависеть от размера выборки, от того, как они были выбраны по схеме выборки.

Machine Translated by Google

использовалась для составления выборки, и использованная оценка . Чем более репрезентативным выборки, тем меньше ошибка. Выборка методы очень полезны для того, чтобы рисовать образцы, которые приводят к минимальной ошибке. Другой метод заключается в наблюдении за всеми особи популяции. Мы ссылаемся здесь к переписи. Опросы имеют стоимость и скорость преимущества перед переписями, которые в значительной степени компенсируют ошибку выборки. Более того, это иногда невозможно провести перепись, например, когда факт наблюдения за единицей населения означает его уничтожение или когда население бесконечно.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Перепись
- Сбор данных
- Выведенный статистика
- Панель
- Население
- Образец
- Выборка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Gourieroux, C .: Theorie des sondages. Эко номика, Париж (1981)

Grosbras, JM: Статистические методы сондажи. Экономика, Париж (1987)

Систематическая выборка

Систематическая выборка – это случайный тип выборка. Особи берутся из популяции через фиксированные промежутки времени, пространство, или порядок появления, причем первая особь рисуется случайным образом.

ИСТОРИЯ

См. выборку.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим популяцию , состоящую из N особей. Систематическая выборка составляется лиц, численность которых составляет арифметическая прогрессия.

Первое числоb выбирается случайным образом из 1 и r, где r = n — размер $\frac{H}{H}$ выборки , а также обычная разность между последовательными формами арифметической прогрессии.

Лица, отобранные таким образом, будут иметь следующие номера:

$$б, б + г, б + 2г, \ldots , b + (n \quad 1)r .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Систематическая выборка возможна только в том случае, если лица классифицируются в определенном порядке.

Главное преимущество этого метода заключается в его простота и, следовательно, его дешевизна.

С другой стороны, у него есть недостаток без учета возможной периодичности изучаемого признака. Серьезные ошибки могут происходят от этого, особенно если период кратное арифметическому отношению прогресс.

ПРИМЕРЫ

Рассмотрим население в 600 студентов. сгруппированы в алфавитном порядке, из которого мы хотим взять выборку из 30 человек по систематической выборке.

Отношение арифметической прогрессии равно равно:

$$p = \frac{H}{н} \text{ "=" } \frac{600}{30} = 20 .$$

Первое число выбирается случайным образом из 1 и 20, например, b = 17.

Выборка будет состоять из студентов с цифрами:

$$17, 37, 57,..., 597 .$$

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Выборочное обследование
- Выборка

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Деминг, В. Э.: Дизайн образцов в бизнес-исследованиях. Уайли, Нью-Йорк (1960)

Machine Translated by Google

Целевая аудитория

Понятие целевой группы
определяется как популяция , к которой мы
нравится применять результаты расследования.

ПРИМЕРЫ

Мы хотим провести расследование в отношении иностранцев
живет в городе X. Население представлено списком
иностранцев, зарегистрированных в
город.
Целевая популяция определяется совокупностью
иностранцев, проживающих в городе X, в том числе
не включенных в список (необъявленных, родового
назначения и т.п.).

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Население

Испытание независимости

Тест на независимость — это проверка гипотезы
где объективист должен определить, являются ли
две случайные величины независимыми или нет.

ИСТОРИЯ

См. Коэффициент ранговой корреляции Кендалла,
Коэффициент ранговой корреляции Спирмена,
и критерий независимости хи-квадрат.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть (X1, X2,...,Xn) и (Y1, Y2,...,Yn)
две выборки размерности n. Цель
заключается в проверке следующих двух гипотез:

Нулевая гипотеза H0: Две переменные X
и Y независимы.

Альтернативная гипотеза H1: две переменные X и Y
не являются независимыми.

Тест на независимость состоит в сравнении
эмпирического распределения с теоретическим
путем вычисления показателя
тор.

Критерий независимости, примененный к двум
непрерывным переменным, основан на рангах
наблюдения. Так бывает, если тесты
на основе ранговой корреляции Спирмена
коэффициент или на корре ранга Кендалла
ляционный коэффициент.

В случае двух категориальных переменных
Самый распространенный тест – это тест хи-квадрат.
независимости.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Сделать тест на независимость каждой пары
(Xi, Yi), i = 1, 2,...,n должны исходить из
такая же двумерная популяция.

ПРИМЕРЫ

См. критерий независимости хи-квадрат.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Критерий независимости хи-квадрат
Проверка гипотезы
Коэффициент ранговой корреляции Кендалла
Непараметрический тест
Коэффициент ранговой корреляции Спирмена

Временная последовательность

Временной ряд – это последовательность наблюдений измеряется в последовательные моменты времени. Временные ряды ежемесячно, триместрально или ежегодно, иногда еженедельно, ежедневно или ежечасно (изучение дорожного движения, телефонный трафик), либо раз в два года, либо раз в десять лет. Анализ временных рядов состоит из методов, которые попытаться понять такие временные ряды, чтобы делать предсказания.

Временной ряд можно разбить на четыре компонентов, каждый из которых выражает определенное аспект движения ценностей
Временные ряды.

Вот эти четыре компонента:

- Светская тенденция, описывающая движение мент в течение срока;
- Сезонные колебания, которые представляют сезонные изменения;
- Циклические колебания, соответствующие к периодическим, но не сезонным колебаниям;
- Неправильные вариации, другие er неслучайные источники вариаций ряд.

Анализ временных рядов состоит в математическом описании этих элементов, т. е. в оценивании отдельно четырех
компоненты.

ИСТОРИЯ

В статье 1936 г. Функхаузер Х.Г. воспроизвел диаграмму 10-го века, которая показывает наклонение орбиты сев en планеты в функции времени. Аккорд Согласно Кендаллу, это самая старая временная диаграмма, известная в западном мире. Достаточно обратиться к астрономии за происхождение временного ряда. Наблюдение _ звезд был популярен в древности.

В 16 веке астрономия снова источник важных открытий. Работы Браге, Тихо (1546–1601), чей сбор данных о движении планет позволил Кеплеру, Иоганну (1571–1630) сформулировать свои законы движения планет. Статистический анализ данных, как эмпирический, так и это могло быть в то время, было в основе этого великий труд ученого.

Развитие математики в 18 в. и 19 века сделали возможным обход графическая визуализация, чтобы прийти к первому методы анализа временных рядов. Используя частотный подход, ученые разработали сначала работает в этом домене.

Частотный анализ (также называемый гармоническим анализ) изначально был разработан, чтобы выделить один или несколько циклических «компонентов» временной ряд.

В 1919 г. Персонс В.М. предложил декомпозицию временных рядов на тенденции (вековые тренды), циклические циклические колебания), сезонные (сезонные колебания), случайные (нерегулярные вариации) компоненты.

Многие работы посвящены определение и устранение одного или другой из этих компонентов.

Определение тенденций часто были связаны с проблемами регрессионного анализа, а также развивали исследователей относительно нелинейные тенденции, как, например, Кривая Гомперца (установлена в 1825 г. Гомперц, Бенджамин) или логистическая кривая (введен в 1844 году Ферхюльстом, Пьером Франсуа). Другой подход к определению тенденций связан с использованием движущихся средние, потому что они позволяют рассматривать

Machine Translated by Google

проблема тенденций, когда они служат для устранения нерегулярных или периодических вариаций. Использование скользящих средних развивалось в двух направлениях: для актуарных задач и для изучения временных рядов в экономике. Использование скользящих средних особенно развито в рамках методов сезонной корректировки. Точнее, отделение сезонных переменных от других

Компоненты временного ряда были первоначально изучены Коуплендом, М.Т. в 1915 г. и Персоном, В.М. в 1919 г., и продолжены Маколеем, Ф.Р. в 1930 г.
Macauley, FR использовал метод отношений к скользящему среднему, который до сих пор широко используется (см. сезонные колебания), см. также Dufour (2006).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть Yt — временной ряд, который можно разложить с помощью этих четырех составляющих. нц:

• Вековой тренд Tt •
Сезонные вариации St •
Циклические флуктуации Ct •
Нерегулярные вариации It

Мы предполагаем, что значения, принимаемые случайной величиной Yt , определяются соотношением между четырьмя предыдущими компонентами.
Мы различаем три модели композиции: 1. Аддитивная модель:

$$Yt = Tt + Ct + St + It .$$

Предполагается, что четыре компонента независимы друг от друга.

2. Мультипликативная модель:

$$Yt = Tt \cdot Ct \cdot St \cdot It .$$

С помощью логарифмов переходим от мультипликативной модели к аддитивной.

3. Смешанная модель:

$$Yt = St + (Tt \cdot Ct \cdot It)$$

или

$$Yt = Ct + (Tt \cdot St \cdot It).$$

Этот тип модели используется очень мало. Выбираем аддитивную модель при сезонности вариации почти постоянны и их влияние влияние на тенденцию не зависит от ее уровня.

Когда сезонные вариации имеют амплитуду, почти пропорциональную амплитуде вековых трендов, мы вводим мультипликативную модель; это вообще так.
Анализ временных рядов заключается в определении значений, принимаемых каждой компонентой. Мы всегда начинаем с сезонных колебаний и заканчиваем циклическими колебаниями. Все колебания, которые нельзя отнести ни к одному из три компонента будут сгруппированы с нерегулярными вариациями.
На основе используемой модели анализ может после определения компонентов скорректировать последующие данные путем вычитания или деления. Таким образом, когда мы оценили светское тенденция Tt, в каждый момент времени t временной ряд корректируется следующим образом:

– Если использовать аддитивную модель Yt = Tt + Кт + Ст + Ит:

$$Yt - Tt = Ct + St + It .$$

– Для мультипликативной модели Yt = Tt ·Ct · Ст · Оно:

$$\frac{Yt}{Tt} = Ct \cdot St \cdot It .$$

Последовательно оценивая каждый из трех компонентов и корректируя временной ряд среди них, мы получаем значения, относящиеся к нерегулярным вариациям.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Когда временные ряды не имеют достаточного периода времени, мы объединяем вековой тренд и циклические колебания в один компонент, называемый внесезонным движением.

Графическое представление

Графическое представление временных рядов помогает анализировать наблюдаемые значения в зависимости от времени.

Чтобы проанализировать эти значения, необходимо принять ряд мер предосторожности: 1. Каждая

точка должна быть репрезентативной для наблюдения.
Если значение представляет акции на фиксированную дату, то точка находится на базовую дату. Если есть

поток, например, месячный, то точка фиксируется в середине месяца.

2. Иногда приходится отображать на одном графике годовые и месячные данные. Репрезентативные точки результатов должны быть в 12 раз чаще, чем точки относительно годовых результатов. Таким образом, мы должны

обратите внимание на репрезентативные точки в нужном месте.

3. Если на одном и том же графике представлено много рядов, они должны быть идентичными по своему характеру.
В противном случае мы должны использовать разные масштабы или располагать графики друг под другом.

Основными задачами исследования временных рядов являются:

- Описание: определение или исключение различных компонентов • Стохастическое

моделирование •

Прогнозирование •

Фильтрация: устранение или преобразование определенные характеристики

- Контроль: наблюдение за прошлым процессом, чтобы реагировать на будущую эволюцию.

Для прогнозирования анализ временных рядов основывается на следующих гипотезах:

- Великая исторически наблюдаемая тенденция
связи поддерживаются в более или менее будущее.
- Измеряемые колебания переменной

воспроизводятся через равные промежутки времени.

Анализ временных рядов будет оправдан до тех пор, пока он позволяет снизить уровень неопределенности при разработке прогнозов.

Мы различаем: 1.

Долгосрочный прогноз: на

основе векового тренда; обычно не принимает во внимание циклические колебания.

2. Среднесрочный прогноз. Чтобы

учесть вероятное влияние циклической составляющей, мы должны умножить прогнозируемое значение векового тренда на оценку относительной дисконтированной вариации, связанной с циклическими колебаниями.

3. Краткосрочный прогноз. Как

правило, мы не пытаемся прогнозировать неравномерные колебания или делать краткосрочные прогнозы.

слепки или прогнозы относительно более длительных периодов. Таким образом, мы умножаем прогнозируемые значения векового тренда на оценку циклических колебаний и на сезонный ход соответствующей даты.

При анализе временных рядов мы сталкиваемся со следующими проблемами:

1. Подходит ли классическая модель?

Мы ожидаем, что качество результатов анализа будет пропорционально точности и точности самой модели .

Мультипликативная модель Yt = Tt·St·Ct·It предполагает, что четыре компонента независимы, что в действительности редко соответствует действительности.

2. Верны ли наши гипотезы о постоянстве и регулярности?

Анализ должен быть скорректирован с учетом субъективных и качественных факторов, существующих практически в любой ситуации.

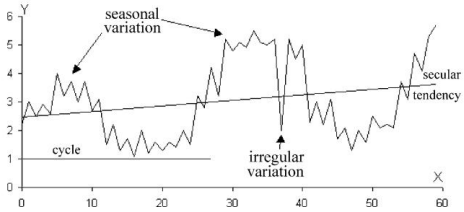
Machine Translated by Google

3. Можем ли мы доверять имеющимся данным?

Проблема может возникнуть из-за разрывов рядов из-за недостаточности данных или изменения качества переменных.

ПРИМЕРЫ

Временной ряд обычно представляется следующим образом:



Мы выделяем четыре компонента: • Вековой тренд, незначительно увеличивающийся в Настоящий случай • Сезонные колебания, хорошо заметные • Циклические колебания в форме циклов приблизительной амплитуды в 27 единиц времени

• Нерегулярные вариации, как правило, достаточно слабые, за исключением t = 38, что представляет собой очень резкое падение, которое было бы нерациональным.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Циклическое колебание
- Прогнозирование
- Графическое представление
- Неправильная вариация
- Сезонная вариация
- Светская тенденция

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Коупленд, М.Т.: Статистические показатели условий ведения бизнеса. QJ Экон. 29, 522–562 (1915)

Дюфур, IM: Histoire de l'analyse de Séries Chronologues, Монреальский университет, Канада (2006 г.)

Функхаузер, Х. Г.: Заметка о графике 10 века. Осирис 1, 260–262 (1936)

Кендалл, MG: временные ряды. Гриффин, Лондон (1973)

Macauley, FR: Сглаживание временных рядов. Национальное бюро экономических Исследования, 121–136 (1930).

Persons,WM: Индекс общих условий ведения бизнеса. Преподобный эконом. Стат. 1, 111–205 (1919)

След

Для квадратной матрицы A порядка n определим след матрицы A как сумму членов, расположенных по диагонали. Таким образом, след матрицы является скалярной величиной.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть A — квадратная матрица порядка n, A = где i, j = 1, 2,...,n. Мы определяем aij , след A следующим образом:

$$\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Поскольку нас интересуют только диагональные элементы, трасса определена только для квадратных матриц.

1. След суммы двух матриц равно сумме следов матрицы се:

$$\operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B).$$

2. След произведения двух матриц не меняется при перестановках матриц. Математически: $\operatorname{tr}(A \cdot B) = \operatorname{tr}(B \cdot A)$.

Machine Translated by Google

ПРИМЕРЫ

Пусть A и B соединяют две квадратные матрицы порядка 2:

$$A = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} \quad \text{и } B = \begin{matrix} 5 & 2 \\ 2 & 3 \end{matrix} .$$

Расчет следов дает:

$$\text{тр}(A) = \text{тр} \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix} = 1 + 3 = 4 ,$$

$$\text{тр}(B) = \text{тр} \begin{matrix} 5 & 2 \\ 2 & 3 \end{matrix} = 5 + 3 = 8 ,$$

и

$$\text{тр}(A + B) = \text{тр} \begin{matrix} 6 & 4 \\ 1 & 6 \end{matrix} = 6 + 6 = 12 .$$

Проверим, что tr(A) + tr(B) = 12. На другая рука:

$$A \cdot B = \begin{matrix} 1 & 8 \\ 1 & 11 \end{matrix} \quad \text{и}$$

$$B \cdot A = \begin{matrix} 7 & 16 \\ 1 & 5 \end{matrix} ,$$

$$\text{tr}(A \cdot B) = \text{tr} \begin{matrix} 1 & 8 \\ 1 & 11 \end{matrix} = 1 + 11 = 12 ,$$

$$\text{tr}(B \cdot A) = \text{tr} \begin{matrix} 7 & 16 \\ 1 & 5 \end{matrix} = 7 + 5 = 12 .$$

Таким образом, $\text{tr}(A \cdot B) = \text{tr}(B \cdot A) = 12$.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Матрица

Трансформация

Трансформация – это изменение одного или нескольких переменных в статистическом исследовании.

Преобразуем переменные, например, заменив их своими логарифмами (логарифмическое преобразование).

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Мы преобразуем переменные по многим причинам:

1. Когда у нас есть набор данных относительно многих переменных, и мы хотим выразить одну из них, Y (зависимую переменную), с помощью других (называемых независимыми переменными), и мы хотим линеаризовать связь между переменными. Например, когда отношение между независимые переменные мультипликативны, логарифмическое преобразование _ данные делают его аддитивным и позволяют использовать линейной регрессии для оценки параметров модели .
2. Когда у нас есть данные установить относительно переменной и ее дисперсии непостоянна, обычно можно сделать его почти постоянным, преобразовав эта переменная.
3. Когда случайная величина следует за любым распространение, адекватное преобразование позволяет получить новую случайную вариацию способныйпримерно соответствует нормальному распределение.

Преимущества преобразования набора данных относительно переменной в основном сводятся к следующему: следует:

- Упрощает существующие отношения с другими переменные.
- Исправление выбросов, линейности и гомоскедастичность.
- Он дает переменную, распределенную приблизительно нормально.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Зависимая переменная
- Независимая переменная
- Логарифмическое преобразование
- Нормальное распределение
- Регрессивный анализ
- Дисперсия

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кокс, Д.Р., Бокс, ГЭП: анализ трансформаций (с обсуждением). Дж. Рой. Стат. соц. сер. Б 26, 211–243 (1964)

Транспонировать

(Матрица), транспонированная матрицы A порядка (m × n), представляет собой матрицу (n × m), , обозначаемую A , полученную путем записи строк A в столбцах A и наоборот.

Таким образом, мы производим простую «инверсию» таблицы.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Если A = aij , i = 1, 2,...,m и j = 1, 2,...,n — матрица порядка (m × n), то транспонирование матрицы A — это матрица A порядка (n × m), определяемый как:

$$A = aji \text{ c } j = 1, 2,...,n \text{ и } i = 1, 2,...,m.$$

Транспозиция — обратимая и реципрокная операция; таким образом, транспонируя транспонированную матрицу, мы находим исходную матрицу.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Мы используем транспонирование матрицы или, точнее , транспонирование вектора при вычислении скалярного произведения.

ПРИМЕРЫ

Пусть A — следующая матрица порядка (3 × 2):

$$A = \begin{matrix} 1 & 2 \\ 0 & 3 \\ 2 & 5 \end{matrix}.$$

Транспонирование A - это матрица (2 × 3):

$$A = \begin{matrix} 1 & 0 & 2 \\ 2 & 3 & 5 \end{matrix}.$$

Мы также можем проверить, что транспонированная матрица A равна исходной матрице A:

$$A = A .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Матрица
- Вектор

Уход

В экспериментальном плане лечение представляет собой определенную комбинацию уровней различных факторов.

ПРИМЕРЫ

Часто проводятся эксперименты для сравнения двух или более способов лечения, например, двух разных удобрений для определенного типа растений или многих типов лекарств для лечения определенного заболевания. Другой пример состоит в измерении времени свертывания проб крови 16 животных, поддерживавших разные режимы A, Б, В и Г. При этом разными примерами обработок являются, соответственно, удобрения, препараты и режимы.

С другой стороны, в экспериментах по испытанию конкретного удобрения, например азота, на урожае пшеницы, мы можем рассматривать разные количества одного и того же удобрения как разные обработки. Здесь у нас есть один фактор (азот) на разных уровнях (количествах), например, 30 кг, 50 кг, 100 кг и 20
Каждый из этих уровней соответствует угощению мент.

Machine Translated by Google

Цель эксперимента — определить, есть ли реальная (значимая) разница между этими методами лечения.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

План экспериментов

Эксперимент

Экспериментальная установка

Фактор

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Сенн, Сент-Луис: перекрестные клинические испытания Исследовать. Уайли (1993)

Тьюки, Джон Уайлдер

Тьюки, Джон Уайлдер родился в Бедфорде, Массачусетс, в 1915 году. Он изучал химию в Университете Брауна и получил докторскую степень по математике в Принстонском университете в 1939 году. В возрасте 35 лет он стал профессором математики в Принстоне. Он был ключевой игрок в формировании факультета статистики Принстонского университета в 1966 году. Он был ее председателем в 1966–1969 гг. Он является автором книги «Исследовательский анализ данных» (1977 г.) (перевод на русский язык) и восьми томов статей. Он является соавтором Статистических проблем отчета Кинси о сексуальном поведении мужчин; Анализ данных и регрессия (также переведена на русский язык), Index to Statistics and Probability. Он является соредактором следующих книг: Understanding Robust and Exploratory Data Analysis, Exploring Data Tables, Trends Shapes, Configural Polysampling и Fundamentals of Exploratory Analysis of Variance. Он умер в Нью-Джерси в июле 2000 года.

Избранные основные работы и статьи Тьюки, Джон Уайлдер:

1977 Исследовательский анализ данных. 1-е изд. Аддисон-Уэсли, Рединг, Массачусетс.

1980 г. Нам нужны как исследовательские, так и подтверждающие данные. Являюсь. Стат. 34, 23–25.

1986 Анализ данных и наука о поведении, или научиться нести человеческое бремя количественных вычислений, избегая плохих повелений. В: Собрание сочинений Джона У. Тьюки, Том III: Философия и принципы анализа данных, 1949–1964. (Джонс, Л.В., изд.) Wadsworth Advanced Books & Software, Монтерей, Калифорния.

1986 Собрание сочинений Джона В. Тьюки, Том III: Философия и принципы анализа данных, 1949–1964. (Джонс, Л.В., изд.) Wadsworth Advanced Books & Software, Монтерей, Калифорния.

1986 г. Собрание сочинений Джона В. Тьюки, том IV: философия и принципы анализа данных, 1965–1986 гг. (Джонс, Л.В., изд.) Wadsworth Advanced Books & Software, Монтерей, Калифорния.

1988 Собрание сочинений Джона В. Тьюки, Том V: Графика, 1965–1985. (Кливленд, Вашингтон, изд.) Уолдсворт и Брукс/Коул, Пасифик-Гроув, Калифорния.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Коробчатый сюжет

Исследовательский анализ данных

Двусторонний тест

Двусторонний тест, относящийся к совокупности, — это тест гипотезы , который применяется, когда мы хотим сравнить оценку параметра с заданным значением с альтернативной гипотезой, не равной значению состояния.

Machine Translated by Google

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Двусторонний тест — это проверка гипотезы на популяции типа:

Нулевая гипотеза H0 : $\theta = \theta 0$

Альтернативная гипотеза H1 : $\theta = \theta 0$,

где θ — параметр популяции значение которого неизвестно, а θ0 – предполагаемое значение этого параметра. Для теста, касающегося сравнения двух популяции, гипотезы таковы:

Нулевая гипотеза H0 : $\theta 1 = \theta 2$

Альтернативная гипотеза H1 : $\theta 1 = \theta 2$,

где θ1 и θ2 — неизвестные параметры из двух базовых популяций.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

В двустороннем тесте область отторжения разделить на две части, левую и правую стороны рассматриваемого параметра.

ПРИМЕРЫ

Компания производит стальные тросы. Он хочет проверить, соответствует ли диаметр μ большей части производимого кабеля стандартному диаметру эфир 0,9 см.

Берется выборка из 100 кабелей, среднее значение которых диаметр 0,93 см со стандартным отклонением S = 0,12 . _

Гипотезы следующие:

Нулевая гипотеза H0 : $\mu = 0{,}9$

Альтернативная гипотеза H1 : $\mu = 0{,}9$,

Поскольку объем выборки достаточно велик, выборочное распределение среднего может быть аппроксимируется нормальным распределением. Для уровня значимости α = 5% получаем критическое значение zα/2 в нормальной таблице zα/2 = 1,96.

Находим область принятия нуля гипотезы (см. область отклонения):

$$\mu \pm z\alpha/2 \cdot \frac{\sigma}{\sqrt{n}},$$

где σ — стандартное отклонение

Население. Поскольку стандартное отклонение σ равно неизвестно, мы оцениваем его с помощью эталона отклонение S образца (S = 0,12), которое дает:

$$0{,}9 \pm 1{,}96 \cdot \frac{0{,}12}{\sqrt{100}} = 0{,}8765 \; 0{,}9235 .$$

Область принятия нулевой гипотезы равна интервалу [0,8765, 0,9235]. Поскольку среднее значение выборки (0,93 см) вне этого интервала, мы должны отбросить нулевая гипотеза для альтернативной гипотезы. Делаем вывод, что на уровне значимости 5% и в зависимости от исследуемой выборки, диаметр кабелей не соответствует к норме.

Проводится исследование для сравнения среднего срок службы двух типов шин. Образец 50 шин марки 1 дают средний срок службы 72000 км, а выборка 40 шин марки 2 имеет средний срок службы 74400 км. Предположим, что стандартное отклонение для марки 1 составляет σ1 = 3200 км, а марки 2 σ2 = 2400 км, мы хотим знать, есть ли значительная разница в продолжительности жизни между два бренда, на уровне значимости α = 1%.

Гипотезы следующие:

Нулевая гипотеза H0 : $\mu 1 = \mu 2$

Альтернативная гипотеза H1 : $\mu 1 = \mu 2$.

Размер образца достаточно велик, и стандартные отклонения популяций известны, так что мы можем приблизиться к

T

Machine Translated by Google

выборочное распределение с использованием нормального распределения.

Область принятия нулевой гипотезы определяется (см. Область отклонения):

$$\mu 1 \quad \mu 2 \pm z\alpha/2 \cdot \sigma \bar{x}1 \quad \bar{x}2 ,$$

где $\sigma \bar{x}1 \quad \bar{x}2$ — стандартное отклонение выборочного распределения разности двух средних или стандартная ошибка двух средних.

Значение $z\alpha/2$ находится в обычной таблице. Для двустороннего теста с $\alpha = 1\%$ мы получаем $z\alpha/2 = 2{,}575$, что дает:

$$0 \pm 2{,}575 \cdot \sqrt{\frac{\sigma 2_1}{n1} \sigma 2 + \frac{2}{n2}}$$

$$= 0 \pm 2{,}575 \cdot \sqrt{\frac{32002}{50} \frac{24002}{40} +}$$

$$= 0 \pm 2{,}575 \cdot 590{,}59$$

$$= \pm 1520{,}77 .$$

Область приемлемости нулевой гипотезы задается интервалом [-1520,77, 1520,77]. Поскольку разница в средних значениях выборок (72 000 - 74 400 = -2400) находится за пределами допустимой области, мы должны отвергнуть нулевую гипотезу в пользу альтернативной гипотезы. Таким образом, можно сделать вывод о существенной разнице между сроками службы шин

проверенные бренды.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Область приема
- Альтернативная гипотеза
- Проверка гипотезы
- Нулевая гипотеза
- Односторонний тест
- Область отторжения
- Выборочное распределение
- Уровень значимости

Двусторонний дисперсионный анализ

Двусторонний дисперсионный анализ представляет собой расширение однофакторного дисперсионного анализа, в котором есть два независимых фактора (переменных). Каждый фактор имеет два или более уровней, и лечение формируется путем составления всех возможных комбинаций уровней двух факторов.

ИСТОРИЯ

Именно Фишер Р.А. (1925) дал название «факториальные эксперименты» сложным экспериментам.

Йейтс Ф. (1935, 1937) разработал концепцию и анализ этих факторных экспериментов.

См. дисперсионный анализ.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Дан эксперимент с двумя факторами: фактор A имеет уровни a , а фактор B имеет уровни b .

Если план, связанный с этим факторным экспериментом , представляет собой полностью рандомизированный план, модель выглядит следующим образом:

$Yijk = \mu + \alpha i + \beta j + (\alpha\beta)ij + \varepsilon ijk$, $i = 1, 2,\ldots,a$ (уровни фактора A) $j = 1, 2,\ldots,b$ (уровни фактора B) $k = 1, 2,\ldots,c$ (количество наблюдений получающих лечение ij)

где μ — общее среднее , общее для всех обработок, αi — эффект уровня i фактора A, βj — эффект уровня j фактора B, $(\alpha\beta)ij$ — эффект взаимодействия между

Machine Translated by Google

αi и βj, а εijk — экспериментальная ошибка наблюдения Yijk .

Эта модель подвергается основным предположениям , связанным с дисперсионным анализом, если предполагается, что ошибки εijk являются независимыми случайными величинами, подчиняющимися аномальному распределению N(0, σ2).

Затем можно проверить три гипотезы :

1.

H0 : α1 = α2 = ··· = αa

H1 : По крайней мере один αi отличается из αj, i = j.

2.

H0 : β1 = β2 = H1 : ··· = βb

По крайней мере один βi отличается из βj, i = j.

3.

H0 : (αβ)11 = (αβ)12 = ··· = (αβ)1b = (αβ)21 = ··· = (αβ)ab

H1 : По крайней мере один из взаимодействие отличается от других.

Распределение Фишера используется для проверки первая гипотеза. Это распределение требует создание отношения, числитель которого оценка дисперсии фактора A и знаменатель которой является оценкой переменной ошибка в лечении, известная также как ошибка или остаточное.

Это отношение, обозначаемое F, соответствует закону Фишера. распределение со степенями a 1 и ab(c 1) свободы.

Нулевая гипотеза H0 : α1 = α2 = ··· = αa

будет отвергнут на значимом уровне α, если отношение F больше или равно значению в таблице Фишера, то есть если

F Fa 1,ab(c 1),α .

Для проверки второй гипотезы создадим отношение, числитель которого является оценкой дисперсия фактора B и номинал которого то есть оценка вариации в менты.

Это отношение, обозначаемое F, соответствует закону Фишера. распределения со степенями b 1 и ab(c 1) свободы.

Нулевая гипотеза H0 : β1 = β2 = ··· = βb

будет отклонено, если коэффициент F больше или равно значению таблицы Фишера , т. е. если

F Fb 1,ab(c 1),α .

Для проверки третьей гипотезы составим соотношение числитель которого является оценкой дисперсия взаимодействия между фак торы A и B, знаменатель которых оценка дисперсии в рамках лечения.

Это отношение, обозначаемое F, соответствует закону Фишера. распределение с (a-1) (b-1) и ab (c-1) степени свободы.

Нулевая гипотеза H0 : (αβ)11 = (αβ)12 = ··· = (αβ)1b = (αβ)21 = ··· = (αβ)ab

будет отклонено, если коэффициент F больше или равно значению в таблице Фишера, т. е. если

F F(a 1)(b 1),ab(c 1),α .

Дисперсия фактора A

Для дисперсии фактора A сумма квадраты должны быть рассчитаны для фактора A (SSA), который получается следующим образом:

$$CCA = б \cdot c \sum_{я=1}^{a} (\bar{Y}i.. \quad \bar{Y}...)^2,$$

где $\bar{Y}i..$ — среднее значение всех наблюдений уровня i фактора A и $\bar{Y}...$ является общим среднее значение всех наблюдений.

T

Machine Translated by Google

Число степеней свободы , связанных с этой суммой, равно a 1.

Тогда дисперсия фактора A равна:

$$c_A^2 \text{ "=" } \frac{\text{ССА}}{\text{а - 1}} .$$

Дисперсия фактора B

Для дисперсии фактора B сумма квадраты должны быть рассчитаны для фактора B (SSB), который получается следующим образом:

$$\text{SSB} = \text{a} \cdot \text{c} \sum_{\text{j=1}}^{\text{б}} (\bar{Y}.\text{j.} \quad \bar{Y}...)^2 ,$$

где $\bar{Y}$i.. — среднее значение всех наблюдений уровня i фактора B, а $\bar{Y}$ ... является общим среднее значение всех наблюдений.

Число степеней свободы , связанных с этой суммой, равно b 1.

Тогда дисперсия фактора B равна :

$$c_Б^2 \text{ "=" } \frac{\text{SSB}}{\text{б - 1}} .$$

Дисперсия взаимодействия AB

Для дисперсии взаимодействия AB сумма квадратов необходимо вычислить для взаимодействия AB (SSAB), которое получается следующим образом:

$$\text{SSAB} = \text{c} \sum_{\text{я=1}}^{\text{a}} \sum_{\text{j=1}}^{\text{б}} (\bar{Y}\text{ij.} \quad \bar{Y}\text{i..} \quad \bar{Y}.\text{j.} + \bar{Y}...)^2 ,$$

где $\bar{Y}$ij. является средним значением всех наблюдений уровня взаимодействия AB,$\bar{Y}$i..среднее значение все наблюдения уровня i фактора A, $\bar{Y}$.j. является средним значением всех наблюдений уровня j фактора B, а $\bar{Y}$... — общее среднее всех наблюдения.

Число степеней свободы , связанных с этой суммой, равно (a 1)(b 1).

Тогда дисперсия взаимодействия AB равна равно:

$$c_{АБ}^2 \text{ "=" } \frac{\text{ССАБ}}{\text{(а - 1) (б - 1)}} .$$

Расхождение в лечении или ошибка

Для дисперсии внутри обработок сумма квадратов внутри обработок получается как следует:

$$\text{ССЭ} = \sum_{\text{я=1}}^{\text{a}} \sum_{\text{j=1}}^{\text{б}} \sum_{\text{к=1}}^{\text{c}} (\text{Yijk} - \bar{Y}\text{ij.})^2 ,$$

где a и b – соответственно число уровней факторов A и B , c – число наблюдений, получающих обработку ij (постоянная для любого iandj), Yijk – наблюдениек из уровень i фактора A и уровень j фактора B, и $\bar{Y}$ij. это среднее значение всех наблюдений уровень ij взаимодействия AB.

Число степеней свободы , связанных с этой суммой, равно ab(c 1).

Таким образом , дисперсия внутри лечения равна к:

$$c_E^2 \text{ "=" } \frac{\text{ССЭ}}{\text{аб (с - 1)}} .$$

Общая сумма квадратов (SST) равна:

$$\text{ТСТ} = \sum_{\text{я=1}}^{\text{a}} \sum_{\text{j=1}}^{\text{б}} \sum_{\text{к=1}}^{\text{c}} \text{Y2}_{\text{эйк}} .$$

Число степеней свободы , связанных с этой суммой, равно N, т. е. общее количество наблюдений.

Сумма квадратов среднего равна :

$$\text{SSM} = \text{N}\bar{Y}^2 \text{ ...} .$$

Число степеней свободы , связанных с этой суммой, равно 1.

Общая сумма квадратов (SST) может быть выражается со всеми остальными суммами квадратов следующим образом:

$$\text{SST} = \text{SSM} + \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE} .$$

Тесты Фишера

Теперь мы можем рассчитать различные коэффициенты F.

Machine Translated by Google

Чтобы проверить нулевую гипотезу

$$H0 : \alpha 1 = \alpha 2 = \quad \cdots \quad = aa ,$$

составляется первое отношение F , числитель которого является оценкой дисперсии коэффициента _ _ tor A, а знаменатель – оценка разница в лечении:

$$\Phi = \frac{c_A^2}{c_E^2} .$$

Если F больше или равно значению в Таблица Фишера для степеней a 1 и ab(c 1) свободы нулевая гипотеза отвергается и делается вывод, что хотя бы одно αi отличается от αj, i = j.

Чтобы проверить нулевую гипотезу

$$H0 : \beta 1 = \beta 2 = \quad \cdots \quad = \beta b ,$$

составляется второе отношение F , числитель которого является оценкой дисперсии фактора B а знаменатель есть оценка переменной в рамках лечения:

$$\Phi = \frac{c_B^2}{c_E^2} .$$

Если F больше или равно значению в Таблица Фишера для степеней b 1 и ab(c 1) свободы нулевая гипотеза отвергается и делается вывод, что H верно.

Чтобы проверить нулевую гипотезу

$$H0 : (\alpha\beta)11 = (\alpha\beta)12 = = (\alpha\beta)21 = \cdots = (\alpha\beta)1b$$
$$\cdots = (\alpha\beta)ab$$

составляется третье отношение F , числитель которого равен оценка дисперсии взаимодействия _ _ _ AB, а знаменателем которого является оценка формирование дисперсии внутри обработок:

$$\Phi = \frac{c_{AB}^2}{c_E^2} .$$

Если F больше или равно значению в Таблица Фишера для (a 1)(b 1) и ab(c 1) степеней свободы , нулевая гипотеза отклоняется и делается вывод, что H, верно.

Таблица дисперсионного анализа

Вся информация, необходимая для расчета два коэффициента F можно свести в таблицу дисперсионный анализ:

Источник разнообразия ция	Степени бесплатно дом	Сумма квадраты	Среднее значение квадраты	Ф
Среднее 1		ССМ		
Фактор A a - 1		ССА	c_A^2	$\frac{c_A^2}{c_E^2}$
Фактор B b - 1		SSB	c_B^2	$\frac{c_B^2}{c_E^2}$
Интер действие АБ	(a - 1) · (б - 1)	SSAB s2	АБ	$\frac{c_{AB}^2}{c_E^2}$
В пределах обращаться менты	аб(с-1) SSI		c_E^2	
Общий	азбука	SST		

Если c = 1, это означает, что существует только один наблюдение за обработку ij (или за ячейку ij), невозможно атрибутировать часть ошибки «внутри групп» к взаимодействию.

ПРИМЕРЫ

Директор школы хочет проверить четырех марки пишущих машинок. Он просит пять профессиональных секретарей попробовать каждую марку и повторить упражнение на следующий день. Все они должны ввести тот же текст, а через 15 мин средний фиксируется количество слов, набранных за 1 мин. Вот результаты:

Machine Translated by Google

Печатная машинка	Секретарь				
	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			

Проведем двусторонний дисперсионный анализ, первый фактор — бренды пишущих машинок, а второй — секретарши.

Нулевые гипотезы следующие:

1.

$H0 : \alpha 1 = \alpha 2 = \alpha 3 = \alpha 4$

H1 : По крайней мере один αi отличается из αj, i = j.

2.

$H0 : \beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5$

H1 : По крайней мере один βi отличается из βj, i = j.

3.

$H0 : (\alpha\beta)11 = (\alpha\beta)12 = = (\alpha\beta)21 \cdots = (\alpha\beta)15 = \cdots = (\alpha\beta)45$

H1 : По крайней мере одно из взаимодействий отличается от других.

Дисперсия фактора А «Марки пишущих машинок»

Дисперсия фактора А

Сумма квадратов фактора А равна:

$$CCA = 5 \cdot 2 \sum_{я=1}^{4} (\bar{Y}i.. - \bar{Y}...)^2$$

$$= 5 \cdot 2 (33 - 34)^2 + (35,5 - 34)^2 + (36 - 34)^2 + (31,5 - 34)^2 = 135 .$$

Число степеней свободы , связанных с этой суммой, равно 4 1 = 3.

Дисперсия фактора А (среднее квадратов) тогда равно:

$$c_A^2 = \frac{CCA}{a - 1} = \frac{135}{3} = 45 .$$

Дисперсия фактора В «Секретарши»

Сумма квадратов фактора В равна:

$$ССБ = 4 \cdot 2 \sum_{j=1}^{5} (\bar{Y}.j. - \bar{Y}...)^2$$

$$= 4 \cdot 2 (34 - 34)^2 + (33,5 - 34)^2 + ... + (35,5 - 34)^2 + (34,5 - 34)^2 = 40 .$$

Число степеней свободы , связанных с этой суммой, равно 5 1 = 4.

Дисперсия фактора В (среднее значение квадратов) тогда равно:

$$c_Б^2 = \frac{SSB}{б - 1} = \frac{40}{4} = 10 .$$

Дисперсия взаимодействия АВ «Марки пишущих машинок – Секретарши»

Сумма квадратов взаимодействия АВ равна равно:

$$ССАБ = 2 \sum_{я=1}^{4} \sum_{j=1}^{5} (\bar{Y}ij. - \bar{Y}i.. - \bar{Y}.j. + \bar{Y}...)^2$$

$$= 2 (34,5 - 33 - 34 + 34) 2+ (31 - 31,5 - 34,5 + 34) 2 + ... +(33 - 31,5 - 34,5 + 34)^2$$

Machine Translated by Google

$$= 2\,1{,}52 + (\quad 1{,}5)^{2} + \ldots + 12$$

$$= 83 .$$

Число степеней свободы , связанных с этой суммой, равно (4 1)(5 1) = 12.

Тогда дисперсия взаимодействия AB равна равно:

$$\sigma^{2}_{\text{АБ}} = \frac{\text{SSAB}}{(a - 1)(б - 1)} = \frac{83}{12} = 6{,}92 .$$

Расхождение внутри лечения (ошибка)

Сумма квадратов внутри обработок равна:

$$\text{ССЭ} = \sum_{я=1}^{4}\sum_{j=1}^{5}\sum_{k=1}^{2}(Yijk - \bar{Y}_{ij.})^{2}$$

$$2\,2 = (33 - 34{,}5) + (36 - 34{,}5) + (31 \quad 31)^{2} + \ldots + (33 \quad 33)^{2} = 58 .$$

Число степеней свободы , связанных с этой суммой, равно 4·5(2 1) = 20:

$$\sigma^{2}_{\text{Е}} = \frac{\text{SSI}}{\text{аб}\,(c - 1)} = \frac{58}{20} = 2{,}9 .$$

Сумма квадратов равна:

$$\text{ТСТ} = \sum_{i=1}^{4}\sum_{j=1}^{5}\sum_{k=1=}^{2}(\text{Йийк})^{2}$$

$$332 + 362 + 312 + \ldots + 272 + 332 + 332 = 46556 .$$

Число степеней свободы , связанных с этой суммой, равно количеству N наблюдений, т. е. 40.

Сумма квадратов среднего равна :

$$\text{SSM} = N\bar{Y}^{2}_{\ldots} = 40 \cdot 342 = 46240 .$$

Число степеней свободы , связанных с этой суммой, равно 1.

Тесты Фишера

Чтобы проверить нулевую гипотезу

$$H0 : \alpha1 = \alpha2 = \alpha3 = \alpha4 ,$$

мы формируем первое отношение F , числитель которого является оценкой дисперсии коэффициента _ _ tor A, а знаменатель – оценка разница в лечении:

$$\Phi = \frac{c^{2}_{\text{A}}}{c^{2}_{\text{E}}} = \frac{45}{2{,}9} = 15{,}5172 .$$

Значение таблицы Фишера с 3 и 20 степеней свободы для значимого уровня α = 0,05 равно 3,10.

Поскольку F F3,20,0,05, мы отвергаем гипотезу H0 и делаем вывод, что марки пишущих машинок оказывают существенное влияние на количество слов, набираемых за 1 мин.

Чтобы проверить нулевую гипотезу

$$H0 : \beta1 = \beta2 = \beta3 = \beta4 = \beta5 ,$$

мы образуем второе отношение Fratio , числитель которого является оценкой дисперсии фактора B и чьим знаменателем является оценка

T

Machine Translated by Google

разница в лечении:

$$\Phi = \frac{c_B^2}{c_E^2} \text{ "=" } \frac{10}{2,9} = 3,4483 .$$

Значение таблицы Фишера с 4 и 20 степеней свободы для α = 0,05 равна до 2,87.

Поскольку F F4,20,0,05, мы отвергаем H0 гипотезы и сделать вывод, что секретари существенно влияет на количество слов, набранных за 1 мин.

Чтобы проверить нулевую гипотезу

$$H0 : (\alpha\beta)11 = (\alpha\beta)12 = = (\alpha\beta)21 = \cdots = (\alpha\beta)15 \cdots = (\alpha\beta)45 ,$$

мы образуем третье отношение F , числитель которого является оценкой дисперсии взаимодействия _ AB, а знаменателем которого является оценка формирование дисперсии внутри обработок:

$$\Phi = \frac{c_{AB}^2}{c_E^2} \text{ "=" } \frac{6,92}{2,9} = 2,385 .$$

Значение таблицы Фишера с 12 и 20 степеней свободы для α = 0,05 равна до 2,28.

Поскольку F F4,20,0,05, мы отвергаем H0 гипотезы и сделать вывод, что влияние Взаимодействие секретаря и Тип пишущей машинки значим по количеству слов, набранных за 1 мин.

Таблица дисперсионного анализа

Вся эта информация сведена в таблица дисперсионного анализа:

Источник разнообразия ция	Степени бесплатно дом	Сумма квадраты	Среднее значение квадраты	Ф
Иметь в виду	1	46240		
Фактор А 3		135 45		15.5172
Фактор В 4		40 10		3.4483
Интер действие АБ	12	83	6,92 2,385	
В пределах обращаться менты	20	58	2,9	
Общий	40	46556		

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Контраст
- Таблица Фишера
- Тест Фишера
- Взаимодействие
- Тест на наименьшее существенное различие
- Односторонний дисперсионный анализ

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Кокс, Д.Р., Рид, Н.: Теория дизайна экспериментов. Чепмен и Холл, Лондон (2000)

Фишер, Р.А.: Статистические методы для Исследовательские работники. Оливер и Бойд, Эдин Бург (1925)

Йейтс, Ф.: Сложные эксперименты. Доп. Дж. Рой. Стат. соц. 2, 181–247 (1935)

Йейтс, Ф.: Планирование и анализ факторных экспериментов. Техническое сообщение Бюро Содружества почв 35, Сельскохозяйственное содружество Бюро, Фарнхэм Роял (1937)

Ошибка типа I

Когда проводится проверка гипотез , ошибка первого рода возникает, когда человек отвергает

Machine Translated by Google

нулевая гипотеза , когда она верна, а α
вероятность отклонения нулевой гипотезы
H0 , когда это правда:

$$\alpha = P(\text{отклонить } H0 \mid H0 \text{ истинно}) .$$

Вероятность ошибки I рода α равна
уровень значимости проверки гипотезы.

ИСТОРИЯ

В 1928 г. Нейман, Дж. и Пирсон, Э.С.
первые авторы, осознавшие, что рациональный
выбор проверки гипотезы должен
принимать во внимание не только гипотезу о том, что
хочется проверить, но и альтернатива
гипотеза. Они ввели ошибку первого рода
и ошибка второго рода.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Ошибка первого рода — одна из двух ошибок ,
статистик сталкивается с гипотезой
тестирование: это тип ошибки, которая может возникнуть
при принятии решения, если нулевая гипотеза
истинный.

Если нулевая гипотеза неверна, другой тип
возникновения ошибки называется ошибкой типа II ,
что означает принятие нулевой гипотезы H0, когда она
является ложным.

Различные типы ошибок могут быть представлены в
следующей таблице:

Ситуация	Решение	
	Принимать ЧАС0	Отклонять ЧАС0
H0 правда	1 - а	α
H0 ложь	β	1 - β

где α – вероятность ошибки I рода
β – вероятность ошибки второго рода.

ПРИМЕРЫ

Проиллюстрируем на примере из
повседневной жизни, различные ошибки, которые можно
посвятить себя принятию решения.

Выходя утром из дома, мы
интересно какая будет погода. Если мы
думаем, что пойдет дождь, берем зонт.
Если мы думаем, что будет солнечно, мы не
взять что-нибудь на плохую погоду.
Поэтому мы сталкиваемся со следующим
гипотезы :

Нулевая гипотеза H0: собирается дождь.

Альтернативная гипотеза H1: будет
быть солнечным.

Предположим, мы принимаем гипотезу дождя и
возьми зонтик.

Если действительно идет дождь, мы поступили правильно
решение, но если солнечно, мы допустили ошибку:
ошибка принятия ложной гипотезы.
В обратном случае, если мы отбросим дождь
гипотеза, мы приняли правильное решение
если солнечно, но мы допустили ошибку
если идет дождь: ошибка отклонения истинной
гипотезы.

Мы можем представить эти различные типы
ошибки в следующей таблице:

Ситуация	Решение	
	Принимать ЧАС0 (Я беру зонтик)	Отклонять ЧАС0 (не беру зонт)
H0 правда (Идет дождь)	Хороший решение	Ошибка I типа
H0 ложь (солнечно)	Ошибка типа II β	Хорошая решение

Вероятность отклонения верной гипотезы называется
уровнем значимости.
Вероятность β принятия ложного
гипотеза - вероятность типа II
ошибка.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Уровень значимости
- Ошибка второго рода

T

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Нейман, Дж., Пирсон, Э.С.: Об использовании и интерпретации некоторых критериев тестирования для целей статистического вывода, Части I и II. Биометрика 20A, 175–240, 263–294 (1928).

Ошибка типа II

При проверке гипотезы ошибка II рода возникает при принятии нулевой гипотезы при условии, что она ложна, а β — вероятность принятия нулевой гипотезы H0 , когда она ложна:

$$\beta = P(\text{примем } H0 \mid H0 \text{ ложно}) .$$

ИСТОРИЯ

См. ошибку I рода.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Вероятность ошибки II рода обозначается как β, вероятность отклонения нулевой гипотезы , если она ложна, определяется как 1 β:

$$\beta = P(\text{отклонить } H0 \mid H0 \text{ ложно}) .$$

Вероятность 1 β называется мощностью теста.

Поэтому всегда необходимо пытаться свести к минимуму риск ошибки второго рода , что означает увеличение мощности проверки гипотезы.

ПРИМЕРЫ

См. ошибку I рода.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Ошибка I типа

Machine Translated by Google

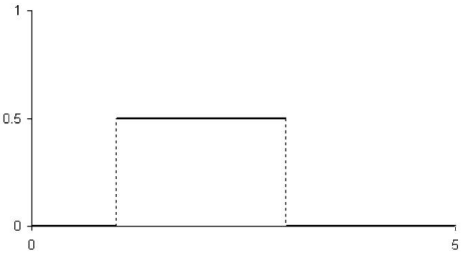
Равномерное распределение

Равномерное распределение используется для описания совокупность , равномерно распределенная на интервале.

Говорят, что случайная переменная X равномерно распределена в интервале [a, b], если ее функция плотности определяется выражением:

$$\phi(x) = \begin{cases} \frac{1}{б - а} & \text{если } а \quad x \quad б \\ 0 & \text{иначе .} \end{cases}$$

Равномерное распределение из-за прямоугольной формы функции плотности также иногда называют прямоугольным.



Равномерное распределение, а = 1, б = 3

Равномерное распределение является непрерывным распределение вероятностей.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Функция распределения равномерного распределение следующее:

$$\Phi(x) = \begin{cases} 0, \text{если } x \quad а \\ \frac{x - а}{б - а} & \text{если } а < x < б \\ 1 & \text{иначе .} \end{cases}$$

Ожидаемое значение случайной величины , равномерно распределенной в интервале [a, b], равно:

$$Э[X] = \int_{а}^{б} \frac{икс}{б - а} \, 2 \, дх = \frac{б2 - а2}{2(б - а)} \overset{"="}{} \frac{а + б}{2} .$$

Этот результат интуитивно очень легко понять, поскольку функция плотности симметрична относительно средней точки интервала [a, b].

Мы можем рассчитать

$$E[X2] = \int_{а}^{б} x2f(x) \, dx = \frac{b2 + ab + a2}{3} .$$

Таким образом , дисперсия равна :

$$Var(X) = E[X2] \quad (E[X])^2 \overset{"="}{} \frac{(б - а)^2}{12} .$$

Равномерное распределение является частным случаем бета-распределения , когда α = β = 1.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Применения равномерного распределения многочисленны при построении моделей для физические, биологические и социальные явления. Равномерное распределение часто используется как

Machine Translated by Google

модель ошибок аппроксимации при измерениях . Например, если длина некоторых объектов измеряется с точностью до 1 см, длина 35 см может представлять все объекты длиной от 34,5 до 35,5 см. Ошибки аппроксимации подчиняются равномерному распределению в интервале

[0,5, 0,5].

Равномерное распределение часто используется для генерации случайных чисел любого дискретного или непрерывного распределения вероятностей. Метод, используемый для создания непрерывного распределения, заключается в следующем.
Рассмотрим X как непрерывную случайную величину и F(x) его функция распределения. Предположим, что F непрерывна и строго возрастает.
Тогда U = F(X) — равномерная и непрерывная случайная величина. U принимает свои значения в интервале [0,1]. Тогда мы имеем X = F 1(U), где F 1 — обратная функция функции распределения.

Следовательно, если u1, u2,...,un ряд случайных чисел, то:

x1 = F 1(u1) , x2

= F 1(u2) ,

... ,

xn = F 1(un)

ряд случайных чисел, сгенерированный из распределения случайной величины X.

ПРИМЕРЫ

Если случайная величина X следует отрицательному экспоненциальному распределению среднего, равного к 1, то его функция распределения

F (x) знак равно 1 - exp (-x) .

Так как F(x) = u, то u = 1 e x.
Чтобы найти x, мы ищем F 1:

и = 1 - e-x ,

и - 1 = -e-x ,

1 - и = e-x ,

пер(1 - и) = -x ,

ln(1 u) = x .

Обратите внимание, что если u — случайная величина , равная равномерная на [0,1], то (1 u) также является случайной величиной , равномерной на [0,1]. Применяя вышеупомянутое к ряду случайных чисел, мы находим выборку , взятую из X:

U	Икс
0,14	1,97
0,97	0,03
0,53	0,63
0,73	0,31
⋮	⋮

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Бета-распределение
- Непрерывное распределение вероятностей
- Дискретное равномерное распределение

Унимодальное распространение

См. распределение частот.

Machine Translated by Google

B

Ценить

Значение – это количественная или качественная мера, связанная с переменной.

Мы говорим о наборе значений, принимаемых переменной. Значения могут быть количественными числами, если они связаны с количественной переменной или порядковыми номерами, если они связаны с качественной категориальной переменной.

ПРИМЕРЫ

Значения, принимаемые некоторой переменной, могут быть следующего типа: •

Рассмотрим количественную переменную «количество детей». Ценности ассоциируются к этой переменной относятся натуральные числа числа 0, 1, 2, 3, 4 и т. д.

• Рассмотрим качественную категориальную переменную «пол». Связанные значения категории «мужское» и «женское».

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Качественная категориальная переменная

Количественная переменная

Переменная

Переменная

Переменная – это измеримая характеристика которые относятся ко многим различным переменным.

Мы различаем много типов переменных: • Количественная переменная, из которых мы различаем:

– Дискретную переменную, например, результат тест

– Непрерывная переменная, например, вес или доход

• Качественная категориальная переменная, например, семейное положение

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Данные

Дихотомическая переменная

Качественная категориальная переменная

Количественная переменная

Случайная переменная

Ценить

Дисперсия

Дисперсия — это мера дисперсии распределения случайной величины.

Эмпирически дисперсия количественной переменной X определяется как сумма квадратов отклонений каждого наблюдения относительно среднее арифметическое , деленное на количество наблюдений.

ИСТОРИЯ

Дисперсионный анализ в том виде, в каком мы его понимаем и практикуем сегодня, был принципиальным.

Machine Translated by Google

разработан Фишером, Рональдом Эйлмером (1918, 1925, 1935). Он также ввел термины дисперсия и дисперсионный анализ.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Дисперсия обычно обозначается S2 , когда она относится к выборке , и σ2 , когда она относится к генеральной совокупности.
Мы также обозначаем дисперсию через Var(X), когда говорим о дисперсии случайной величины.

Пусть совокупность Noobservations относится к количественной переменной X. По определению, дисперсия совокупности рассчитывается следующим образом:

$$\sigma 2 = \frac{\sum_{я=1}^{H} (кси - м)^2}{H},$$

где N - размерность населения, а μ - среднее значение наблюдения:

$$м = \frac{\sum_{я=1}^{H} хи}{H}.$$

Когда наблюдения упорядочены в виде частотного распределения расчет дисперсии производится по следующей схеме:

манера мычать:

$$\sigma 2 = \frac{\sum_{я=1}^{к} fi \cdot (xi - \mu)2}{\sum_{я=1}^{к} фи},$$

где xi — различные значения переменной , fi — частоты, связанные с этими значениями, а k — число различных значений
уэс.

Для расчета дисперсии частотного распределения количественного X, где значения сгруппированы в классы, считаем, что все наблюдения, принадлежащие определенному классу, принимают значения центра класса

Оно верно только в том случае, если подтверждается гипотеза о том, что наблюдения равномерно распределены внутри каждого класса. Если это гипотеза не проверена, полученное значение дисперсии будет лишь приблизительным.

Для значений, сгруппированных в классы, имеем:

$$\sigma 2 = \frac{\sum_{я=1}^{к} fi \cdot (\delta i - \mu)^2}{\sum_{я=1}^{к} фи},$$

где δi — центры классов, fi — частоты, связанные с каждым классом, а k — количество классов.

Формула расчета дисперсии может быть изменена для уменьшения времени расчета ции и повысить точность. Таким образом, лучше рассчитывать дисперсию с помощью следующих формула мычания:

$$\sigma 2 = \frac{H \cdot \sum_{я=1}^{H} x2_я - \left(\sum_{я=1}^{H} хи\right)^2}{N2}.$$

Дисперсия соответствует центральному моменту второго порядка.

Дисперсия, измеренная на выборке, представляет собой оценку дисперсии генеральной совокупности. Рассмотрим выборку из n наблюдений относительно количественной переменной X и обозначим через $\bar{x}$ среднее арифметическое этой выборки:

$$\bar{x} = \frac{\sum_{я=1}^{н} хи}{н}.$$

Чтобы дисперсия выборки была оценкой без смещения дисперсии генеральной совокупности, ее необходимо вычислить путем деления на

Machine Translated by Google

(n 1), а не на n:

$$S2 = \frac{\sum_{я=1}^{н} (xi \quad \bar{x})^2}{п - 1} .$$

ПРИМЕРЫ

Пять студентов сдали два экзамена, заработав следующие оценки:

Экзамен 1:

3 544,5 3,5 4,5

Экзамен 2:

2,5 5,5 3,5 4,5 4

Обратите внимание, что среднее арифметическое $\bar{x}$ набора двух наборы наблюдений идентичны:

$$\bar{x}1 = \bar{x}2 = 4 \quad .$$

Разброс наблюдений вокруг среднее не то же самое .

Чтобы определить дисперсию, сначала вычислим отклонения каждого наблюдения от

среднее арифметическое, а затем возвести в квадрат отклонения:

Экзамен 1:

Оценка	(xi ик1)	(xi ик1)2
3,5	0,5	0,25
4	0,0	0,00
4,5	0,5	0,25
3,5	0,5	0,25
4,5	0,5	0,25
$\sum_{я=1}^{5} xi \quad \bar{x}1^2$	1,00	

Экзамен 2:

Оценка	xi ик2	xi ик2 2
2,5	1,5	2,25
5,5	1,5	2,25
3,5	0,5	0,25
4,5	0,5	0,25
4	0,0	0,00
$\sum_{я=1}^{5} xi \quad \bar{x}2^2$	5.00	

Разница оценок каждого экзамена равно:

Экзамен 1:

$$S2_1 = \frac{\sum_{я=1}^{5} (xi \quad \bar{x}1)^2}{п - 1} = \frac{1}{4} = 0,25$$

Экзамен 2:

$$S2_2 = \frac{\sum_{я=1}^{5} (xi \quad \bar{x}2)^2}{п - 1} = \frac{5}{4} = 1,25$$

Поскольку разброс оценок второго экзамена больше, на втором экзамене оценки более разбросаны вокруг среднего арифметического, чем на первом экзамене.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Мера рассеивания
  - Среднеквадратичное отклонение
- Дисперсия случайной величины

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Фишер, Р.А.: Корреляция между родственниками на основе предположения Менделя. наследование. Транс. Рой. соц. Эдинб. 52, 399–433 (1918)

Фишер, Р.А.: Статистические методы для Исследовательские работники. Оливер и Бойд, Эдин Бург (1925)

Фишер, Р.А.: План экспериментов. Оливер и Бойд, Эдинбург (1935)

Machine Translated by Google

Дисперсионный анализ

См. дисперсионный анализ.

Дисперсия случайной величины

Дисперсия количественной случайной величины измеряет значение среднего отклонения значений этой случайной вариации можно из математического ожидания (то есть из его среднего значения).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В зависимости от того, является ли случайная величина дискретной или непрерывной, говорят о дисперсии дискретной или непрерывной случайная переменная.

В случае, когда случайная величина dis creteX принимает n значений x1, x2, ... ,xn с относительными частотами f1, f2,...,fn, дисперсия определяется формулой:

$$\sigma 2 = \mathrm{Var}(X) = E\,(X - \mu)^2$$

$$= \sum_{i=1}^{n} f_i\,(x_i - \mu)^2 ,$$

где μ представляет собой математическое ожидание X.
Можно установить другую формулу, для которой расчет дисперсии производится таким образом:

$$\mathrm{Var}(X) = E\,(X - \mu)^2$$

$$= E\,X2 - 2\mu X + \mu 2$$

$$= E\,X2 - E\,[2\mu X] + E\,\mu 2$$

$$= E\,X2 - 2\mu E\,[X] + \mu 2$$

$$= E\,X2 - 2\mu 2 + \mu 2$$

$$= E\,X2 - (E\,[X])^2$$

$$= \sum_{i=1}^{n} x_i^2 f_i - \left(\sum_{i=1}^{n} x_i f_i\right)^2 .$$

На практике эта формула обычно лучше подходит для проведения расчетов.
Если случайная величина X непрерывна в интервале D выражение дисперсии принимает вид:

$$\sigma 2 = \mathrm{Var}(X) = E\,(X - \mu)^2$$

$$= \int_D (x - м)^2 f\,(x)\,dx .$$

Мы также можем преобразовать эту формулу теми же операциями, что и в дискретном случае, и тогда дисперсия будет равна:

$$\mathrm{Вар}(X) = \int_D x2 f\,(x)\,dx - \left(\int_D x\phi(x)dx\right)^2 .$$

Свойства дисперсии 1. Пусть a и b — две константы, а X а побежал. переменная дом:

$$\mathrm{Var}(aX + b) = a2 \cdot \mathrm{Var}(X) .$$

2. Пусть X и Y — две случайные величины:

$$\mathrm{Вар}(X + Y) = \mathrm{Вар}(X) + \mathrm{Вар}(Y) + 2\mathrm{Cov}\,(X, Y),$$

$$\mathrm{Вар}\,(X - Y) = \mathrm{Вар}\,(X) + \mathrm{Вар}\,(Y) - 2\mathrm{Cov}\,(X, Y),$$

где Cov (X, Y) — ковариация между X и Y. В частности, если X и Y независимы:

$$\mathrm{Вар}(X + Y) = \mathrm{Вар}(X) + \mathrm{Вар}(Y)$$

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

Machine Translated by Google

ПРИМЕРЫ

Рассмотрим два примера, один из которых касается дискретная случайная величина, другая — непрерывная случайная величина.

Бросаем кубик несколько раз подряд. Предположим, мы выиграли 1 евро, если результат четный и 2 евро, если результат 1 или 3, и что мы проигрываем 3 евро, если результат 5.
Случайная величина X описывает количество евро выиграли или проиграли. В следующей таблице представлены различные значения X и их соответствующие вероятности:

Икс	3	1	2
П (Икс)	1/6	3/6	2/6

Дисперсия дискретной случайной величины X равно:

$$\begin{aligned}\operatorname{Var}(X) &= \sum_{i=1}^{3} \text{исправить}_i^2 \cdot \text{фикси} \\ &= 2^2 \cdot \tfrac{1}{6} + (-1)^2 \cdot \tfrac{3}{6} + (-2)^2 \cdot \tfrac{2}{6} \\ &\quad - \left(3 \cdot \tfrac{1}{6} + 1 \cdot \tfrac{3}{6} + 2 \cdot \tfrac{2}{6}\right)^2 \\ &= \frac{26}{9}.\end{aligned}$$

Рассмотрим непрерывную случайную величину X функция плотности которой равномерна на (0, 1):

$$\phi(x) = \begin{cases} 1 \text{ для } 0 < x < 1 \\ 0 \text{ иначе .} \end{cases}$$

Математическое ожидание этой случайной величины равно $\frac{1}{2}$.
Мы можем вычислить дисперсию случайного переменная X:

$$\begin{aligned}\operatorname{Var}(X) &= \int_0^1 x^2 f(x)\,dx - \left(\int_0^1 x\phi(x)\,dx\right)^2 \\ &= \int_0^1 x^2 \cdot 1\,dx - \left(\int_0^1 x \cdot 1\,dx\right)^2 \\ &= \left[\frac{x^3}{3}\right]_0^1 - \left[\frac{x^2}{2}\right]_0^1{}^2 \\ &= \frac{1}{3} - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{12}.\end{aligned}$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Ожидаемое значение
- Случайная переменная

Матрица дисперсии-ковариации

Матрица дисперсии-ковариации V между n переменных X1,...,Xn — матрица порядка (n × n), где диагональ содержит дисперсии каждой переменной Xi и вне диагонали – это ковариации между пары переменных (Xi, Xj) для каждого i = j. Для n переменных, мы вычисляем n дисперсий и n(n 1) ковариации. В результате матрица V является матрицей размерности n × n. В литре природе мы часто сталкиваемся с этой матрицей при другое название: дисперсионная матрица или ковариационная матрица.

Если переменные стандартизированы (см. стандартизированные данные), дисперсии переменных равны 1, а ковариации становятся корреляции, и поэтому матрица становится корреляционная матрица.

ИСТОРИЯ

См. дисперсию и ковариацию.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть X1,...,Xn — n случайных величин.
Матрица дисперсии-ковариации представляет собой матрицу порядка (n × n); он симметричен и содержит

B

Machine Translated by Google

по диагонали — дисперсии каждой переменной, а в качестве других членов — ковариации между переменными. Матрица V _

Низкая матрица:

$$\begin{matrix} \text{Вар}(X1) & \text{Cov}(X1, X2) \cdots & \text{Cov}(X1, Xn) \\ \text{Ков}(X2, X1) & \text{Вар}(X2) & \cdots \text{Cov}(X2, Xn) \\ \vdots & \vdots & \ddots \vdots \\ \text{Cov}(Xn, X1) & \text{Cov}(Xn, X2) & \cdots \text{Var}(Xn) \end{matrix},$$

с

$$\text{Var}(Xi) = E\,X2_{я} \quad (Э\,[Си])^2 \quad \text{и}$$

$$\text{Cov}\,Xi, Xj = E\,(Xi \quad E[Xi])\,Xj \quad E[Xj]\,.$$

ПРИМЕРЫ

Рассмотрим следующие три переменные

X1, X2 и X3 приведены в следующей таблице:

Икс1	Икс2	Икс3
1	2	3
2	3	4
1	2	3
5	4	3
4	4	4

В случае n = 3 каждая случайная величина состоит из пяти слагаемых, которые мы обозначим N = 5. Мы хотим рассчитать дисперсию – ковариационная матрица V этих переменных. Мы выполните следующие шаги.

Мы оцениваем математическое ожидание каждой переменной по ее эмпирическому среднему $\overline{xi}$ и получаем следующие результаты:

$$\overline{X1} = 2{,}6$$

$$\overline{X2} = 3$$

$$\overline{X3} = 3{,}4\,.$$

Дисперсия для переменной Xj определяется как:

$$\frac{\sum_{я=1}^{H} (\text{хдж} - \overline{\text{хдж}})^2}{4}\,.$$

Ковариация между двумя переменными Xr и Xs рассчитывается следующим образом:

$$\frac{\sum_{я=1}^{H} (\text{xir} - \overline{\text{xr}})\,(\text{xis} - \overline{\text{xs}})}{4}\,.$$

Затем мы должны вычесть среднее Xi $\overline{\text{каждой}}$ переменной Xi. Получаем новую переменную Yi $\overline{= \text{Xi}}$ Xi.

Д1	Д2	Д3
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

Обратите внимание, что среднее значение каждой новой переменной Yi равно 0. Мы находим эти результаты в матричной форме:

$$Y = \begin{matrix} 1{,}6 & 1 & 0{,}4 \\ 0{,}6 & 0 & 0{,}6 \\ 1{,}6 & 1 & 0{,}4 \\ 2{,}4 & 1 & 0{,}4 \\ 1{,}4 & 1 & 0{,}6 \end{matrix}\,.$$

Из этой матрицы мы вычисляем матрицу Y · Y, симметричную матрицу, в следующем виде:

$$\begin{matrix} \sum_{я=1}^{H} y2_{1} & \sum_{i=1}^{H} yi2 \cdot yi1 & \sum_{i=1}^{H} yi3 \cdot yi1 \\ \sum_{i=1}^{H} yi2 \cdot yi1 & \sum_{я=1}^{H} y2_{2} & \sum_{i=1}^{H} yi3 \cdot yi2 \\ \sum_{i=1}^{H} yi3 \cdot yi1 & \sum_{i=1}^{H} yi3 \cdot yi2 & \sum_{я=1}^{H} y2_{3} \end{matrix}\,.$$

Эта матрица содержит значения суммы произведений квадратов каждой переменной Yi (по диагонали) и перекрестного произведения каждой пары переменных (Yi, Yj) на остальную часть матрицы. Для исходных переменных Xi следующая матрица содержит на диагонали сумму квадратов отклонений от среднего для

Machine Translated by Google

каждой переменной, а на внешних условиях перекрестные произведения этих отклонений.

Численно получаем:

$$\begin{matrix} 13.270.8 \\ 7\ 41 \\ 0{,}8\ 11{,}2 \end{matrix} .$$

Среднее значение суммы квадратов отклонений от среднего значения случайной величины Xi из N терминов называется дисперсией . Значение скрещенных произведений этих отклонений называется ковариацией . Таким образом, чтобы получить дисперсионно-ковариационную матрицу V переменной

Если X1,...,Xn, мы должны разделить элементы матрицы Y · Y на N 1, что математически равносильно:

$$\begin{matrix} \text{н} & \text{н} & \text{н} \\ \frac{\sum_{i=1}^{N} y2_{i1}}{N\ 1} & \frac{\sum_{i=1}^{N} yi2 \cdot yi1}{N\ 1} & \frac{\sum_{i=1}^{N} yi3 \cdot yi1}{N\ 1} \\ \text{н} & \text{н} & \text{н} \\ \frac{\sum_{i=1}^{N} yi2 \cdot yi1}{N\ 1} & \frac{\sum_{i=1}^{N} y2_{i2}}{N\ 1} & \frac{\sum_{i=1}^{N} yi3 \cdot yi2}{N\ 1} \\ \text{н} & \text{н} & \text{н} \\ \frac{\sum_{i=1}^{N} yi3 \cdot yi1}{N\ 1} & \frac{\sum_{i=1}^{N} yi3 \cdot yi2}{N\ 1} & \frac{\sum_{i=1}^{N} y2_{i3}}{N\ 1} \end{matrix} .$$

Факт деления на N 1, а не на N для получения среднего связан с тем, что мы, как правило, работаем с выборкой случайных величин, не зная реальной совокупности, и для которой мы делаем несмещенную оценку дисперсии и ковариации.

Численно получаем:

$$B = \begin{matrix} 3{,}3 & 1{,}75 & 1{,}75 \\ 1{,}75 & 1 & 0{,}2 \\ 0{,}2 & 0{,}25 & 0{,}3 \end{matrix} .$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Коэффициент корреляции
- Ковариация
- Стандартизированные данные
- Дисперсия

Вектор

Назовем вектор размерности n матрицей (n × 1) , то есть таблица, имеющая только один столбец и n строк. Элементы в этом таблицы называются коэффициентами или векторными ком компоненты.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть x = (xi) и y = (yi) для i = 1, 2,...,n — два вектора размерности n.

Мы определяем:

- Сумма двух векторов, составленная покомпонентно:

$$x + y = (xi) + (yi) = (xi + yi) .$$

- Произведение вектора x на скаляр k:

$$к \cdot x = (к \cdot xi),$$

где каждая компонента умножается на k. • Скалярное произведение x и y :

$$x \cdot y = (xi) \cdot (yi) = \sum_{i=1}^{\text{н}} xи \cdot йи .$$

• Норма x : _

$$x = \sqrt{x \cdot x}$$

$$x = \sqrt{\sum_{i=1}^{\text{н}} x2_i} .$$

Мы называем единичный вектор вектором нормы (или длины). 1. Мы получаем единичный вектор, деля вектор на его норму. Так, например,

$$\frac{1}{\text{Икс}} (xi)$$

является единичным вектором, имеющим то же направление как x.

Machine Translated by Google

• i -й вектор канонических базисов в пространство n измерений определяется вектором измерения n , имеющим «1» в строке i и «0» для других составных частей. нц.

• Вектор — это матрица определенной размерности (n × 1). Можно умножить вектор на матрицу (m × n) в соответствии с правилами умножения, определенными для матрицы. сек.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Вектор можно рассматривать как столбец числа берс, откуда термин вектор-столбец, точно так же, как мы используем термин вектор-линия для обозначения транспонирования вектора .

Таким образом, скалярное произведение двух векторов соответствует вектору-строке, умноженному на вектор-столбец.

ПРИМЕРЫ

В трехмерном пространстве рассмотрим векторы:

$$x = \begin{matrix}1\\2\\3\end{matrix} \quad \text{и} \quad y = \begin{matrix}0\\1\\0\end{matrix},$$

где y — второй вектор канонического базиса.

– Норма x определяется по формуле :

$$\|x\|_2 = \begin{matrix}1\\2\\3\end{matrix} \cdot \begin{matrix}1\\2\\3\end{matrix} = [1\ 2\ 3] \cdot \begin{matrix}1\\2\\3\end{matrix} = 12 + 22 + 32 = 1 + 4 + 9 = 14,$$

откуда $x = \sqrt{14} \approx 3{,}74.$

– Таким же образом получаем y = 1.

– Сумма двух векторов x и y равна:

$$x + y = \begin{matrix}1\\2\\3\end{matrix} + \begin{matrix}0\\1\\0\end{matrix} = \begin{matrix}1+0\\2+1\\3+0\end{matrix} = \begin{matrix}1\\3\\3\end{matrix}.$$

– Умножение вектора x на скалярное произведение 3 дает:

$$3 \cdot x = 3 \cdot \begin{matrix}1\\2\\3\end{matrix} = \begin{matrix}3\cdot 1\\3\cdot 2\\3\cdot 3\end{matrix} = \begin{matrix}3\\6\\9\end{matrix}.$$

– Скалярное произведение x и y равно:

$$x \cdot y = \begin{matrix}1\\2\\3\end{matrix} \cdot \begin{matrix}0\\1\\0\end{matrix} = [1\ 2\ 3] \cdot \begin{matrix}0\\1\\0\end{matrix} = 1 \cdot 0 + 2 \cdot 1 + 3 \cdot 0 = 2.$$

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Матрица
- Транспонировать

Фон Мизес, Рихард

Фон Мизес Рихард родился в 1883 году в Лемберге в Австро-Венгерской империи (ныне Львов, Украина) и умер в 1953 году.

Machine Translated by Google

Мизес изучал машиностроение в Техническом университете в Вене до 1906 года.
Он стал ассистентом Георга Хамеля, профессора механики, в Брюэнне (ныне Брно, Чехия), где в 1908 году получил Venia Legendi. Затем он работал адъюнкт-профессором прикладной математики. После Первой мировой войны, в 1920 г., он стал директором Института прикладной математики в Берлине.

Фон Мизес, Рихард известен прежде всего своими работами по основам вероятности и статистики, которые были реабилитированы в 1960-е годы. Он основал прикладную школу математики и написал свою первую работу по философский позитивизм в 1939 году.

Избранные основные работы и статьи фон Мизес, Ричард:

1928 Wahrscheinlichkeit, Statistik und Wahrheit. Спрингер, Берлин, Гейдельберг, Нью-Йорк.

1939 г. Вероятность, статистика и правда, пер. Нейман Дж., Шолль Д. и Рабинович Э. Ходж, Глазго.

Machine Translated by Google

Взвешенное среднее арифметическое

Взвешенное среднее арифметическое является мерой центральной тенденции набора количественных наблюдений, когда не все наблюдатели

ции имеют одинаковое значение.

Мы должны присвоить вес каждому наблюдению в зависимости от его важности по сравнению с другими наблюдениями.

Средневзвешенное арифметическое равно сумме наблюдений, умноженной на их веса, деленной на сумму их весов.

ИСТОРИЯ

Взвешенное среднее арифметическое было введено Котесом Роджером в 1712 году. Его работа была опубликована в 1722 году, через шесть лет после его смерти.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ Пусть

x1, x2,...,xn — набор из n величин или n наблюдений относительно количественной переменной X , которой мы присваиваем веса

w1, w2,...,wn.

Средневзвешенное арифметическое равно:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i}.$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Средневзвешенное арифметическое в настоящее время используется в экономике, особенно в индексах потребительских цен и цен производителей и т. д.

ПРИМЕРЫ

Предположим, что во время курса обучения в компании управление было составлено из трех относительно разных требований разной важности:

Индивидуальный проект	30%
Промежуточный экзамен	20%
Заключительный экзамен	50%

Каждый учащийся получает оценку по шкале от 1 до 10. Один учащийся получает 8 баллов за свою индивидуальную работу.

визуальный проект, 9 на промежуточном экзамене и 4 на выпускном экзамене.

Его итоговая оценка рассчитывается по средневзвешенному арифметическому:

$$\bar{x} = \frac{(8 \cdot 30) + (9 \cdot 20) + (4 \cdot 50)}{= 6{,}2} = \frac{. 30 + 20}{+ 50}$$

В этом примере веса соответствуют относительной важности различных требований курса.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Среднее арифметическое

  Иметь в виду

- Мера центральной тенденции

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Котес, Р.: Aestimatio Errorum in Mixta Mathesi, per вариации partium Trian guli plani et sphaerici. В: Смит, Р. (ред.) Opera Miscellania, Кембридж (1722 г.)

Взвешенный метод наименьших квадратов

Метод

Метод взвешенных наименьших квадратов применяют, когда дисперсия ошибок непостоянна, то есть когда нарушается следующая гипотеза метода наименьших квадратов: дисперсия ошибок постоянна (равна неизвестному значению σ2) для любого наблюдения i (то есть независимо от значения соответствующего xij). Таким образом, вместо того, чтобы для каждого i = 1,...,n иметь Var(εi) = σ2, мы имеем:

$$\mathrm{Var}(\varepsilon i) = \sigma 2wi\ ,$$

где веса wi > 0 могут быть разными для каждого i = 1,...,n.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

В матричной форме имеем модель

$$Y = X\beta + \varepsilon \quad ,$$

где Y — вектор (n × 1) наблюдений относительно зависимой переменной (n наблюдений), β — вектор (p × 1) оцениваемых параметров, ε — вектор (n × 1) ошибок , и

$$X = \begin{matrix} 1\ X11 \ldots X1(p\ \ 1) \\ \vdots\ \ \vdots \qquad\quad \vdots \\ 1\ Xn1 \ldots Xn(p\ \ 1) \end{matrix}$$

– матрица (n × p) плана относительно независимых переменных. Кроме того, у нас есть:

$$\mathrm{Var}(\varepsilon) = \sigma 2V \quad ,$$

где

$$B = \begin{matrix} 1/w1\ 0 \cdots 0\ 1/w2 & & 0 \\ \ldots & & 0 \\ \vdots \qquad \vdots & \ldots & \vdots \\ 0 \qquad 0 \cdots 1/\text{вн} & & \end{matrix} .$$

Заявление

$$Yw = WY \quad ,$$

$$X\text{в} = BX \quad ,$$

$$\varepsilon w = W\varepsilon \quad ,$$

с

$$\text{Вт} = \begin{matrix} \overline{w1\ 0\ 0} & \cdots & 0 \\ w2 \cdots & & 0 \\ \vdots \qquad \vdots & \ldots & \vdots \\ 0 \qquad 0 & \ldots & \overline{\text{вн}} \end{matrix} ,$$

таких как WW = V 1, мы получаем эквивалент альтернативная модель:

$$Yw = X\text{в}\beta + \varepsilon w\ ,$$

где

$$\mathrm{Var}(\varepsilon w) = \mathrm{Var}(W\varepsilon) = W\mathrm{Var}(\varepsilon)\ W = \sigma 2 \cdot WVW = \sigma 2In\ ,$$

где In – единичная матрица размерности n. Поскольку для этой новой модели дисперсия ошибок постоянна, используется метод наименьших квадратов. Вектор оценок:

$$\begin{aligned} \beta\hat{}\ \text{ш} &= X\text{в}Xw\ _^{1}\ X\ wYw \\ &= XW\ WX \quad 1\ XW\ WY \\ &= XV \quad 1X \quad {}^{1}\ X\ V\text{-}1Y\ . \end{aligned}$$

Векторы оценочных значений для Yw = WY становится:

$$\begin{aligned} Yw &= X\text{в}\beta\hat{}w\ _ \\ &= WX\ X\ V \quad 1X \quad {}^{1}\ X\ V\text{-}1Y\ . \end{aligned}$$

Machine Translated by Google

Мы можем получить вектор оценочных значений для установки Y = W 1Yw :

$$\begin{aligned} Y &= W \quad 1Yw \\ &= W \quad 1WX\,X\,V \quad 1X \qquad\qquad X\,V\text{-}1Y \\ &= X\hat{\beta}_{ш}\,. \end{aligned}$$

Дисперсия σ2 оценивается по формуле:

$$\begin{aligned} {}^{2}c_{ж} &= \frac{Yw - Yw\;Yw - Yw}{п - p} \\ &= \frac{Y - YWWY - Y}{п - p} \\ &= \frac{\sum_{я=1}^{н} ви\,(уи - уи)^{2}}{п - p}\,. \end{aligned}$$

В случае простой регрессии из выборки из n наблюдений (xi, yi) и весов wi мы имеем:

$$X\,V\text{-}1X = \frac{\sum_{я=1}^{н} с_{виси} \sum_{я=1}^{н} виси}{\sum_{я=1}^{н} виси \sum_{я=1}^{н} wix2_{я}},$$

$$XV\text{-}1Y = \frac{\sum_{я=1}^{н} я = 1}{\sum_{я=1}^{н} висии},$$

и, таким образом, если мы положим $\hat{\beta}_{w} = \begin{matrix} \beta\,0\,Вт \\ \beta\,1\,Вт \end{matrix}$, мы нашли:

$$\beta\,0w = \frac{\sum_{я=1}^{н} wx2_{я} \sum_{я=1}^{н} виси \sum_{я=1}^{н} висии}{\sum_{я=1}^{н} виси \sum_{я=1}^{н} wix2_{я}} - \frac{\sum_{я=1}^{н} виси \sum_{я=1}^{н} висии}{\sum_{я=1}^{н} виси \sum_{я=1}^{н} wix2_{я}},$$

$$\beta\,1w = \frac{\sum_{я=1}^{н} с_{виксии - я} \sum_{я=1}^{н} виси \sum_{я=1}^{н} я = 1}{\sum_{я=1}^{н} виси \sum_{я=1}^{н} wix2_{я}}.$$

С учетом взвешенного означает:

$$\bar{x}_{ш} = \frac{\sum_{я=1}^{н} виси}{\sum_{я=1}^{н} я}, \qquad \bar{y}_{в} = \frac{\sum_{я=1}^{н} вийи}{\sum_{я=1}^{н} я},$$

у нас есть оценщики:

$$\beta\,1w = \frac{\sum_{я=1}^{н} wixiyi - wixwyw \sum_{я=1}^{н} я = 1}{\sum_{я=1}^{н} wix2_{я} - wix2_{ж}}$$

$$= \frac{\sum_{я=1}^{н} wi\,(xi \quad xw)\,yi \quad yw}{\sum_{я=1}^{н} ви\,(кси - хв)^{2}}$$

$$\beta\,0w = \bar{yw} \quad \beta\,1w\bar{xw}$$

которые сильно напоминают оценки методом наименьших квадратов, за исключением того, что мы придаем больший вес наблюдениям с большим wi (то есть наблюдениям, для которых ошибки имеют меньшую дисперсию в исходной модели). Расчетные значения даны по формуле:

$$yi = \beta\,0w + \beta\,1wxi = \bar{yw} + \beta\,1w\,(xi \quad \bar{xw})$$

а дисперсия σ2 оценивается по формуле:

$${}^{2}c_{ш} = \frac{\sum_{я=1}^{н} ви\,(йи - и)^{2}}{п - 2}$$

$$= \frac{\sum_{я=1}^{н} wi\,yi \quad \bar{yw}\,i=1^{2} \quad 2\,\beta\,1w\,i=1 wi(xi \quad \bar{xw})^{2}}{п - 2}.$$

Примеры весов wi, выбранных в простой регрессии, даны как wi = 1/xi в случаях, когда дисперсия εi пропорциональна xi, или когда wi = 1/x2 , когда дисперсия εi пропорциональна x2 .

я .

Вт

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Самый большой недостаток взвешивания метод наименьших квадратов, который многие предпочитают не знать о, вероятно, факт, что он основан на предположении, что вес известен точно. Это почти на практике это никогда не происходит, и вместо этого используются оценочные веса. Как правило, это трудно оценить эффект от использования ориентировочный вес, но опыт показывает что результаты большинства регрессионных анализов не очень чувствительны к используемому весу. Преимущества взвешенного анализа часто получают в больших масштабах, хотя и не всегда целиком, с приблизительными весами. Это Важно знать об этой проблеме и избегать ее, а также использовать только те веса, которые можно точно оценить относительно друг друга. Взвешенная регрессия наименьших квадратов, как другие методы наименьших квадратов также чувствительны к выбросам.

ПРИМЕРЫ

Рассмотрим набор данных, содержащий 10 наблюдений с объясняющей переменной X , принимающей значения Xi = i c i = 1, 2,..., 10. Переменная Y генерируется с использованием модели:

$$yi = 3 + 2xi + \varepsilon i ,$$

где εi нормально распределены с E (εi) = 0 и Var(εi) = (0,2xi) 2. Таким образом, дисперсия ошибки первого наблюдения соответствует [(1) (0,2)] ^{2} = 0,04, как и те из десятого наблюдения соответствуют [(10) (0,2)] ^{2} = 4. Сгенерированные таким образом данные представлены в следующей таблице:

Значения xi и yi, сгенерированные моделью yi = 3 + 2xi + εi

я	xi	йи
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

Сначала мы выполняем невзвешенную регрессию. Получаем уравнение регрессии:

$$yi = 3,49 + 2,09xi ,$$

β 0 = 3,49 при tc = 0,99 β 1 = 2,09 , при tc = 3,69 .

Дисперсионный анализ

Источник вариация	Степени бесплатно дом	Сумма квадраты	Среднее значение квадраты	Ф
Регрессия	1	360,68	360,68	13,59
Рези двойной	8	212,39	26,55	
Общий	9	573.07		

R2 этой регрессии составляет 62,94 %. Сначала взгляд, результаты кажутся хорошими. Оценочный коэффициент, связанный с объясняющей переменной, явно значим. Этот результат совсем не удивителен, потому что тот факт, что дисперсия ошибок не кон константа редко влияет на оцениваемые коэффициенты. Средний квадрат остатков трудно объяснить, учитывая, что нет единственного отклонение от оценки. Более того, стандарт ошибки и длины доверительных интервалов при 95 % для расчетных условных средние относительно постоянны для всех наблюдений. Невзвешенный анализ не

Machine Translated by Google

учитывать наблюдаемые неравенства между дисперсиями ошибок.

Теперь мы делаем взвешенную регрессию, используя в качестве весов значения 1/x2. Веса известны, потому что они пропорциональны реальным дисперсиям, которые равны (0,2x) 2. Результаты следующие:

$$yi = 2,53 + 2,28xi ,$$

$$\beta 0 = 2,53 \text{ при } tc = 3,07 ,$$

$$\beta 1 = 2,28 \text{ при } tc = 7,04 .$$

Дисперсионный анализ

Источник из вариация	Степени свободных дом	Сумма квадраты	Среднее значение квадраты	Ф
Регрессия	1	23,29 23,29		49,6
Остаток 8		3,75	0,47	
Общий	9	27.05		

R2 этой регрессии составляет 86,11 %. На этом основании можно сделать следующие комментарии:

- Все числа, относящиеся к сумме квадратов зависимой переменной, подвержены влиянию весов и несопоставимы с числами, полученными с помощью невзвешенной регрессии.
- Тест Фишера и R2 можно сравнить между двумя моделями. В нашем примере два значения больше, чем невзвешенная регрессия, но это не всегда так.
- Суммарные коэффициенты относятся близки к показателям невзвешенной регрессии. Так бывает часто, но не повсеместно.
- Стандартные ошибки и доверие между валы для условных средних отражают тот факт, что точность оценок уменьшается с увеличением значения x.

Этот результат является первой причиной для использования взвешенная регрессия.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Обобщенная линейная регрессия

Наименьших квадратов

Регрессивный анализ

ССЫЛКИ

Seber, GAF (1977) Линейный регрессионный анализ. Уайли, Нью-Йорк

Уилкоксон, Фрэнк

Уилкоксон, Франк (1892–1965) получил степень магистра химии в Университете Рутгерса в 1921 году и докторскую степень по физике в 1924 году в Корнельском университете, где он также работал над докторской диссертацией с 1924 по 1925 год. С 1928 по 1929 год он работал в Nichols Copper Company в Маспете в Квинсе, штат Нью-Йорк, а затем был нанят Институтом исследований растений Бойса Томпсона в качестве руководителя группы, изучающей действие инсектицидов и фунгицидов. Он остается в институте до начала

Вторая мировая война; в это время и в течение следующих двух лет он перешел в Atlas Powder Company в Огайо, где руководил контрольной лабораторией.

В 1943 году Франк Уилкоксон был назначен руководителем группы в лаборатории American Cyanamid Company (Стэмфорд, Коннектикут), изучавшей инсектициды и фунгициды. Затем у него появился интерес к статистике: в 1950 году его перевели в отдел Lederle Laboratories.

компании Cyanamid в Перл-Ривер, штат Нью-Йорк, где он организовал консультационную группу по статистике. С 1960 года до своей смерти в 1965 году преподавал

Machine Translated by Google

прикладная статистика в Департаменте статистики для студентов, изучающих естественные науки в штате Флорида. Университет.

Избранные основные работы и статьи Фрэнк Уилкоксон:

1945 Индивидуальные сравнения по рейтингу методы. Биометрия 1, 80–83.

1957 Некоторые быстрые приближенные статистические процедуры. американский цианамид, Стэмфордские исследовательские лаборатории, Стэмфорд, Коннектикут.

Подписанная таблица Уилкоксона

Таблица со знаком Уилкоксона дает теоретические значения статистического T знака Уилкоксона испытание , примененное к двум парным образцам, на гипотеза о том, что две популяции следуют такое же распределение.

ИСТОРИЯ

Оуэн, Д.Б. (1962) опубликовал Уилкоксона таблица со знаком для n 20 (где n — количество парных наблюдений) для критерия Уилкокса на знаке, разработанного Уилкоксоном, Ф. (1945).
Обычно используется таблица Хартера. HL и Owen, DB (1970) созданы для n 50.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим набор из n пар наблюдений ((X1, Y1),...,(Xn, Yn)). Мы рассчитываем абсолютные разности |Xi Yi| к которому мы присваиваем ранг (то есть мы связываем ранг 1 с наименьшее значение, ранг 2 до следующего наивысшего значение и так далее, вплоть до ранга n для великого максимальное значение). Тогда придаем этому рангу знак соответствующей разности Xi Yi. Таким образом вычисляем сумму положительных рангов:

$$T = \sum_{i=1}^{k} R_i ,$$

где k — количество пар, представляющих положительная разность (Xi Yi > 0).
Для различных значений m и α (уровень значимости) таблица Уилкоксона со знаком дает теоретические значения статистики T знаковый критерий Уилкоксона, основанный на гипотезе о том, что две лежащие в основе популяции одинаково распределены.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Таблица Уилкоксона со знаком используется в непараметрических тестах, которые используют ранг и особенно в подписанном тесте Уилкоксона.

ПРИМЕРЫ

Ниже приведена выдержка из письма Уилкоксона, подписанного таблица m от 10 до 15 и α = 0,05 до α = 0,975:

м	a = 005	α = 0975
10	11	46
11	14	55
12	18	64
3	22	73
14	26	83
15	31	94

Для примера того, как Уилкоксон подписал используется таблица, см. подписанную таблицу Уилкоксона.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Статистическая таблица
- Подписанный тест Уилкоксона

Machine Translated by Google

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Хартер, Х.Л., Оуэн, Д.Б.: Избранные таблицы по математической статистике, том. 1. Маркхэм, Чикаго (5.7 Приложение) (1970)

Оуэн, Д.Б.: Справочник по статистическим таблицам. (3.3, 5.7, 5.8, 5.10). Аддисон-Уэсли, Рединг, Массачусетс (1962)

Уилкоксон, Ф.: Индивидуальные сравнения методами ранжирования. Биометрия 1, 80–83 (1945) (5.1, 5.7).

Подписанный тест Уилкоксона

Знаковый критерий Уилкоксона является непараметрическим критерием. В качестве знакового теста он служит для проверки любых различий между двумя парами наблюдений. Он сообщает, положительна или отрицательна разница между двумя парами наблюдений, принимая во внимание амплитуду этой разницы. вс.

ИСТОРИЯ

Знак Уилкоксона был создан Ф. Уилкоксоном в 1945 г. В 1949 г. он предложил другой метод обработки нулевых разностей.

которое может появиться между двумя наблюдениями рассматриваемых выборок.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть (X1, X2,...,Xn) и (Y1, Y2,...,Yn) — две выборки размера n. Рассмотрим n пар наблюдений (x1, y1), (x2, y2),...,(xn, yn).
Обозначим через |di| разница в абсолютном значении между xi и yi:

$$|\text{ди}| = |\text{yi} \quad \text{xi}|, \qquad \text{для } i = 1, 2,\ldots,n.$$

Вычитаем из размерности все пары наблюдений, не представляющие разности (di = 0). Обозначим через m число оставшихся пар. Мы присваиваем затем пары наблюдений ранжируются от 1 до m в зависимости от размера |di|, то есть мы присваиваем ранг 1 наименьшему значению |di| и ранжируем m до наибольшего значения |di|. Назовем Ri, i = 1,...,m определенным таким образом рангом. Если несколько пар переменных представляют одну и ту же абсолютную разность |di|, им присваивается средний ранг. При наличии средних рангов статистика теста рассчитывается по следующему выражению:

$$T1 = \frac{\sum_{\text{я}=1}^{\text{м}} \text{Ри}}{\sqrt{\sum_{\text{я}=1}^{\text{м}} R2_{\text{я}}}},$$

где Ri — ранг пары (xi, yi) без знака разности di.

При отсутствии средних рангов правильнее использовать сумму рангов, связанных с положительной разностью:

$$T = \sum_{\substack{\text{(для i}\\ \text{таких, как di} > 0)}} \text{Ри}\ .$$

Гипотезы

Гипотезы, лежащие в основе знакового теста Уилкоксона, основаны на dk, медиане di. В зависимости от того, является ли это двусторонним тестом или односторонним тестом, мы имеем следующее случаи:

A: Двусторонний тест:

$$H0 : \text{дк} = 0$$

$$H1 : \text{дк} = 0$$

B: Односторонний тест:

$$H0 : \text{дк} \quad 0$$

$$H1 : \text{дк} > 0$$

C: Односторонний тест:

$$H0 : \text{дк} \quad 0$$

$$H1 : \text{дк} < 0$$

Вт

Machine Translated by Google

В случае A гипотеза представляет собой ситуацию, когда нет различий между две популяции.

В случае B мы выдвигаем гипотезу о том, что значения совокупности X имеют тенденцию быть больше, чем значения совокупности Y.

В случае C мы выдвигаем противоположную гипотезу, т. е. что значения совокупности X имеют тенденцию быть меньше, чем значения совокупности Y.

Правила принятия

решений Если используется T и если m < 50, то для проверки нулевой гипотезы H0 будет использоваться таблица Уилкокса на подписанной таблице .

Случайная величина T представляет собой линейную комбинацию нация m независимых бернуллиевских случайных величин (но не одинаково распределенных). Он имеет как среднее значение , так и стандартное отклонение:

$$м = \frac{м (м + 1) 4}{} \quad и$$

$$\sigma = \sqrt{\frac{}{м (м + 1) (2м + 1) 24}} .$$

Выводим случайную величину Z:

$$Z = \frac{T - мю}{о} .$$

Обозначим через wα значение таблицы со знаками Вилкоксона с параметрами m и α (где α — уровень значимости).

Если m > 20, то wα можно аппроксимировать с помощью соответствующих значений нормального распределения:

$$w\alpha = \frac{м (м + 1) 4}{} + за\alpha \sqrt{\frac{}{м (м + 1) (2м + 1) 24}} ,$$

где zα — значения в нормальной таблице на уровне α.

Если используется T1 , мы будем основывать наш расчет непосредственно на обычной таблице, чтобы проверить нулевое значение гипотезаH0. Правила принятия решений различаются в зависимости от гипотез. У нас есть решающие правила A, B и C относительно предыдущих случаев A, B и C. Если используется T1 , мы должны заменить T на T1 и wα на zα в следующих правилах.

Случай А

Мы отвергаем H0 на уровне значимости α, если T больше w1 α/2 или меньше wα/2, т. е. если

$$T < w\alpha/2 \text{ или } T > w1 \quad \alpha/2 .$$

Случай Б

Мы отвергаем H0 на уровне значимости α, если T больше, чем w1 α, т. е. если

$$T > w1 \quad \alpha .$$

Случай С

Мы отвергаем H0 на уровне значимости α, если T меньше wα, т. е. если

$$T < w\alpha .$$

ОБЛАСТИ И ОГРАНИЧЕНИЯ Чтобы

использовать подписанный критерий Уилкоксона, должны выполняться следующие критерии: 1. Распределение di должно быть симметричным. 2. di должен быть независимым. 3. di должен измеряться в истинных значениях.

Знаковый критерий Уилкоксона также используется в качестве критерия медианы .

Данные X1, X2,...,Xn состоят из одной выборки размерности n.

Гипотезы, соответствующие предыдущим случаям A, B и C , следующие:

A: Двусторонние случаи:

H0: Медиана X равна предполагаемому значению Md

H1: Медиана X не равна Md

Machine Translated by Google

B: Односторонний случай:

H0: Медиана X Md

H1: медиана X < Md

C: Односторонний случай:

H0: медиана X Md

H1: Медиана X > Md

Обратите внимание, что «медиану» можно заменить на «средним» , потому что мы приняли гипотезу о симметрии распределения X.
Что касается расчетов, то формируем пары (X1, Md), (X2, Md),..., (Xn, Md) и относитесь к ним так же, как и раньше. Остальные теста идентичны.

ПРИМЕРЫ

Мы проводим тесты на восьми парах близнецов. цель состоит в том, чтобы увидеть, умнее ли первенец, чем второй.
Данные представлены в следующей таблице, более высокие баллы соответствуют лучшему тесту Результаты.

Пара двойняшки	Первый рожденный Си	Второй рожденный И	Ранг ди	ди	Ри
1	90	88	2	2	
2	75	79	4	4	4
3	99	98	1	1	1
4	60	66	6	5	5
5	72	64	8	6	6
6	83	83	0	–	–
7	86	86	0	–	–
8	92	95	3	3	3

Поскольку пары 6 и 7 не представляют разницы (d6 = d7 = 0), количество оставшихся пар равно m = 6. Поскольку все di разные, средние ранги вычислять не нужно, т.к. и используем сумму положительных рангов:

$$T = \sum_{(\text{для меня такой так как di} > 0)} Ри ,$$

где Ri — ранг пары (xi, yi) , представляющей положительный di. Таким образом, мы имеем:

$$T = 4 + 5 + 3 = 12 \quad .$$

Гипотезы следующие:

H0: Нет никакой разницы в уровне разведка между первым и вторым близнец (dm = 0)

H1: есть разница (dm = 0)

Эти гипотезы, соответствующие случаю A и решающее правило следующие:
Отклонить H0 на уровне α, если T < wα/2 или T > w1 α/2.
Если мы выберем уровень значимости α = 0,05, получаем из знака Уилкоксона таблица для m = 6:

$$w1 \quad \alpha/2 = 20 \text{ и } w\alpha/2 = 1 \quad ,$$

где T не больше 20 и не меньше чем 1, и мы не можем отвергнуть нулевую гипотезу H0.
Теперь предположим, что результаты для пар 7 и 8 изменились с новой гипотезой в следующей таблице:

Пара двойняшки	Первый рожденный Си	Второй рожденный И	Ранг ди	ди	Ри
1	90	88	2	4	4
2	75	79	4	5	5
3	99	98	1	2	2
4	60	66	6	6	6
5	72	64	8	7	7
6	83	83	0	–	–
7	86	87	1	2	2
8	92	91	1	2	2

Теперь пара 6 не представляет разницы (d6 = 0); таким образом, m равно 7.
Пары 3, 7 и 8 представляют один и тот же абсолютный разница: |d3| = |d7| = |d8| = 1. Мы

Вт

Machine Translated by Google

присвоить средний ранг парам, равным
ранг ex aequo, к которому мы добавляем половину
число ex aequo минус один. Таким образом:

$$\text{средний ранг} = \text{ранг } |d3| + \frac{1}{2}(3 \quad 1) = 2 \quad .$$

В этом случае мы используем статистический тест T1 ,
рассчитанный по формуле:

$$T1 = \frac{\sum_{я=1}^{м} Ри}{\sum_{я=1}^{м} R2_{я}} \text{ "=" } \frac{2}{11,75} = -0,17 .$$

Правило принятия решения следующее:
Отклонить H0 на уровне α, если T1 < zα/2 или T1 > z1 α/2.
Если выбрать уровень α = 0,05, то
значения обычной таблицы :

$$z1 \quad \alpha/2 = \quad 1,96 \text{ и } z\alpha/2 = 1,96 .$$

Следовательно, T1 не превышает
1,96 и не меньше -1,96.

Получаем тот же результат, что и в предыдущем
случае и не отвергаем нулевую гипотезу H0.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Непараметрический тест
- Подписанная таблица Уилкоксона

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Уилкоксон, Ф.: Индивидуальные сравнения
методы ранжирования. Биометрия 1, 80–83
(1945) (5.1, 5.7).

Уилкоксон, Ф.: Некоторые быстрые приближенные
статистические процедуры. американский цианамид,
Стэмфордские исследовательские лаборатории, Стэмфорд,
Коннектикут (1957 г.)

Таблица Уилкоксона

В таблице Уилкоксона приведены теоретические
значения статистики T теста Уилкоксона при
гипотеза об отсутствии разницы
между распределением двух сравниваемых
населения.

ИСТОРИЯ

Критические значения для N 20 (где N
общее число наблюдений) были рассчитаны
Wilcoxon, F. (1947).

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть (X1, X2,...,Xn) — выборка размерности n из
генеральной совокупности 1 и
(Y1, Y2,...,Ym) выборка размерности m
происходящие из населения 2.
Таким образом, мы получаем N = n+m наблюдений, что
мы хотим классифицировать в порядке возрастания
без учета их принадлежности к
образцы. Каждому значению присваиваем ранг:
ранг 1 до наименьшего значения, ранг 2 до следующего
значение, и так далее до ранга N до наибольшего
ценить.
Мы определяем статистический тест T следующим образом.
в манере:

$$T = \sum_{я=1}^{н} P(Си) ,$$

где R(Xi) — ранг, присвоенный наблюдению Xi, i =
1, 2,...,n, относительно множества
две выборки (X1,...,Xn) и (Y1,...,Ym).
Для различных значений n, m и α (уровень
значимости) таблица Уилкоксона дает
теоретические значения статистики T
критерий Уилкоксона, согласно гипотезе о том, что
две основные популяции идентичны
распределенный.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Таблица Уилкоксона используется в непараметрических тестах, использующих ранг, особенно в Тест Уилкоксона.

ПРИМЕРЫ

Вот выдержка из таблицы Вилкоксона для n = 14, m = 11 и α = 0,05, что равно 0,95:

н	м	α	критический
14	11	0,05	152
14	11	0,95	212

Для примера использования Уилкоксона таблица, см. критерий Вилкоксона.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Статистическая таблица

Тест Уилкоксона

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Уилкоксон, Ф.: Таблицы вероятностей для индивидуальных сравнений с помощью методов ранжирования. Биометрия 3, 119–122 (1947).

Тест Уилкоксона

Критерий Уилкоксона является непараметрическим критерием. Он используется, когда у нас есть два образца из двух популяций. Цель состоит в том, чтобы проверить если есть различие между популяциями на основе случайных выборок, взятых из этих популяций.

ИСТОРИЯ

Wilcoxontest назван в честь своего автора и был введен в 1945 г.

См. также критерий Манна-Уитни.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Пусть (X1, X2,...,Xn) — выборка размерности n из совокупности 1, и пусть (Y1, Y2,...,Ym) — выборка размерности m происходящие из населения 2.

Таким образом, мы получаем N = n+m наблюдений, что будем классифицировать по возрастанию без учета их принадлежности к выборкам. Затем мы присваиваем ранг 1 наименьшему оценочному значению, ранг 2 — следующему по величине значению, и так далее до ранга N, который присваивается самое высокое значение. Если много наблюдений точно такое же значение, то мы присвоим средний ранг. Обозначим через R(Xi) ранг присваивается Xi, i = 1,...,n.

Статистика W теста определяется:

$$Вт = \sum_{я=1}^{н} P(Си).$$

Если размер выборки большой (N 12), то используем, по Гиббонсу, JD (1971), аппроксимация с использованием статистики W1, которая должна следовать стандартному нормальному распределению N (0, 1):

$$W1 = \frac{Вт - мк}{о},$$

где μ и σ — среднее и стандартное отклонение случайного переменная W:

$$м = \frac{п(Н + 1)}{2}$$

и

$$\sigma = \sqrt{\frac{мин(N + 1)}{12}}.$$

Если использовать приближение для больших отсчетов W1, а если воспользоваться тем, что существует ранги ex aequo среди N наблюдений, мы заменяем стандартное отклонение на:

$$\sigma = \sqrt{\frac{мн}{12}\left(Н + 1 - \frac{\sum_{j=1}^{г} тдж^{2} - 1}{Н(Н - 1)}\right)},$$

Вт

Machine Translated by Google

где g - количество групп равных (равных) ранги и размерность группы j. (Если здесь нет равных (галстичных) рангов, наблюдения рассматривается как множество групп размерности 1. Следовательно, g = N и tj = 1 для j = 1,...,N, и σ уменьшается до $\sqrt{\frac{\text{мин}\,(N+1)}{12}}$.)

Гипотезы

Тест Уилкоксона можно провести на основе двустороннего теста или одностороннего теста в зависимости от типа гипотезы:

A: Двусторонний тест:

$$H0 : P(X < Y) = \frac{1}{2}$$
$$H1 : P(X < Y) = \frac{1}{2}$$

B: односторонний тест:

$$H0 : P(X < Y) \quad \frac{1}{2}$$
$$H1 : P(X < Y) > \frac{1}{2}$$

C: Односторонний тест:

$$H0 : P(X < Y) \quad \frac{1}{2}$$
$$H1 : P(X < Y) < \frac{1}{2}$$

Случай A выражает гипотезу о том, что существует нет разницы между двумя популяциями.
Случай B представляет собой гипотезу о том, что популяция 1 (откуда мы взяли выборку X) обычно принимает большие значения, чем популяция 2 (откуда мы взяли выборку из Y).
Случай C выражает противоположную гипотезу, т. е. что значения генеральной совокупности 1 имеют тенденцию быть меньше значений генеральной совокупности 2.

Правила принятия решений

Случай A

Мы отвергаем нулевую гипотезу H0 на уровне значимости α, если W меньше значения ue таблицы Уилкоксона с параметрами n, m, а $\frac{\alpha}{2}$ обозначается через tn,m,α/2 или если W больше затем значение в таблице для n, m и 1 α $\frac{}{2}$, по tn,m,1 α/2, т.е. если

$$W < tn,m,\alpha/2 \text{ или } W > tn,m,1 \quad \alpha/2 .$$

Если тест использует статистику W1, сравнение сделано с обычной таблицей:

$$W1 < z\alpha/2 \text{ или } W1 > z1 \quad \alpha/2 ,$$

при zα/2 и z1 α/2 значения нормальной таблицы с параметрами соответственно $\frac{\alpha}{2}$ и 1 — $\frac{\alpha}{2}$.

Случай Б

Отклоняем нулевую гипотезу H0 на уровне значимости α, если W меньше значения таблицы Вилкоксона с параметрами n, m, и α, обозначаемый через tn,m,α, т.е. если

$$W < tn,m,\alpha ,$$

и в случае, когда используется статистика W1 :

$$W1 < z\alpha ,$$

где zα — значение нормальной таблицы с параметром α.

Случай C

Отклоняем нулевую гипотезу H0 на уровне значимости α, если W больше значения в таблице Вилкоксона с параметрами n, m, и 1 α, обозначаемое через tn,m,1 α, т.е. если

$$W > tn,m,1 \quad \alpha ,$$

и со статистикой W1:

$$W1 > z1 \quad \alpha ,$$

где z1 α — значение нормальной таблицы с параметром α.

Machine Translated by Google

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Для использования теста Уилкоксона должны быть соблюдены следующие критерии:

1. Две выборки должны быть случайными. выборки, взятые из соответствующих популяций.
2. Помимо независимости внутри каждой выборки, между двумя выборками должна быть взаимная независимость.
3. Шкала измерения должна быть не ниже ordi нал.

ПРИМЕРЫ

В классе у нас 25 учеников: 14 мальчиков и 11 девочек. Мы проверяем их на умственном счете посмотреть, есть ли у мальчиков в этом классе склонность быть лучше девушек.

Данные представлены в следующей таблице, самые высокие баллы соответствуют лучшему тесту.

Результаты.

Мальчики (Си)		Девушки (И)	
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	

Упорядочим баллы в порядке возрастания и присвоить ранг:

Результаты	Образец	Классифицировать	Р (Си)
7.1	Д	1	–
7.7	Икс	2	2
8,6	Д	3	–
9.2	Икс	4	4
10,6	Д	5,5	–
10,6	Икс	5,5	5,5
11.1	Д	7	–
11.3	Икс	8	8
12.3	Икс	9	9
13.3	Д	10	–
14,0	Икс	11	11
14.1	Д	12	–
15.2	Икс	13	13
15,6	Икс	14,5	14,5
15,6	Д	14,5	–
16,0	Икс	16	16
16,4	Икс	17	17
17,5	Икс	18	18
17,7	Д	19	–
17,9	Икс	20	20
19,8	Икс	21	21
20,3	Д	22	–
21,0	Д	23	–
21.1	Икс	24	24
23,6	Д	25	–

Рассчитываем статистику W:

$$Вт = \sum_{я=1}^{14} Р\,(Си)$$

$$= 2 + 4 + 5{,}5 + 8 + 9 + 11 + 13 + 14{,}5 + 16 + 17 + 18 + 20 + 21 + 24 = 183 .$$

Если выбрать приближение для больших выборки с учетом средних рангов, делаем корректировку:

$$W1 = \frac{Вт - \frac{п\,(Н + 1)}{2}}{\sqrt{\frac{мн}{12}\left(м + п + 1 - \frac{\sum_{j=1}^{г} тдж\,т^{2}\,1}{(м + п)\,(м + п - 1)}\right)}} .$$

Здесь g = 2 и tj = 2 при j = 1, 2. Таким образом, мы иметь:

$$W1 = \frac{183 - \frac{14\,(25 + 1)}{2}}{\sqrt{\frac{11 \cdot 14}{12}\left(11 + 14 + 1 - \frac{2(22^{3} - 1^{3} + 2^{3} - 2^{3})}{(11+14)(11+14^{3} - 1)}\right)}}$$

$$= 0{,}0548 .$$

Вт

Machine Translated by Google

Гипотезы, которые мы хотим проверить:

H0: Мальчики в этом классе не стремятся быть лучше девочек в умственном счете.

H1: Мальчики в этом классе имеют тенденцию быть лучше, чем девочки в этом классе в умственных вычислениях.

Это выражается через H0: P (X < Y) 1 и 2 соответствует случаю C. Правило принятия решения следующее:
отклонить H0 на уровне значимости α, если

$$W > t_{n,m,1-\alpha} \text{ или } W1 > z_{1-\alpha}.$$

Если мы выберем α = 0,05, теоретическое значение в таблице Уилкоксона будет $t_{n,m,1-\alpha} = 212$.
Вычисленное значение W (W = 183) не превышает $t_{n,m,1-\alpha} = 212$, и мы не можем отвергнуть нулевую гипотезу H0.
Значение в нормальной таблице $z_{1-\alpha}$ равно 1,6449.
Следовательно, W1 не больше, чем $z_{1-\alpha}$ (0,0548 < 1,6449), и мы делаем такой же вывод на основании

тик W: мальчики в этом классе не стремятся быть лучше девочек в умственном счете.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Проверка гипотезы
- Тест Манна – Уитни
- Непараметрический тест
- Подписанный тест Уилкоксона
- Таблица Уилкоксона

ССЫЛКИ Гиббонс,

Д.Д.: Непараметрический статистический вывод. Макгроу-Хилл, Нью-Йорк (1971)

Уилкоксон, Ф.: Индивидуальные сравнения методами ранжирования. Биометрия 1, 80–83 (1945).

Уиллкокс, Уолтер Фрэнсис

Уиллкокс, Уолтер Фрэнсис (1861–1964) окончил Амхерстский колледж в Массачусетсе (получив степень бакалавра в 1884 г. и степень магистра в 1888 г.). Он продолжил учебу в Колумбийском университете в Нью-Йорке в конце 1880-х годов. Он получил свой документ

в Колумбийском университете в 1891 г.
Его докторская степень диссертация называлась «Развод».

Задача: Исследование по статистике. С 1891 года Уиллкокс стал профессором экономики и статистики Корнельского университета.

г., где он оставался до выхода на пенсию в 1931 г.

Его основной вклад в статистику относится к области демографии, а также к развитию системы федеральной статистики.

С 1899 по 1902 год он руководил 12-й переписью населения ООН и председательствовал в Американской статистической ассоциации в

1912 г., Американская экономическая ассоциация в 1915 г. и Международный статистический институт в 1947 г.

Избранные основные работы и статьи Уиллкокс, Вальтер Фрэнсис:

1897 г. Проблема развода: исследование статистики. Исследования в области истории, экономики и публичного права. Колумбийский университет

Versity Press, Нью-Йорк.

1940 Исследования по американской демографии, издательство Корнельского университета, Итака, Нью-Йорк.

Machine Translated by Google

Алгоритм Йейтса

Алгоритм Йейтса — это процесс, используемый для вычисления оценок основных эффектов и взаимодействий в факторном эксперименте.

ИСТОРИЯ

Алгоритм Йейтса был создан Йейтсом Франком в 1937 году.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

Рассмотрим факторный эксперимент с n факторами, каждый из которых имеет два уровня. Обозначим этот эксперимент через 2n.

Факторы обозначаются прописными буквами .

Мы используем строчную букву для высокого уровня фактора и опускаем букву для нижнего уровень. Когда все факторы находятся на нижнем уровне уровень, мы используем символ (1).

Алгоритм представлен в таблице . В первом столбце находим все комбинации уровней различных факторов в стандартном порядке (за введением буквы следуют комбинации со всеми предыдущими комбинациями) . Второй столбец содержит наблюдения всех комбинаций.

Для последующих n столбцов процедура такая же: первая половина столбца получается суммированием пар последовательных значений предыдущего столбца; вторая половина столбца получается вычитанием получение первого значения из второго значения каждой пары.

Столбец n содержит контрасты основных эффектов и взаимодействий, приведенных в первом столбце. Чтобы найти оценки этих эффектов, разделим n-й столбец на m·2n 1, где m — число повторений эксперимента , а n — число факторов. Беспокойство

В первой строке мы находим в столбце n сумму всех наблюдений.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Алгоритм Йейтса можно применять к экспериментам с факторами , имеющими более два уровня.

Можно также использовать этот алгоритм для нахождения суммы квадратов, соответствующих основным эффектам и взаимодействиям 2n- факториала . эксперимент. Они получаются возведением в квадрат n-го столбца и делением результата на m · 2n (где m — число повторений, а n — число факторов).

ПРИМЕРЫ

Чтобы вычислить оценки основных эффектов и трех взаимодействий, давайте рассмотрим факторный эксперимент 23 (имеется в виду три фактора с двумя уровнями) относительно производства сахарной свеклы. Первый фактор, обозначенный A, касается типа удобрения (навоз или ненавоз); в

Machine Translated by Google

второй, обозначенный B, касается глубины вспашка (20 или 30 см); и последний фактор C, сорт сахарной свеклы (сорт 1 или 2). Поскольку каждый фактор имеет два уровня, можно говорят об эксперименте 2 × 2 × 2. Например, взаимодействие между фактором A и фактором B обозначается A×B, а взаимодействие между всеми тремя факторами пишется A × B × C.

Восемь наблюдений указывают вес урожая в килограммах за полторы гектар, давая следующую таблицу:

A	Навоз		Ненавоз	
Б	20 см	30 см	20 см	30 см
Вар. 1	80	96	84	88
C				
Вар. 2	86	96	89	93

Проводя этот эксперимент, мы хотим количественно оценить не только влияние удобрений, вспашки, и сорта свеклы (простые эффекты), но также влияние комбинированного воздействия (воздействия взаимодействия) на уборке урожая. Чтобы составить таблицу алгоритма Йейтса, восемь наблюдений должны быть классифицированы в Заказ Йейтса. Классификационная таблица в этом случае удобна; символы «-» и «+» представляют собой, соответственно, нижний и верхний уровень факторов. Например, первая строка указывает на наблюдение «80», соответствующее нижним уровням для всех трех факторов (A: навоз, B: 20 см, C: сорт 1).

Таблица классификации наблюдений в порядке Йейтса

Обс.	Фактор	Соответствующий
	азбука	наблюдение
1	80	
2 +	84	
3 +	96	
4 ++	88	
5 +	86	
6 ++	89	
7 ++	96	
8 +++	93	

Теперь мы можем составить таблицу алгоритма Йейтса:

Таблица, соответствующая алгоритму Йейтса

Комби-нация	обс.	Столбцы 123	Оценка	Средний эффект
(1)		80 164 348 712 89		—
а		84 184 364 4 1 A		
б		96 175 4 34 8,5 Б		
аб		88 189 0 18 4,5 A × B		
с		86 4 20 16 4		C
ac		89 8 14 4 1 A × C		
bc		96 3 12 6 1,5 B × C		
абв	93	3 6 6 1,5 A × B × C		

Члены в столбце 2, например, вычисляются следующим образом:

- Первая половина столбца (сумма элементов каждая пара предыдущего столбца, в этом случае столбец 1):

348 = 164 + 184

364 = 175 + 189

4 = 4 + (8)

0 = 3 + (3)

- Вторая половина столбца (вычитание элементов каждой пары в обратном порядке):

20 = 184 - 164

14 = 189 - 175

12 = 8 4

6 = 3 3

Machine Translated by Google

Оценка основного эффекта _

тор A равен 1. Как видно, значение 4 (столбец 3) было разделено на m · 2n 1 = 1 · 23 1 = 4, где m равно 1, так как эксперимент проводился только один раз (без повторения), а n соответствует числу факторов, которых в данном

случай.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

- Дисперсионный анализ
- Контраст
- План экспериментов
- Факторный эксперимент
- Взаимодействие
- Основной эффект

ССЫЛКИ Йейтс,

Ф.: Дизайн и анализ факторных экспериментов. Техническое сообщение Бюро Содружества

почв 35, Сельскохозяйственное содружество Бюро, Фарнхэм Роял (1937)

Юден, Уильям Джон

Юден, Уильям Джон (1900–1971) получил степень магистра химии в Колумбийском университете в 1924 году. С 1924 по 1948 год он работал в Институте ИССЛЕДОВАНИЯ РАСТЕНИЙ Бойса Томпсона в

Йонкерс, Нью-Йорк. С 1948 г. до выхода на пенсию в 1965 г. Юден В. Дж. был членом отдела прикладной математики в Национальном бюро стандартов в Вашингтоне, округ Колумбия. В этот период Юден разработал тест и написал «Статистические методы для химиков» .

Основные работы Юдена, Уильяма Джона:

1951 Статистические методы для химиков.

Уайли, Нью-Йорк.

Коэффициент Юла и Кендалла

Коэффициент Юла используется для измерения асимметрия частотного распределения. Он учитывает относительные положения квартилей по отношению к медиане и сравнивает распространение кривой вправо и влево от медианы.

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ

При симметричном распределении квартили находятся на одинаковом расстоянии по обе стороны от медианы. Это значит, что

$$(Q3 - M) - (M - Q1) = 0 \quad ,$$

где M — медиана, Q1 — первый квартиль, а Q3 — третий квартиль.
Если распределение асимметрично, то предыдущее равенство неверно.
Левую часть уравнения можно преобразовать следующим образом:

$$Q3 + Q1 - 2M \quad .$$

Чтобы получить коэффициент асимметрии, равный независимо от единицы измерения уравнение следует разделить на величину (Q3 Q1).

Тогда коэффициент Юла равен:

$$КИ = \frac{Q3 + Q1 - 2M}{Q3 - Q1} \quad .$$

Если CY положителен, то распределение распространяется вправо; если CY отрицательна, распределение смещается влево.

ОБЛАСТИ И ОГРАНИЧЕНИЯ

Если коэффициент положительный, то он представляет собой кривую, расширяющуюся вправо. Если он положительный, то он представляет собой кривую, расширяющуюся влево. Коэффициент, который

Machine Translated by Google

близкое к нулю означает, что кривая приблизительно симметрична.

Коэффициент Юла и Кендалла может варьироваться между -1 и +1. В самом деле, поскольку медиана M всегда находится между первой и третий квартиль, следующие два крайних возникают возможности:

1. M = Q1
2. M = Q3

В первом случае коэффициент принимает вид:

$$КИ = \frac{Q3 + Q1 - 2Q1}{Q3 - Q1} = 1 \quad ,$$

а во втором случае получается:

$$КИ = \frac{Q3 + Q1 - 2Q3}{Q3 - Q1} = \quad 1 \quad .$$

В этих условиях коэффициент должен быть близко к нулю, чтобы распределение было согласованным двусторонне симметричный.

Этот коэффициент, как и другие уверенности в асимметрии представляют интерес, только если они может сравнивать формы двух или нескольких дистрибутивы. Результаты сильно различаются от одной формулы к другой. Очевидно, что сравнения должны производиться с использованием одних и тех же формула.

ПРИМЕРЫ

Предположим, что мы хотим сравнить форму распределения дневного оборота в 75 пекарни более 2 лет. Затем мы вычисляем в в обоих случаях коэффициенты Юла и Кендалла.

Данные следующие :

Частота оборота	1 год	Частота 2 год
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

Для первого года первый квартиль, медиана и третий квартиль соответственно:

$$Q1 = 268{,}46 \quad ,$$
$$M = 288{,}18 \quad ,$$
$$Q3 = 310 \; .$$

Тогда коэффициент будет следующим:

$$CY = \frac{Q3 + Q1 - 2M}{Q3 - Q1}$$
$$"=" \frac{310 + 268{,}46 - 2(288{,}18)}{310 - 268{,}46}$$
$$= 0{,}05 \; .$$

Для года 2 эти значения соответственно:

$$Q1 = 230 \quad ,$$
$$M = 251{,}67 \quad ,$$
$$Q3 = 293{,}12 \; .$$

Тогда коэффициент Юла и Кендалла равен равно:

$$CY = \frac{Q3 + Q1 - 2M}{Q3 - Q1}$$
$$"=" \frac{293{,}12 + 230 - 2(251{,}67)}{293{,}12 - 230}$$
$$= 0{,}31 \; .$$

За 1 год, так как коэффициент дает результат близко к нулю, можно признать, что распределение дневного оборота в 75 пекарни очень близко к симметричному распределению. Для 2-го года коэффициент выше er; это означает, что распространение распространяется вправо.

Machine Translated by Google

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Мера формы
Мера асимметрии

ССЫЛКИ Юл,
Г.В., Кендалл, М.Г.: Введение в теорию статистики,
14-е изд. Griffin, Лондон (1968)

Юл, Джордж Удни

Юл Джордж Удни родился в 1871 году в Бич-Хилл недалеко от Хаддингтона (Экоссе) и умер в 1951 году в Кембридже.
Коэффициент ассоциации Юла, парадокс Юла (позже названный также парадоксом Симпсона) и процедура Йола носят его имя в статистике, которые также внесли свой вклад.

к теории Менделя и временным рядам.

Избранные основные произведения Юла, Джорджа Удный:

1897 г. К теории корреляции. Дж. Рой.
Стат. соц. 60, 812–854.
1900 г. Об ассоциации атрибутов в
статистика: с иллюстрацией из
материал общества детства.Phi los. Транс.
Рой. соц. Лонд. сер. А 194, 257–319.

1926 г. Почему мы иногда получаем бессмысленные корреляции между временными рядами?
Исследование выборки и природы
временных рядов. Дж. Рой. Стат. соц. (2)
89, 1–64.
1968 (совместно с Кендаллом, М.Г.) Введение в
теорию статистики. 14-е изд.
Griffin, Лондон.

ДАЛЬНЕЙШЕЕ ЧТЕНИЕ

Мера асимметрии
Коэффициент Юла и Кендалла

Machine Translated by Google

Приложение

Machine Translated by Google

Приложение А: Таблица случайных чисел, десятичные дроби числа π

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

Приложение B: Биномиальная таблица

Число испытаний		Вероятность успеха								
		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

Число испытаний		Вероятность успеха								
		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	0,9901	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

Количество испытаний	Вероятность успеха								
	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

Приложение C: Таблица Фишера для α = 005

	v1									
v2	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
3,84		3.00	2,60	2,37	2.21	2.10	2.01	1,94	1,88	1,83

Machine Translated by Google

					ν1				
ν2	12	15	20	24	30	40	60	120	
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
1,75		1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Machine Translated by Google

Приложение D: Таблица Крускала-Уоллиса

H_1	H_2	H_3	Критическое значение	α
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

H_1	H_2	H_3	Критическое значение	α
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	Критическое значение	α
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	Критическое значение	α
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

Приложение E: Стол для учащихся

	α					
ν	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
1,282		1,645	1.960	2,326	2,576	3.090

Machine Translated by Google

Приложение F: Таблица хи-квадрат

	Степень свободы					
ν	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

Приложение G: Стандартная таблица нормальных значений

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

Библиография

Абди, Х.: Биномиальное распределение: биномиальные и знаковые тесты. В: Салкинд, Нью-Джерси (ред.) Энциклопедия измерений и статистики. Мудрец, Тысяча дубов (2007)
Биномиальный тест

Абди, Х .: Расстояние. В: Салкинд, Нью-Джерси (ред.) Энциклопедия измерений и статистики. Мудрец, Тысяча дубов (2007) Расстояние

Агрести, А.: Категориальный анализ данных. Wiley, Новый анализ Йорк (1990) категориальных данных

Эйри, Великобритания: письмо профессора Эйри, королевского астронома, редактору. Астроном. Журнал 4, 137–138 (1856 г.) Выброс

Аллан, Ф.Э., Уишарт, Дж.: Метод оценки урожайности отсутствующего участка в полевых экспериментальных работах. Дж. Сельское хозяйство. науч. 20, 399–406 (1930) Недостающие данные

Альтман, Д.Г.: Практическая статистика для медицинских исследований. Чепмен и Холл, Лондон (1991)
Логистическая регрессия

Андерсон, Т.В., Дарлинг, Д.А.: Асимптотическая теория некоторых критериев согласия, основанная на стохастических процессах. Анна. Мат. Стат. 23, 193–212 (1952)
Тест Андерсона-Дарлинга

Андерсон, Т.В., Дарлинг, Д.А.: Тест на соответствие. Варенье. Стат. доц. 49, 765–769 (1954)
Тест Андерсона-Дарлинга

Андерсон, Т.В., Слове, С.Л.: Введение в Статистический анализ данных. Хоутон Миффлин, Бостон (1978) Диапазон

Андерсон, В.Л., Маклин, Р.А.: Дизайн эксперименты: реалистический подход. Марсель Деккер, Нью-Йорк (1974) Модель

Эндрюс Д. Ф., Герцберг А. М.: Данные: Сборник задач из многих областей для студентов и научных сотрудников. Спрингер, Берлин Гейдельберг, Нью-Йорк (1985)
Ср Критерий

Anscombe, FJ, Tukey, JW: Анализ остатков. Технометрика 5, 141–160 (1963). Анализ Остатки

Anscombe, FJ: Исследование остатков. проц. 4-й Беркли Симп. Мат. Статист. Проб. 1, 1–36 (1961)
Анализ остатков

Анскомб, Ф.Дж.: Графики в статистическом анализе. Являюсь. Стат. 27, 17–21 (1973) Анализ остатков

Антилль Г., Уджвари А.: Статистическая практика. Inferentielle. ПАН, Невшатель, Швейцария (1991)
Сбор данных

Арбутнотт, Дж.: Аргумент в пользу Божественного Провидения, взятый из постоянной регулярности, наблюдаемой в рождении обоих полов. Филос. Транс. 27, 186–190 (1710)
Проверка гипотез, непараметрический тест, знаковый тест

Арматт, М.: Построение понятий. d'estimation et de vraisemblance chez Ronald A. Фишер. В: Maresse, J. (ed.) Estimation et sondages. Экономика, Париж (1988) Оценка

Армитидж П., Колтон Т.: Энциклопедия биостатистики. Уайли, Нью-Йорк (1998)
Биостатистика

Артханари, Т.С., Додж, Ю.: Математика Программирование в статистике. Уайли, Нью-Йорк (1981)
Линейное программирование, оптимизация

Азимов, Д.: Большой тур: инструмент для просмотра многомерные данные. SIAM J. Sci. Стат. вычисл. 6 (1), 128–143 (1985) Статистическое программное обеспечение

Аттея, О., Дюбуа, Ж.-П., Вебстер, Р.: Геостатистический анализ загрязнения почвы в швейцарской Юре. Окружающая среда. Загрязн. 86, 315–327 (1994)
Геостатистика, пространственные данные

Бейли, Д.Э.: Вероятность и статистика: модели для исследования. Уайли, Нью-Йорк (1971) Образец пространства

Machine Translated by Google

Балестра, П.: Матричный расчет для экономистов.
Editions Castella, Альбёв, Швейцария (1972)
Матрица

Барлетт, MS: Введение в стохастик
Процессы. Издательство Кембриджского университета, Кембридж
(1960) Стохастический процесс

Барнард, Г.А., Плакетт, Р.Л., Пирсон, С.: «Студент»: статистическая
биография Уильяма Сили Госсета: на основе работ Э.С.
Пирсона. Издательство Оксфордского университета,
Оксфорд (1990) Стандарт
Отклонение

Барнетт, В., Льюис, Т.: Выбросы в статистических данных.
Уайли, Нью-Йорк (1984) Выброс

Бартлетт, М.С.: О теоретической спецификации
выборочные свойства автокоррелированных временных рядов.
Дж. Рой. Стат. соц. сер. Б 8, 27–41 (1946)
Серийная корреляция

Бартлетт, М.С.: Свойства достаточности и статистических критериев.
проц. Рой. соц. Лонд. сер. А 160, 268–282 (1937)
Анализ категорийных данных

Бартлетт, М.С.: Некоторые примеры статистических методов
исследования в сельском хозяйстве и прикладной биологии.
Дж. Рой. Стат. соц. (Приложение) 4, 137–183 (1937).
Критерий хи-квадрат, ковариационный анализ,
отсутствующие данные

Бассет, Г., Кенкер, Р.: Асимптотическая теория регрессии с
наименьшими абсолютными ошибками. Варенье. Стат. доц.
73, 618–622 (1978) Оценка L1 Bates,

DM, Watts DG: нелинейная регрессия
анализа и его приложений. Уайли, Нью-Йорк (1988)
Нелинейная регрессия

Байеса, Т.: Очерк решения проблемы в учении о шансах. Филос.
Транс. Рой. соц. Лонд.
53, 370–418 (1763). Опубликовано по инициативе Прайса Р.
через 2 года после его смерти. Переиздано с биографией
Барнарда, Джорджа А. в 1958 году, и в Пирсоне, Э.С., Кендалле,
М.Г.: Исследования в области
История статистики и вероятности. Гриффин, Лондон,
стр. 131–153 (1970). Теорема Байеса,
Логическая статистика

Bealle, EML: Введение в оптимизацию. Wiley, Operations Research
Нью-Йорк (1988) Becker, RA, Chambers, JM,
Wilks, AR: Новый язык S. Уолдсворт и Брукс/Коул, Пасифик-Гроув,
Калифорния (1988)
Статистическое
программное обеспечение Белсли, Д., Ку, Э., Уэлш, RE: Регрессия
диагностика Выявление влиятельных данных и источников
коллинеарности. Уайли, Нью-Йорк (1976)
Мультиколлинеарность

Белсли Д.А., Кух Э., Уэлш Р.Э.: Диагностика регрессии. Уайли,
Нью-Йорк, стр. 16–19 (1980).
Шляпа Матрица

Бенигер, Дж. Р., Робин, Д. Л.: Количественная графика в
статистика: краткая история. Являюсь. Стат. 32, 1–11 (1978)
Количественный график

Бензекри, JP: Histoire et préhistoire de l'analyse des données, les
cahiers de l'analyse des données 1, no. 1–4. Дюно, Париж (1976)
Переписка
Анализ, анализ данных

Бензекри, JP: Histoire et préhistoire de l'analyse des données. Дюно,
Париж (1982) Вероятность

Benzécri, JP: L'Analyse des données. Том. 1: Таксиномия. Том. 2:
L'Factoryelle des корреспонденции. Дюно, Париж (1976)
Анализ соответствия, анализ данных Бернардо, Дж. М.,

Смит, ACM: байесовская теория.
Уайли, Чичестер (1994) Байесовская статистика

Бернулли, Д.: Что является причиной физического наклона
планет (...). Рек. Отчет о предметах. Приз акад. Рой.
науч. 3, 95–122 (1734)
Проверка гипотез

Бернулли, Д.: Наиболее вероятный выбор между несколькими
противоречивыми наблюдениями, переводом и
комментариями (1778 г.). В: Пирсон, Э.С. и Кендалл,
М.Г. (ред.) Исследования по истории статистики и
вероятности. Гриффин, Лондон (1970)
Оценка

Бернулли, Дж.: Ars Conjectandi, Opus Posthumum.
Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de
ludo Pilae recticularis. Impensis Thurnisiorum, Fratrum, Базель
(1713 г.) Бернулли
Испытание, Теорема Бернулли, Биномиальное
Распределение, оценка, закон большого
Числа, нормальное распределение

Берри, Д.А.: Статистика, байесовская перспектива.
Уодсворт, Белмонт, Калифорния (1996) Байесовская
статистика

Бикель, Р. Д.: Выбор оптимальной области отторжения для
множественного тестирования. Рабочий документ,
Управление биостатистики и биоинформации, Медицинский
колледж Джорджии (2002 г.) Район отбраковки

Bienaymé, IJ: Соображения по поводу Appui de la
découverte de Laplace sur la loi de probabilité dans la méthode des
moindres carrés. Comptes Rendus de l'Académie des Sciences,
Paris 37, 5–13 (1853 г.); переиздано в 1867 году в журнале
Лиувилля, предшествующем ла преуве де Негалитес де
Бьенеме-Чебышева, Journal de Mathématiques

Machine Translated by Google

Pures et Appliquées 12, 158–176 Law of Large Числа

Биллингсли, П.: Статистический вывод для Маркова Процессы. Издательство Чикагского университета, Чикаго (1961) Стохастический процесс

Биркес, Д., Додж, Ю., Сили, Дж.: Охватывающие наборы для заметные контрасты в моделях классификации. Анна. Стат. 4, 86–107 (1976) Недостающие данные

Биркес Д., Додж Ю.: Альтернативные методы регрессии. Уайли, Нью-Йорк (1993) Коллинеарность, регрессия с наименьшим абсолютным отклонением, стандартизированные данные Bishop, YMM, Fienberg, SE, Holland, PW: дискретный многомерный анализ: теория и практика. MIT Press, Кембридж, Массачусетс (1975) Двоичные данные, обобщенная линейная регрессия, Логистическая регрессия

Бьерхаммар, А.: Прямоугольные взаимные матрицы со специальной ссылкой на геодезические расчеты. Бык. Геод. 20, 188–220 (1951) Обобщенный обратный

Бок, Р.Д.: Многомерные статистические методы в Поведенческие исследования. Макгроу-Хилл, Нью-Йорк (1975) Коллинеарность

Boscovich, RJ et Maire, C.: De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus. Палладис, Рим (1755 г.) Модель Боскович, RJ: De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex instanceria etiam sensorum impressa. Bononiensi Scientiarium et Artium Instituto Atque Academia Commentarii, vol. IV, стр. 353–396 (1757 г.) L1 Оценка, медиана

Бурбонне, Р.: Econométrie, manuel et exercices corrigés, 2-е изд. Дюно, Париж (1998) Автокорреляция, критерий Дарбина-Ватсона

Бурсен, Ж.-Л.: Lesstructures du hasard. Сёй, Париж Сбор данных (1986)

Бауэрман, Б.Л., О'Коннел, Р.Т.: Временные ряды и прогнозирование: прикладной подход. Даксбери, Белмонт, Калифорния (1979) Неправильная вариация

Боули, Алабама: Обращение президента к экономическим секция Британской ассоциации. Дж. Рой. Стат. соц. 69, 540–558 (1906) Доверительный интервал, выборка

Box, GEP, Draper, NR: основа для выбора конструкции поверхности отклика. Варенье. Стат. доц. 54, 622–654 (1959) Критерий общего среднего Squared Error Box,

GEP, Hunter, WG, Hunter, JS: Статистика для Экспериментаторы. Введение в проектирование, анализ данных и построение моделей. Уайли, Нью-Йорк (1978) Эксперимент, Основной эффект, Стандарт Ошибка

Бокс, ГЭП, Дженкинс, ГМ: Анализ временных рядов: прогнозирование и контроль (ряды в анализе временных рядов). Холден Дэй, Сан-Франциско (1970) Модели ARMA, автокорреляция, цикличность Флуктуация, серийная корреляция

Box, GEP, Tiao, GP: байесовский вывод в Статистический анализ. Аддисон-Уэсли, Рединг, Массачусетс (1973) Блок байесовской статистики , GEP: ненормальность и тесты на дисперсию. Биометрика 40, 318–335 (1953). Крепкий Оценка

Браун, Л.Д., Олкин, И., Сакс, Дж., Винн, НР: Джек Карл Кифер, Сборник статей III: Планирование экспериментов. Спрингер, Берлин Гейдельберг, Нью-Йорк (1985) Оптимальный дизайн

Карли, GR: Del valore и т. Д. B: Opere Scelte di Carli, vol. я, с. 299. Кустоди, Милан (1764 г.) Индекс Число

Карлин, Б., Луис, Т.А.: Байесовский и эмпирический байесовский анализ Методы. Чепмен и Холл /CRC, Лондон (2000) Байесовская статистика

Казелла, Г., Бергер, Р.Л.: Статистический вывод. Даксбери, Пасифик-Гроув (2001) Соединение Функция распределения вероятностей

Коши, А.Л.: На основе результатов наблюдений за природой мемов и на основе результатов плюс вероятности. ЧР акад. науч. 37, 198–206 (1853 г.) Распределение Коши

Селе, Г., Дидей, Э., Говар, Г., Лешевалье, Ю., Раламбондрани, Х.: Автоматическая классификация основ - аспекты статистики и информации. Дюно, Париж (1989) Кластерный анализ, мера непохожести палат, Дж.

М. и Хасти, Т.К. (ред.): Статистические данные. модели в S. Waldsworth и Brooks/Cole, Pacific Гроув, Калифорния (1992) Статистическое программное обеспечение

Чемберс, Дж. М., Кливленд, В. С., Кляйнер, Б., Тьюки, Пенсильвания: Графические методы анализа данных. Уодсворт, Белмонт, Калифорния (1983) График QQ (график квантиля к квантилю)

Чемберс, Дж. М.: Вычислительные методы анализа данных. Уайли, Нью-Йорк (1977) Алгоритм

Machine Translated by Google

Чатфилд, К.: Анализ временных рядов: An
Введение, 4-е изд. Чепмен и Холл (1989)
Автокорреляция

Чатфилд, К.: Анализ временных рядов. Введение. Чепмен
и Холл (2003) Светская
тенденция

Чаттерджи, П., Хади, А.С.: Регрессионный анализ на примере.
Уайли, Нью-Йорк (2006)
Множественная линейная регрессия

Cochran, WG: Sampling Techniques, 2-е изд.
Уайли, Нью-Йорк (1963) Размер выборки,
простая случайная выборка, стандартная ошибка

Коллатц, Л., Веттерлинг, В.: Проблемы оптимизации.
Спрингер, Берлин Гейдельберг, Нью-Йорк (1975)
Исследование операций

Кук, Р.Д., Вайсберг, С.: Введение в
Графика регрессии. Уайли, Нью-Йорк (1994)
Анализ остатков

Кук, Р.Д., Вайсберг, С.: Прикладная регрессия
Включая вычислительную технику и графику. Уайли, Нью-
Йорк (1999) Анализ остатков

Кук, Р.Д., Вайсберг, С.: Остатки и влияние в регрессии. Чепмен и
Холл, Лондон (1982)
Анализ остатков

Коупленд, MT: Статистические показатели условий
ведения бизнеса. QJ Экон. 29, 522–562 (1915)
Временная последовательность

Корнфилд, Дж.: Метод оценки сравнительных
показатели по клиническим данным. Применение при раке
легких, молочной железы и шейки матки. Дж. Натл. Рак
инст. 11, 1269–75 (1951) атрибутивный риск,
Предотвратимый риск, уровень распространенности, относительный
Риск, Риск

Корнфилд, Дж.: Методы отбора проб человеческого населения.
Определение размера выборки.
амер. Дж. Публ. Здоровье 41, 654–661 (1951)
Размер образца

Котес, Р.: Aestimatio Errorum in Mixta Mathesi, per вариации partium
Trianguli plani et sphaerici. В: Смит, Р. (ред.) Opera Miscellania,
Кембридж (1722 г.)
Взвешенное среднее арифметическое

Кокс, Д.Р., Бокс, GEP: анализ
преобразования (с обсуждением). Дж. Рой. Стат. соц.
сер. Б 26, 211–243 (1964) Трансформация

Кокс, Д.Р., Хинкли, Д.В.: Теоретическая статистика.
Чепмен и Холл, Лондон (1973) Вероятность
Соотношение Тест

Кокс, Д.Р., Миллер, HD: Теория стохастика
Процессы. Уайли, Нью-Йорк; Метуэн, Лондон (1965)
Стохастический процесс

Кокс, Д.Р., Рид, Н.: Теория дизайна
эксперименты. Чепмен и Холл, Лондон (2000)
Односторонний дисперсионный анализ, двусторонний
Анализ дисперсии Кокс,

Д.Р., Снелл, Э.Дж.: Анализ бинарных данных, 2-е изд. Чепмен и
Холл, Лондон (1990)
Анализ категориальных данных

Кокс, Д.Р., Снелл, Э.Дж.: Анализ двоичных данных.
Чепмен и Холл (1989) Анализ двоичных
данных, двоичных данных, логистической регрессии

Кокс, Д.Р.: Оценка методом двойной выборки.
Биометрика 39, 217–227 (1952). Размер выборки

Кокс, Д.Р.: Взаимодействие. Междунар. Стат. 52 , 1–25 (1984)
Взаимодействие

Кокс, PR: Демография, 5-е изд. Издательство
Кембриджского университета, Кембридж (1976)
Демография

Кресси, NAC: Статистика пространственных данных. Wiley,
Нью-Йорк (1991) Геостатистика, Пространственная
Статистика

Кром, AFW: Географически-статистика
Darstellung der Staatskräfte. Вейганд, Лейпциг (1820 г.)
Графическое представление

Crome, AFW: Ueber die Grösse und Bevölkerung der sämtlichen
Europäischen Staaten. Вейганд, Лейпциг (1785 г.)
Графическое представление

Денес, Дж., Кидуэлл, А.Д.: Латинские квадраты и их приложения.
Академик, Нью-Йорк (1974)
Graeco-Latin Square Design Dantzig,

Великобритания: Приложения и расширения линейного
программирования. Дюно, Париж (1966)
Линейное программирование

Дэвид, FN: Игра в кости и игры (заметка об истории вероятности).
Биометрика 42, 1–15 (1955).
Вероятность

Дэвисон, А.С., Хинкли, Д.В.: Методы начальной загрузки и их
применение. Издательство Кембриджского университета,
Кембридж (1997) Bootstrap

DeLury, DB: ковариационный анализ. Биометрический
4, 153–170 (1948) ковариационный анализ

Деланиус, Т.: Элементы выборки обследования. SAREC, Sampling
Стокгольм (1985) Deming,

WE: Дизайн выборки в бизнес-исследованиях.
Уайли, Нью-Йорк (1960) Систематическая выборка

Desrosières, A. La partie pour le tout: comment généraliser? La
prehistoire de la contrainte de

Machine Translated by Google

представитель. Оценка и зондаж. Cinq взносы à l'histoire de la statistique. Экономика, Париж (1988) Доверительный интервал, выборка

Deuchler, G.: Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. Zeitung für Pädagogische Psychololegie und Experimentelle Pädagogik, 15, 114–131, 145–159, 229–242 (1914) Кендалл Коэффициент ранговой корреляции

Диггл, П.Дж., П. Хигерти, К.-Ю. Лян и С.Л. Зегер.: Анализ продольных данных, 2-е изд. Издательство Оксфордского университета, Оксфорд (2002) Панель

Диксон, главный редактор WJ: BMDP Statistical Software Communications. Издательство Калифорнийского университета. Беркли, Калифорния (1981) Статистическое программное обеспечение

Додж, Ю. (2004) Применение оптимизации. Springer, Lagrange Париж Multiplier Dodge, Y.:

Генератор естественных случайных чисел, Int. Стат. Обзор случайных чисел , 64, 329–344 Поколение (1996).

Додж, Ю.: Premiers pas en statistique. Спрингер, Париж (1999) Мера центральной тенденции, мера рассеяния

Додж, Ю. (ред.): Статистический анализ данных на основе нормы L1 и родственных методов. Эльзевир, Амстердам (1987) Оценка L1 Додж,

Ю., Хэнд, Д.: Каким должно быть статистическое программное обеспечение будущего? вычисл. Стат. Анализ данных. (1991) Статистическое программное обеспечение

Додж, Ю., Кондилис, А., Уиттакер, Дж.: Расширение PLS1 до регрессии PLAD и использование нормы L1 в мягком моделировании. COMPSTAT 2004, стр. 935–942, Physica/ Verlag-Springer (2004) Частичная регрессия с наименьшим абсолютным

отклонением Додж, Ю., Маджумдар, Д.: Алгоритм поиска обобщенных обратных наименьших квадратов для моделей классификации с произвольными шаблонами. Дж. Стат. вычисл. Симул. 9, 1–17 (1979) Обобщенная обратная, отсутствующие

данные Додж, Ю., Руссон, В.: Многомерное L1-среднее. Метрика 49, 127–134 (1999) Норма вектора

Додж, Ю., Шах, К.Р.: Оценка параметров в латинских квадратах и греко-латинских квадратах с отсутствующими наблюдениями. коммун. Стат. Теория А6(15), 1465–1472 (1977) Греко-латинский квадратный дизайн

Додж, Ю., Томас, Д.Р.: О производительности непараметрической и нормальной теории кратных процедуры сравнения. Санкхья В42, 11–27 (1980) Тест на наименьшие значимые различия

Додж, Ю., Уиттакер, Дж. (ред.): Вычислительный статистика, т. 1 и 2. КОМПСТАТ, Невшатель, Швейцария, 24–28 августа 1992 г. (на английском языке). Physica-Verlag, Гейдельберг (1992) Статистический Программное обеспечение

Додж, Ю., Уиттакер, Дж., Кондилис, А.: Прогресс в области PLAD и PQR. В: Труды 55-й сессии ISI 2005 г., стр. 495–503 (2005 г.). Частичный минимум Абсолютное отклонение регрессии

Додж, Ю.: Анализ приложения регрессии. Дюно, Критерий общего Париж (1999) среднего квадрата Ошибка

Додж, Ю.: Анализ экспериментов с отсутствующими данными. Уайли, Нью-Йорк (1985) Недостающие данные

Додж, Ю.: Математика базы для экономистов. Спрингер, Берлин Гейдельберг, Нью-Йорк (2002) Собственный вектор, норма вектора

Додж, Ю.: Основные планы опыта. В: Эшлиманн, Дж., Бонжур, К., Стокер, Э. (ред.) Методологии и методы планирования опыта: 28 курсов по совершенствованию Ассоциации водолазных шершёров телосложение, Саас-Фе, 2–8 марта 1986 г. (1986 г.) Факторный эксперимент

Додж, Ю.: Некоторые трудности, связанные с непараметрической оценкой функции плотности. Дж. Оффик. Стат. 2 (2), 193–202 (1986) Гистограмма

Дуб, Дж. Л.: Стохастические процессы. Уайли, Нью-Йорк (1953) Стохастический процесс

Дуб, Дж. Л.: Что такое стохастический процесс? Являюсь. Мат. Ежемесячно 49, 648–653 (1942) Стохастический Процесс

Дрейпер, Н.Р., Смит, Х.: Прикладная регрессия Анализ, 3-е изд. Уайли, Нью-Йорк (1998) Анализ остатков, нелинейная регрессия

Дробиш, MW: Über Mittelgrössen und die Anwendbarkeit derselben auf die Berechnung des Steigens und Sinkens des Geldwerts. Berichte über die Verhandlungen der König. Сакс. Гэс. Висс. Лейпциг. математика-физ. Классе, 23, 25 (1871 г.) Индекс

Droesbeke, JJ, Tassi, PH: Histoire de la Statistique. Издания «Que sais-je?» Presses universitaires de France, Париж (1990) Среднее геометрическое

Дюфур, IM: Histoire de l'analyse de Séries Хронологи, Университет Монреаля, Канада (2006) Временные ряды

Machine Translated by Google

Дункан, Г.Дж.: Панельные исследования домохозяйств: перспективы и проблемы. Методология социальных наук, Тренто, Италия (1992 г.) Панель

Дункан, Дж. В., Гросс, А. С.: Статистика для 21 века: предложения по улучшению статистики для лучшего принятия решений. Irwin Professional Publishing, Чикаго (1995) Официальная статистика

Дюпакье, Ж. и Дюпакье, М.: Histoire de la демография. Академия Перрен, Париж (1985) Официальная статистика

Дурбин Дж., Нотт М., Тейлор С.К.: Компоненты статистики Крамера-Фон Мизеса, II. Дж. Рой. Стат. соц. сер. Б 37, 216–237 (1975) Андерсон-Дарлинг Тест

Дурбин, Дж., Уотсон, Г.С.: Проверка последовательной корреляции в регрессии по методу наименьших квадратов, I. Biometrika 37, 409–428 (1950). Тест Дурбина-Ватсона

Дурбин, Дж., Уотсон, Г.С.: Тестирование последовательной корреляции в регрессии по методу наименьших квадратов, II. Биометрика 38, 159–177 (1951). Тест Дурбина-Ватсона, серийный Корреляция

Дурбин, Дж.: Метод, альтернативный d-критерию. 56, 1–15 (1969) Биометрический тест Дурбина-Ватсона

Дюто, К.: Политические размышления о финансах и торговле. Vaillant и Николя Прево, The Гаага (1738 г.) Индекс

Эджворт Ф.Ю.: Новый метод сокращения наблюдения, относящиеся к нескольким величинам. Филос. Маг. (5-я серия) 24, 222–223 (1887) Оценка L1

Эджворт, FY: Методы статистики. Юбилей Том Королевского статистического общества, Лондон (1885 г.) Проверка гипотез

Эдингтон, ES: рандомизационные тесты. Марсель Деккер, Нью-Йорк (1980) Повторная выборка

Эдвардс, AWF: Вероятность. Счет статистическая концепция вероятности и ее применение к научным выводам. Издательство Кембриджского университета, Кембридж (1972) Тест отношения правдоподобия

Эдвардс, Д.: Введение в графическое моделирование. Спрингер, Берлин Гейдельберг, Нью-Йорк (1995) Статистическое программное обеспечение

Эфрон Б., Тибширани Р.Дж.: Введение в Начальная загрузка. Чепмен и Холл, Нью-Йорк (1993) Bootstrap

Efron, B.: Методы начальной загрузки: еще один взгляд на складной нож. Анна. Стат. 7, 1–26 (1979) Bootstrap

Efron, B.: Складной нож, Bootstrap и другие планы передискретизации. Общество промышленных и Прикладная математика, Филадельфия (1982) Метод складного ножа

Эфрон, Б.: Складной нож, бутстрап и другие планы повторной выборки, монография CBMS. № 38. СИАМ, Филадельфия (1982) Повторная выборка

Эггенбергера, Ф. и Полиа, Г.: Über die Statistik verketteter Vorgänge. Zeitschrift für angewandte Mathematische Mechanik 3, 279–289 (1923) Отрицательное биномиальное распределение

Эйзенхарт, К.: Боскович и комбинация Наблюдения. В: Кендалл, М., Плакетт, Р.Л. (ред.) Исследования по истории статистики и вероятностей, том. II. Гриффин, Лондон (1977) Регрессивный анализ

Элдертон, В.П.: Таблицы для проверки соответствия теории наблюдениям. Биометрика 1, 155–163 (1902). Таблица хи-квадрат

Эйлер, Л.: Исследования по вопросу о неравенствах движения Сатурна и Юпитера, часть, о которой сообщается о премии 1748 года, парижская Королевская академия наук. Republié en 1960, в исполнении Леонарди Эйлери, Opera Omnia, 2ème série. Туричи, Базе, 25, стр. 47–157 (1749 г.) Анализ остатков, медиана, модель

Эйлер, Л.: Recherches sur une nouvelle espèce de carrés magiques. Верх. Zeeuw. Генерал Ветен. Флиссенген 9, 85–239 (1782 г.) Греко-латинский квадратный дизайн

Эвелин, сэр Г.С.: отчет о некоторых попытках установить стандарт веса и меры. Филос. Транс. 113 (1798) Индекс

Эверитт, Б.С.: Кластерный анализ. Холстед, Лондонская (1974) классификация, кластерный анализ, Анализ данных

Фехнер, Г.Т.: Коллективмасслере. В. Энгельманн, Лейпциг (1897 г.) Корреляция рангов Кендалла Коэффициент

Федерер, В.Т., Валаам, Л.Н.: Библиография по План эксперимента и лечения до 1968 года. Хафнер, Нью-Йорк (1973) План экспериментов

Федерер, WT: Сбор данных. В: Коц, С., Джонсон, Н.Л. (ред.) Энциклопедия статистических наук, том. 2. Уайли, Нью-Йорк (1982) Данных Федоров В.В. Теория оптимальных экспериментов. Академик, Нью-Йорк (1972) Оптимальный дизайн

Фестингер, Л.: Значение разницы между средними значениями без привязки к частоте

Machine Translated by Google

функция распределения. Психометрика 11, 97–105 (1946) Тест Манна – Уитни. Финберг, С. Е.: Графический метод в статистике. Являюсь. Стат. 33, 165–178 (1979) Графический Представление

Fienberg, SE: Анализ перекрестных классификаций Категориальные данные, 2-е изд. MIT Press, Кембридж, Массачусетс (1980) Таблица непредвиденных обстоятельств Финетти, Б. де: Теория вероятностей; Критическая вводная обработка. Транс. Мачи, А., Смит, А. Уайли, Нью-Йорк (1975) Байесовская статистика

Фишер, И.: Создание индексов. Хоутон Миффлин, Бостон (1922 г.) Индекс Фишера

Фишер, Р.А., Маккензи, Вашингтон: Манориал Реакция разных сортов картофеля. Дж. Сельское хозяйство. науч. 13, 311–320 (1923) Нелинейная регрессия

Фишер, Р.А., Йейтс, Ф.: Статистические таблицы для Биологические, сельскохозяйственные и медицинские исследования, 6-е изд. Хафнер (Макмиллан), Нью-Йорк (1963) Дизайн греко-латинских квадратов

Фишер, Р.А., Йейтс, Ф.: Статистические таблицы для Биологические, сельскохозяйственные и медицинские исследования. Оливер и Бойд, Эдинбург и Лондон (1963) Фишер Стол

Фишер, Р.А.: Применение «студенческого» распределения. Метрон 5, 90–104 (1925) Степень свободы, Распределение Студенчества, Таблица Студенчества, Тест Стьюдента

Фишер, Р.А.: Моменты и моменты продукта выборочные распределения. проц. Лонд. Мат. соц. 30, 199–238 (1929) Метод моментов

Фишер, Р.А.: Об абсолютном критерии подгонки частотные кривые. Беспорядок. Мат. 41, 155–160 (1912) Оценка, максимальное правдоподобие

Фишер, Р.А.: Об интерпретации χ2 из таблицы непредвиденных обстоятельств и расчет PJ Roy. Стат. соц. сер. А 85, 87–94 (1922) Критерий независимости Хи-квадрат

Фишер, Р.А.: О математических основах теоретическая статистика. Филос. Транс. Рой. соц. Лонд. сер. А 222, 309–368 (1922) Оценка, нелинейная регрессия

Фишер, Р.А.: Статистические методы и научные выводы. Оливер и Бойд, Эдинбург (1956) Выведенный статистика

Фишер, Р.А.: Статистические методы для научных работников. Оливер и Бойд, Эдинбург (1925) Дисперсионный анализ, блок, ковариация Анализ, распределение Фишера, взаимодействие, планы латинского квадрата, рандомизация, двусторонний дисперсионный анализ, дисперсия Фишер, Р.А.: статистические методы, экспериментальный план и научный вывод. Переиздание «Статистических методов для научных работников», «Планирование экспериментов» и «Статистические методы и научные выводы». Беннетт, Дж. Х., Йейтс, Ф. (ред.) Oxford University Press, Оксфорд (1990) Рандомизация

Фишер, Р.А.: План экспериментов. Оливер и Бойд, Эдинбург (1935) Гипотеза, взаимодействие, тест наименьших значимых различий, рандомизация, повторная выборка, дисперсия

Фишер, Р.А.: Организация полевых экспериментов. Дж. Министерство сельского хозяйства. 33, 503–513 (1926) Дизайн экспериментов, рандомизация

Фишер, Р.А.: Корреляция между родственниками при предположении менделевского наследования. Транс. Рой. соц. Эдинб. 52, 399–433 (1918) Дисперсия

Фишер, Р.А.: Моменты распределения для нормальных выборок мер отклонения от нормальности. проц. Рой. соц. Лонд. сер. А 130, 16–28 (1930) Метод моментов Фишер, Р.А.: Точность дискриминантных функций. Анна. Ойген. (Лондон) 10, 422–429 (1940) Анализ корреспонденции, анализ данных

Фишер, Р.А.: Теория статистического оценивания. проц. Камб. Филос. соц. 22, 700–725 (1925) Оценка

Фишер-Бокс, Дж.: Гиннесс, Госсет, Фишер и малый образцы. Стат. науч. 2, 45–52 (1987) Госсет, Уильям Сили

Фишер-Бокс, Дж.: Р. А. Фишер, жизнь ученого. Уайли, Нью-Йорк (1978) Фишер, Рональд Эйлмер

Фицморис, Г., Лэрд, Н., Уэр, Дж.: Прикладной Продольный анализ. Уайли, Нью-Йорк (2004) Продольные данные

Флюкигер, Ю.: Анализ социально-экономических различия кантоналес де chômage. Le chômage en Suisse: баланс и перспективы. Форум Хелвет. 6, 91–113 (1995) Индекс Джини

Флитвуд, В.: Chronicon Preciosum, 2-е изд. Флитвуд, Лондон (1707 г.) Индекс

Флетчер, Р.: Практические методы оптимизации. Уайли, Нью-Йорк (1997) Нелинейная регрессия

Фуди, В.Н., Хедагат, А.С.: О теории и применение конструкций BIB с повторяющимися блоками.

Machine Translated by Google

Анналы статистики, V = S 932–945 (1977)
Выборка

Фрэнсис И.: Статистическое программное обеспечение: сравнительный обзор.
Северная Голландия, Нью-Йорк (1981) Статистическое
программное обеспечение

Франк, И., Фридман, Дж.: Статистический взгляд на некоторые
инструменты хемометрической регрессии. Технометрика 35, 109–
135 (1993). Частичная регрессия методом
наименьших квадратов

Фредхольм, Дж.: Sur une classe d'equations
функции. Акта Математика. 27, 365–390 (1903)
Обобщенный обратный

Фридман Д., Пизани Р., Пурвес Р.: Статистика.
Нортон, Нью-Йорк (1978) Гистограмма

Фуллер, А.В.: Введение в статистические временные ряды.
Джон Вили'н'Сас (1976) Сезонная вариация

Функхаузер, Х. Г.: Заметка о графике 10 века.
Осирис 1, 260–262 (1936) Временные ряды

Гальтон Ф. Соотношения и их измерение.
главным образом по антропологическим данным. проц. Рой. соц.
Лонд. 45, 135–145 (1888 г.) Корреляция
Коэффициент, генетическая статистика

Гальтон, Ф.: Родство и корреляция. Север Ам. Откр. 150, 419–431 (1890
г.) Коэффициент корреляции

Гальтон, Ф.: Воспоминания о моей жизни. Метуэн, Лондон
(1908) Коэффициент корреляции

Гальтон, Ф.: Естественное наследование. Макмиллан, Лондонский
(1889) коэффициент корреляции, логический вывод
Статистика, регрессионный анализ

Гальтон, Ф.: Среднее геометрическое в естественной и социальной
статистике. проц. Рой. соц. Лонд. 29, 365–367 (1879 г.)
Логнормальное распределение

Гальтон, Ф.: Типичные законы наследственности. проц. Рой. Инст.
Great Britain 8, 282–301 (1877) (перепечатано в: Stigler, SM
(1986) The History of Statistics: The Measurement of Uncertainty Before
1900. Belknap Press, Cambridge, MA, p. 280)
Генетический
Статистика

Gauss, CF: Bestimmung der Genauigkeit der
Beobachtungen, том. 4, стр. 109–117 (1816 г.). В: Gauss, CF Werke
(опубликовано в 1880 г.).
Dieterichsche Universitäts-Druckerei, нормальное распределение
Геттингена , стандартная ошибка

Гаусс, К. Ф.: Работа Гаусса (1803–1826) по теории наименьших квадратов.
Троттер, HF, пер. Исследовательская группа статистических
методов, Технический отчет, № 5, Princeton University Press,
Принстон, Нью-Джерси (1957)
Стандартная ошибка

Гаусс, К. Ф.: В: Работа Гаусса (1803–1826) по теории наименьших
квадратов. Транс. Троттер, HF
Исследовательская группа статистических методов, Технический
отчет № 5. Издательство Принстонского университета,
Принстон, Нью-Джерси (1821 г.) Среднеквадратическая ошибка

Gauss, CF: Méthode des Moandres Carrés.
Mémoires sur la Combinaison des Observations.
Французский перевод Дж. Бертрана.
Малле-Башелье, Париж (1855 г.) Предвзятость

Гаусс, CF: Theoria Combinationis Observationum
Erroribus Minimis Obnoxiae, части 1, 2 и доп.
Верке 4, 1–108 (1821, 1823, 1826) Предвзятость,
Теорема Гаусса – Маркова.

Гаусс, CF: Theoria Motus Corporum Coelestium.
Верке, 7 (1809 г.) Ошибка, оценка,
Логическая статистика, нормальное распределение,
Регрессивный анализ

Гаварре, Ж.: Основные статистические принципы.
медицинский. Beché Jeune & Labé, Париж (1840 г.)
Проверка гипотез Гельман

А., Карлин Дж. Б., Штерн Х. С., Рубин Д. Б.:
Байесовский анализ данных. Чепмен и Холл / CRC, Лондон (2003)
Байесовская статистика

Гиббенс, Р.Дж., Пратт, Дж.В.: Интерпретация и методология Р-значений.
Являюсь. Стат. 29(1), 20–25 (1975) Значение p Гиббонс, Дж.
Д.,

Чакраборти, С.: Непараметрический статистический вывод, 4-
е изд. КПР (2003)
Непараметрическая статистика

М Вывод. Макгроу-Хилл, Нью-Йорк (1971)
Тест Уилкоксона

Глассер, Г.Дж., Винтер, РФ: Критические значения
коэффициент ранговой корреляции для проверки гипотезы
о независимости. Биометрика 48, 444–448 (1961)
Коэффициент ранговой корреляции Спирмена

Гувертс, П.: Геостатистика природных ресурсов.
Оценка. Издательство Оксфордского университета, Оксфорд
(1997) Геостатистика, Пространственные данные, Пространственная
статистика

Гордон, AD: Классификация. Методы исследовательского
анализа многомерных данных.
Чепмен и Холл, Лондон (1981)
Классификация, кластерный анализ Госсет,

SW («Студент»): Об ошибке подсчета гемоцитометром. Биометрика
5, 351–360 (1907).
Отрицательное биномиальное
распределение Госсет, ПО «Студент»: новые таблицы для проверки
значимости наблюдений. Метрон 5, 105–120

Machine Translated by Google

(1925) Студенческое распределение, Студенческий стол, Студенческий тест

Госсет С.В. «Студент»: вероятная ошибка Среднее. Биометрика 6, 1–25 (1908). логический Статистика, Моделирование, Стандартное отклонение, Студенческое распределение, Студенческий стол, Студент Тест

Gourieroux, C .: Theorie des sondages. Экономика, Париж (1981) Survey

Gower, JC: Классификация, геометрия и данные анализ. В: Бок, Х.Х. (ред.) Классификация и родственные методы анализа данных. Эльзевир, Амстердам (1988) Анализ данных

Граунт, Дж.: Естественные и политические наблюдения упомянуты в следующем указателе и сделаны в списках смертных: в отношении правительства, религии, торговли, роста, жизни, болезней и различных изменений в указанном городе. То. Ройкрофт для Джона Мартина, Джеймса Аллестри и Тхо. Дикас (1662 г.) Биостатистика

Грейбилл, Ф.А.: Теория и приложения линейной модели. Даксбери, Северный округ, Массачусетс (Уолдсворт и Брукс / Коул, Пасифик-Гроув, Калифорния) (1976) Теорема Гаусса – Маркова, обобщенная линейная Регрессия, Матрица

Гринакр, М.: Теория и приложения Анализ корреспонденции. Академический, Лондон (1984) Анализ корреспонденции

Гринвальд, Д. (ред.): Encyclopédie économique. Economica, Париж, стр. 164–165 (1984). Индекс Джини

Гринвуд, Дж. А., Хартли, Х.О.: Руководство по таблицам математической статистики. Издательство Принстонского университета, Принстон, Нью-Джерси (1962) Нормальная таблица Гринвуд, М., Юл, Г.У.: Исследование характер частотного распределения, репрезентативного для множественных событий, с особой ссылкой на возникновение множественных приступов болезни или повторяющихся несчастных случаев. Дж. Рой. Стат. соц. сер. А 83, 255–279 (1920) Отрицательное биномиальное распределение

Гринвуд, М.: Медицинская статистика от Граунта до Фарр. Издательство Кембриджского университета, Кембридж (1948) Биостатистика

Grosbras, JM: Статистические методы зондирования. Экономика, Париж (1987) Survey

Guerry, AM: Essai sur la statistique моральный дух Франции. Крошар, Париж (1833 г.) Графическое представление

Хаберман, С.Дж.: Анализ качественных данных. Том. I: Вводные темы. Академик, Нью-Йорк (1978) Анализ категориальных данных Хэдли

Г.: Линейное программирование. Аддисон-Уэсли, Рединг, Массачусетс (1962) Зал линейного программирования , П.: Расширение Bootstrap и Edgeworth. Спрингер, Берлин Гейдельберг, Нью-Йорк (1992) Bootstrap

Hand, DJ: Дискриминация и классификация. Уайли, Нью-Йорк (1981) Мера непохожести

Хансен, М. Х., Гурвиц, В. Н., Мэдоу, М. Г.: Методы выборочного обследования и теория. Том. I. Методы и приложения. Том. II. Теория. Чепмен и Холл, Лондон (1953) Кластерная выборка

Hardy, GH: менделевские пропорции в смешанном населении. Наука 28, 49–50 (1908) Генетический Статистика

Харнетт, Д.Л., Мерфи, Дж.Л.: Вводный статистический анализ. Аддисон-Уэсли, Рединг, Массачусетс (1975) Нерегулярная вариация, размер выборки

Хартер, Х.Л., Оуэн, Д.Б.: Избранные таблицы в Математическая статистика, т. 1, с. 1. Маркхэм, Чикаго (5.7 Приложение) (1970) Подписанная таблица Уилкоксона

Гарвардский университет: таблицы кумулятивных Биномиальное распределение вероятностей, том. 35. Летопись вычислительной лаборатории, издательство Гарвардского университета, Кембридж, Массачусетс (1955). Биномиальное распределение, Биномиальная таблица

Харви, АС: Эконометрический анализ временных рядов. Филип Аллан, Оксфорд (Уайли, Нью-Йорк) Тест Дарбина-Ватсона (1981)

Хаузер, П. М., Дункан, О. Д.: Изучение населения: инвентаризация и оценка. University of Chicago Press, Чикаго, Иллинойс (1959) Демография

Хейс, А. Р.: Статистическое программное обеспечение: обзор и критика его развития. Управление военно-морских исследований, Арлингтон, Вирджиния (1982 г.) Статистический Программное обеспечение

Hecht, J .: L'idee du dénombrement jusqu'à la Революция. Pour une histoire de la statistique, том 1, стр. 21–82. Экономика/INSEE (1978) Перепись

Хедаят, А.С., Синха, Б.К.: Дизайн и вывод в выборке конечной совокупности. Уайли, Нью-Йорк (1991) Выборка

Хелланд: О структуре регрессии PLS, Comm. в Стат-Симул. и комп., 17, 581–607 (1988). Частичная регрессия методом наименьших квадратов

Machine Translated by Google

Heyde, CE, Seneta, E .: IJ Bienaymé. Ожидается статистическая теория. Спрингер, Берлин Гейдельберг, Нью-Йорк (1977) Чебышев Пафнутий Львович

Hirschfeld, HO: Связь между корреляцией и непредвиденными обстоятельствами. проц. Камб. Филос. соц. 31, 520–524 (1935) Анализ корреспонденции, анализ данных

Хоаглин, округ Колумбия, Мостеллер, Ф., Тьюки, Дж. В. (ред.): Понимание надежного и исследовательского анализа данных. Уайли, Нью-Йорк (1983) График QQ (график квантиля к квантилю)

Hoaglin, DC, Welsch, RE: Матрица шляпы в регрессия и дисперсионный анализ. Являюсь. Стат. 32, 17–22 (и исправление на 32, 146) (1978) Шляпа Матрица

Хорл А.Е., Кеннард Р.В., Болдуин К.Ф.: Регрессия хребта: некоторые симуляции. коммун. Стат., вып. 4, стр. 105–123 (1975). Ридж-регрессия

Хорл А.Е., Кеннард Р.В.: Ридж-регрессия: смещенная оценка для неортогональных задач. Технометрика 12, 55–67 (1970). Ридж-регрессия

Hoerl AE: Применение гребневого анализа к Проблемы регрессии. хим. англ. прог. 58 (3), 54–59 (1962) Ридж-регрессия

Хогг, Р.В., Крейг, А.Т.: Введение в математическую статистику, 2-е изд. Макмиллан, Нью Йорк (1965) Оценщик, совместное распределение Функция, Функция предельной плотности, Пара Случайные переменные

Холгейт, П.: Влияние работы Гюйгенса на Динамика его вклада в вероятность. Междунар. Стат. 52 , 137–140 (1984) Вероятность

Хосмер Д.В., Лемешоу С.: Прикладная логистическая регрессия. Ряды Wiley в теории вероятностей и статистике. Уайли, Нью-Йорк (1989) Логистическая регрессия

Хубер, П.Дж.: Надежная статистика. Wiley, New York Leverage (1981) Point, надежная оценка

Хубер, П.Дж.: Надежная оценка местоположения параметр. Анна. Мат. Стат. 35, 73–101 (1964) Надежная оценка

Хубер, П.Дж.: Надежность: где мы сейчас? Студент, об. 1, нет. 2, стр. 75–86 (1995). Крепкий Оценка

Huitema, BE: Анализ ковариации и альтернатив. Уайли, Нью-Йорк (1980) Ковариационный анализ

INSEE: Pour une histoire de la statistique. Том. 1: Взносы. INSEE, Economica, Париж (1977) Официальная статистика

INSEE: Pour une histoire de la statistique. Том. 2: Материалы. Дж. Аффишар (редактор) INSEE, Economica, Париж (1987 г.) Официальная статистика

ISI: Будущее статистики: международный Перспектива. Ворбург (1994) Официальная статистика

Международные математические и статистические библиотеки: Информация о библиотеке IMSL. IMSLK, Хьюстон (1981) Статистическое программное обеспечение

Исаакс, Э. Х., Шривастава, Р. М.: Введение в прикладную геостатистику. Издательство Оксфордского университета, Оксфорд (1989) Геостатистика, Пространственная Статистика

Жаффе, В .: Вальрас, Леон. В: Силлс, Д.Л. (ред.) Международная энциклопедия социальных наук, том. 16. Macmillan and Free Press, Нью-Йорк, стр. 447–453 (1968). Эконометрика

Джамбу, М., Лебо, Миссури: Классификация автоматический анализ данных. Дюно, Париж (1978) Кластерный анализ

Джевонс, В. С.: Установлено серьезное падение стоимости золота и его социальные последствия. Стэнфорд, Лондон (1863 г.) Индекс

Джевонс, В. С.: Принципы науки: трактат о логике и научных методах (два тома). Макмиллан, Лондон (1874 г.) Среднее геометрическое

Джоанс, Д. Н. и Гилл, Калифорния: Сравнение показателей асимметрии пробы и эксцесса, JRSS.D 47 (1), 183–189 (1998). Мера эксцесса

Джонсон, Н.Л., Коц, С., Балакришнан, Н.: Дискретные многомерные распределения, Джон Вили (1997) Распределение вероятностей

Джонсон, Н.Л., Коц, С.: Распределения в статистике: Непрерывные одномерные распределения, тома. 1 и 2. Уайли, Нью-Йорк (1970) Непрерывный Распределение вероятностей, логнормальное Распределение, вероятностное распределение

Джонсон, Н.Л., Коц, С.: Распределения в статистике: Дискретные распределения. Уайли, Нью-Йорк (1969) Дискретное распределение вероятностей, вероятность Распределение

Джонсон, Н.Л., Леоне, ФК: статистика и экспериментальный дизайн в инженерных и физических науках. Уайли, Нью-Йорк (1964) Коэффициент вариации

Machine Translated by Google

Journel, A., Huijbregts, CJ: Горная геостатистика.
Академик, Нью-Йорк (1981) Геостатистика,
пространственная
статистика Каарсемакер, Л., ван Вейнгаарден, А.: Таблицы для
использования в ранговой корреляции. Стат. Нирланд. 7, 53 (1953)
Коэффициент ранговой корреляции Кендалла

Каптейн, Дж., ван Увен, М.Дж.: Кривые перекоса частоты в
биологии и статистике. Братья Хойтсема, Гронинген
(1916 г.) Логнормальное распределение

Каптейн, Дж.: Кривые частоты перекоса в биологии и статистике.
Астрономическая лаборатория Нордхофф, Гронинген (1903
г.) Логнормальное распределение

Каспршик Д., Дункан Г. и соавт. (ред.): Панельные опросы.
Нью-Йорк, Чичестер, Брисбен, Торонто, Сингапур.
Уайли, Нью-Йорк (1989) Панель

Кауфман, Л., Руссеу, П.Дж.: Поиск групп в
Данные: Введение в кластерный анализ. Уайли, Нью-Йорк
(1990) Классификация, Кластер
Анализ

Кендалл, М.Г., Стюард, А.: Расширенная теория статистики, том. 2.
Гриффин, Лондон (1967)
Тест отношения правдоподобия

Кендалл, М.Г., Стюард, А.: Расширенная теория статистики. Том. 1.
Теория распределения, 4-е изд.
Гриффин, Лондон (1977) Метод моментов

Кендалл, М.Г.: Новая мера ранговой корреляции.
Биометрика 30, 81–93 (1938). Кендалл Ранг
Коэффициент корреляции

Кендалл, М. Г.: Методы ранговой корреляции. Грифон,
Лондон (1948) Корреляция рангов Кендалла
Коэффициент

Кендалл, М.Г.: Исследования по истории вероятностей и
статистики: II. Начало исчисления вероятностей. Биометрика
43, 1–14 (1956).
Вероятность

Кендалл, MG: Ранняя история индексов.
В: Кендалл, М., Плакетт, Р.Л. (ред.) Исследования по истории
статистики и вероятностей, том. II. Гриффин, Лондон (1977)
Индекс

Кендалл, MG: временные ряды. Гриффин, Лондон (1973)
Ковариация, временной ряд

Кендалл, М.Г.: С чего должна начаться история статистики?
Биометрика 47, 447–449 (1960).
Статистика

Хорезми, Муса ибн Меусба (9 век). Джабр
ва-аль-мукабала. Алгебра Мохаммеда бен Мусы, Розен, Ф. (изд.
и перевод). Георг Олмс Верлаг, Хильдесхайм (1986)
Алгоритм

Кифер, Дж.: Оптимальные экспериментальные планы. Дж. Рой.
Стат. соц. сер. Б 21, 272–319 (1959) Кифер,
Джек Карл, Оптимальный дизайн

Кляйнен, JPC: Статистические методы моделирования (в двух
частях). Том. 9 статистики: учебники и монографии. Марсель
Деккер, Нью-Йорк (1974)
Моделирование

Кляйнен, JPC: Статистические инструменты для моделирования
Практики. Том. 76 статистики: учебники и
Монографии. Марсель Деккер, Нью-Йорк (1987)
Моделирование

Найт, К.: Математическая статистика. Чепмен и совместная
Холл/CRC, Лондон (2000) плотность
Функция

Кокоска С., Цвиллингер Д.: Стандарт CRC
Вероятность и статистика Таблицы и формулы.
КПР Пресс (2000) Статистическая таблица

Колмогоров А.Н.: Основы теории вероятностей. Издательская
компания Челси, Нью-Йорк (1956)
Условная возможность

Колмогоров, А.Н.: Grundbegriffe der
Wahrscheinlichkeitsrechnung. Springer, Берлин Гейдельберг,
Нью-Йорк (1933) Условный
Вероятность

Колмогоров А.Н.: Матем. Сборник. НС 1, 607–610
(1936) Стохастический процесс

Колмогоров, А.Н.: Эмпирическое определение пути распределения.
Giornale dell'Instituto Italiano degli Attuari 4, 83–91 (6.1) (1933)
Критерий Колмогорова-Смирнова, непараметрический
Тест

Кондилис, А., Уиттакер, Дж.: Использование Bootstrap для
регрессии PLAD. PLS и родственные методы. В: Труды
Международного симпозиума PLS'05 2005 г., стр.
395–402. Издание Aluja et al. (2005)
Частичное наименьшее абсолютное отклонение
Регрессия

Коц, С., Джонсон, Н.Л.: Статистическая энциклопедия.
наук, тт. 1–9. Уайли, Нью-Йорк (1982–1988)
Статистика

Крускал, WH, Танур, JM: Международный
Энциклопедия статистики, тт. 1–2. Свободная пресса, Нью-
Йорк (1978) Статистика

Kruskal, WH, Wallis, WA: Использование рангов в
однокритериальный дисперсионный анализ. Варенье. Стат.
доц. 47, 583–621 и опечатки, там же. 48, 907–911 (1952)
Таблица Краскела-Уоллиса, критерий Краскела-Уоллиса,
Непараметрический тест

Machine Translated by Google

Л. Д. Браун, И. Олкин, Дж. Сакс и Х. П. Винн (ред.): Джек Карл Кифер, Сборник статей. I: Статистический вывод и вероятность (1951–1963). Нью-Йорк (1985) Кифер, Джек Карл

Л.Д. Браун, И. Олкин, Дж. Сакс и Х. П. Винн (ред.): Джек Карл Кифер, Сборник статей. II: Статистический вывод и вероятность (1964–1984). Нью-Йорк (1985) Кифер, Джек Карл

Л.Д. Браун, И. Олкин, Дж. Сакс и Х. П. Винн (ред.): Джек Карл Кифер, Сборник статей. III: План экспериментов. Нью-Йорк (1985) Кифер, Джек Карл

Лагард, Ж. де: Инициация анализа донна. Дюно, Париж (1983) Матрица инерции

Лагранж, JL: Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs наблюдения; данс lequel на изучить ле avantages де сетте метод пар ле исчисления вероятностей; et où l'on résoud différents problèmes relatifs à cette matière. Разное Тауриненсия 5, 167–232 (1776 г.) Ошибка

Ланкастер, НО: предшественники Пирсона хи-квадрат. Ауст. Дж. Стат. 8, 117–126 (1966) Распределение хи-квадрат, гамма-распределение

Лаплас, PS de: Deuxième supplément à la théorie аналитика вероятностей. Курсье, Париж (1818 г.). стр. 531–580 dans les Œuvres complètes de Laplace, vol. 7. Gauthier-Villars, Париж, 1886 г. Надежная оценка

Лаплас, PS de: Mémoire sur l'inclinaison moyenne des orbites des comètes. Мем. акад. Рой. науч. Париж , 7, 503–524 (1773 г.) Проверка гипотез

Лапласа, PS de: Mémoire sur la probabilité des вызывает par les événements. Мем. акад. Рой. науч. (представлено разными учеными) 6, 621–656 (1774) (или Laplace, PS de (1891) Œuvres complètes, vol 8. Gauthier-Villars, Paris, pp. 27–65) Ошибка, Оценка, Распределение Лапласа, Нормальный Распределение, обычная таблица

Лаплас, П.С. о: Память о приближениях формул, которые выполняют функции трех великих имен и о возможных приложениях. Мемуары Королевской академии наук Парижа, 10.

Воспроизведено в: Œuvres de Laplace 12, 301–347 (1810 г.). Центральная предельная теорема, теорема Муавра – Лапласа.

Лаплас, PS de: Mémoire sur les probabilités. Мем. акад. Рой. науч. Париж, 227–332 (1781) (или Лаплас, PS de (1891) Œuvres complètes, том 9. Gauthier-Villars, Париж, стр. 385–485.) Ошибка, Нормальная таблица

Лаплас, PS de: Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pendule. Histoire de l'Académie royale des inscriptions et belles lettres, avec les Mémoires de littérature tirées des registres de Cette académie. Париж (1789 г.) Регрессионный анализ

Лапласа, PS de: Аналитическая теория вероятностей, 3-е изд. Курсье, Париж (1820 г.) Среднеквадратическая ошибка

Laplace, PS de: Theorie Analytique des Probabilités, suppl. к 3-му изд. Курсье, Париж (1836 г.) Гамма-распределение

Лаплас, PS de: Аналитическая теория вероятностей. Курсье, Париж (1812 г.) Оценка, Выведенный статистика

Ласпейрес, Э.: Die Berechnung einer mittleren Waarenpreissteigerung. Ярб. Натл. Стат. 16, 296–314 (1871 г.) Индексный номер, Ласпейрес Индекс

1863 г. и смерть калифорнийско-австралийских Geldentdeckung seit 1848. Jahrb. Натл. Стат. 3, 81–118, 209–236 (1864 г.) Индекс, Индекс Ласпейреса

Ле Кам, Л.М., Ян, К.Л.: Асимптотика в статистике: некоторые основные понятия. Спрингер, Берлин Гейдельберг, Нью-Йорк (1990) Конвергенция

Ле Кам, Л.: Центральная предельная теорема вокруг 1935. Стат. науч. 1, 78–96 (1986) Центральный предел Теорема

Лебар, Л., Салем, А.: Статистический анализ текстовых материалов Données. Дюно, Париж (1988) Анализ корреспонденции

Лежандр, АМ: Nouvelles méthodes pour la определение орбит комет. Курсье, Париж (1805 г.) Оценка, метод наименьших квадратов, нормальные уравнения, регрессионный анализ

Леманн, Эрих Л., Казелла, Г.: Теория точечной оценки. Спрингер-Верлаг, Нью-Йорк (1998) Оценка точки

Леманн, Э.И., Романн, С.П.: Статистическое тестирование Гипотеза, 3-е изд. Спрингер, Нью-Йорк (2005) Альтернативная гипотеза

Леманн, Э.Л.: Проверка статистических гипотез, 2-я гипотеза изд. Уайли, Нью-Йорк (1986) Тестирование

Machine Translated by Google

Леманн, Э.Л.: Теория точечной оценки, 2-е изд.
Уайли, Нью-Йорк (1983) Оценщик,
параметр

Лерман, Л.С.: Классификация и анализ порядкового номера
Донни. Дюно, Париж (1981) Кластерный анализ

Ляпунов, А. М.: Sur une proposition de la théorie
вероятностей. Вестник Императорской академии наук Санкт-
Петербурга 8, 1–24 (1900)
Центральная предельная теорема

Lilienfeld, AM, Lilienfeld, DE: Основы
Эпидемиология, 2-е изд. Кларендон, Оксфорд (1980)
Атрибутивный риск, предотвратимый риск, причина
и эффект в эпидемиологии, эпидемиологии,
Уровень заболеваемости, шансы и отношение шансов,
Уровень распространенности, относительный риск, риск

Линдли, Д.В.: Философия статистики.
Статистик 49, 293–337 (2000). Байесовская
статистика

Липпс, Г.Ф.: Die Psychischen Massmethoden. Ф.
Vieweg und Sohn, Брауншвейг, Германия (1906 г.)
Коэффициент ранговой корреляции Кендалла

Лоу, Дж.: Современное состояние Англии в отношении сельского
хозяйства, торговли и финансов. Келли, Лондон (1822 г.)
Индекс

МакМахон, Б., Пью, Т. Ф.: Эпидемиология: принципы и методы.
Литтл Браун, Бостон, Массачусетс (1970)
Атрибутивный риск, предотвратимый риск, причина
и эффект в эпидемиологии, эпидемиологии,
Уровень заболеваемости, шансы и отношение шансов,
Уровень распространенности, относительный риск, риск

МакМахон, Пенсильвания: комбинаторный анализ, тт. я
и II. Издательство Кембриджского университета, Кембридж
(1915–1916) Комбинаторный анализ

Маколи, Ф. Р.: Сглаживание временных рядов.
Национальное бюро экономических исследований, 121–136.
(1930) Временные ряды

Махаланобис, ПК: Об обобщенном расстоянии в статистике,
Proc. физ. Инст. науч. Индия 2, 49–55
(1936) Расстояние Махаланобиса

Махаланобис, ПК: об испытаниях и измерениях
Расхождение групп. Часть I. Теоретические формулы.
J. Asiatic Soc. Бенгалия 26, 541–588 (1930)
Расстояние Махаланобиса

Майстров, Л.Е.: Теория вероятностей — история
Эскиз (перевод и редакция Kotz, S.). Академик, Нью-Йорк
(1974) Индепенденс

Маллоуз, CL: Выбор переменных в линейной регрессии:
графическое пособие. Представлено на Центральном
региональном собрании Института
Математическая статистика, Манхэттен, Канзас, 7–9 мая.
1964 (1964) Критерий Ср

Маллоус, К.Л.: Некоторые комментарии к Ср.
Технометрика 15, 661–675 (1973). СП
Критерий

Манн, Х.Б., Уитни, Д.Р.: О проверке того, является ли одна из двух
случайных величин стохастически больше другой. Анна. Мат.
Стат. 18, 50–60 (1947)
Критерий Манна-Уитни, непараметрический критерий

Марков, А.А.: Применение функций продолжается в расчете
вероятностей. Касан. Бык. 9 (2), 29–34 (1899 г.)
Марков Андрей Андреевич

Марков, А.А.: Расширение теоретических пределов исчисления
вероятностей относительно некоторых количественных
показателей лжи в цепи. Мем. акад. науч. СПб. 8, 365–397
(1908). Центральная предельная теорема

Мартенс Х. и Нэс Т.: Многомерная калибровка.
Уайли, Нью-Йорк (1989) Частичный метод наименьших квадратов
Регрессия

Massey, FJ: Таблица распределения отклонений
между двумя выборочными кумулятивами. Анна. Мат. Стат. 23,
435–441 (1952) Тест Колмогорова –
Смирнова.

Матерон, Г.: Теория региональных переменных и ее приложений.
Массон, Париж (1965)
Геостатистика, Пространственные данные, Пространственная статистика

Майер, Т .: Abhandlung über die Umwälzung des Monds um
seine Achse und die scheinbare
Bewegung dep Mondfleken. Kosmographische Nachrichten
und Sammlungen auf das Jahr 1748 1, 52–183 (1750)
Анализ остатков, медиана,
модель

Майр, Э., Линсли, Э.Г., Узингер, Р.Л.: Методы и принципы
систематической зоологии. Макгроу-Хилл, Нью-Йорк (1953)
Дендрограмма

Макалистер, Д.: Закон среднего геометрического. проц.
Рой. соц. Лонд. 29, 367–376 (1879 г.)
Логнормальное распределение

Макмаллен, Л., Пирсон, Э.С.: Уильям Сили Госсет, 1876–1937. В:
Пирсон, Э.С., Кендалл, М. (ред.)
Исследования по истории статистики и вероятностей, том. И.
Гриффин, Лондон (1970) Госсет, Уильям Сили

Меррингтон, М., Томпсон, К.М.: Таблицы процентных
пунктов обратного бета- (F) -распределения.
Биометрика 33, 73–88 (1943).
Фишер Стол

Machine Translated by Google

Метрополис, Н., Улам, С.: Метод Монте-Карло. Варенье. Стат. доц. 44, 335–341 (1949) Монте Метод Карло

Мейер, компакт-диск: матричный анализ и прикладная линейная алгебра, SIAM (2000 г.) Собственный вектор Миллера, LH: Таблица процентных пунктов Статистика Колмогорова. Варенье. Стат. доц. 31, 111–121 (1956) Тест Колмогорова-Смирнова

Миллер, Р.Г., младший: Одновременный статистический вывод, 2-е изд. Спрингер, Берлин Гейдельберг, Нью-Йорк (1981) Тест наименьших значимых различий, односторонний тест

Митринович, Д.С. (с Васичем, П.М.): Аналитик Неравенства. Спрингер, Берлин Гейдельберг, Нью-Йорк (1970) Гармоническое среднее

Moivre, A. de: Approximatio ad summam terminorum binomii (a + b)n, in seriem expansi. Приложение II к Miscellanae Analytica, стр. 1–7 (1733 г.). Фотографически перепечатано в редкой брошюре о Муавре и некоторых его открытиях. Опубликовано Арчибальдом, RC Isis 8, 671–683 (1926). Центральная предельная теорема, теорема Муавра – Лапласа, нормальное распределение

Муавр, А. де: Доктрина шансов: или, Метод расчета вероятности событий в игре. Пирсон, Лондон (1718 г.) логический Статистика, нормальное распределение

Моленбергс, Г., Герт, В.: Модели для дискретных продольных рядов данных. Серия Спрингера по статистике. Спрингер, Берлин Гейдельберг, Нью-Йорк (2005) Продольные данные

Монтгомери, округ Колумбия: Дизайн и анализ эксперименты, 4-е изд. Уайли, Чичестер (1997) Односторонний анализ дисперсии

Montmort, PR de: Essai d'analyse sur les jeux de hasard, 2-е изд. Кийо, Париж (1713 г.) Настроение отрицательного биномиального распределения , AM: Введение в теорию статистики. Макгроу-Хилл, Нью-Йорк (1950) Непараметрический тест

Мура, ЕН: Общий анализ, часть I. Мем. Являюсь. Филос. Soc., Филадельфия, Пенсильвания, стр. 147–209 (1935). Обобщенный обратный

Морабия, А.: Причинная эпидемиология. Издания Médecine et Hygiène, Женева (1996 г.) Атрибутивный риск, предотвратимый риск, причина и эффект в эпидемиологии, эпидемиологии, Уровень заболеваемости, шансы и отношение шансов, Уровень распространенности, относительный риск, риск

Морабия, А .: L'Epidémiologie Clinique. Издания «Que sais-je?». Presses Universitaires de France, Париж (1996) Атрибутивный риск, предотвратимый риск, уровень заболеваемости, шансы и отношение шансов, уровень распространенности, относительный риск, риск

Мостеллер, Ф., Тьюки, Дж. В.: Анализ данных и регрессия: второй курс статистики. Аддисон-Уэсли, Рединг, Массачусетс (1977) Коэффициент корреляции, точка рычага Мудхолкар, GS: Множественный коэффициент корреляции. В: Коц, С., Джонсон, Н.Л. (ред.) Энциклопедия статистических наук, том. 5. Уайли, Нью-Йорк (1982) Коэффициент корреляции

Национальное бюро стандартов.: Путеводитель по таблицам нормальный интеграл вероятности. Министерство торговли США. Серия прикладной математики, 21 (1952) Обычная таблица

Национальное бюро стандартов.: Таблицы бинома Распределение вероятностей. Министерство торговли США. Прикладная математика, серия 6 (1950) Биномиальное распределение, биномиальная таблица Нейлор, Т. Х., Балинтфи, Дж. Л., Бердик, Д. С., Чу, К.: Методы компьютерного моделирования. Уайли, Нью-Йорк (1967) Моделирование

Нейман Дж., Пирсон Э.С.: К проблеме наиболее эффективные проверки статистических гипотез. Филос. Транс. Рой. соц. Лонд. сер. А 231, 289–337 (1933) Проверка гипотез

Нейман Дж., Пирсон Э.С.: Об использовании и интерпретация некоторых тестовых критериев для целей статистического вывода, часть I (1928 г.). Перепечатано издательством Кембриджского университета в 1967 г. Уровень значимости

Нейман, Дж., Пирсон, Э.С.: Об использовании и интерпретации некоторых критериев тестирования для целей статистического вывода, Части I и II. Биометрика 20A, 175–240, 263–294 (1928). Критическое значение, Проверка гипотезы, Логическая статистика, Тест отношения правдоподобия, Ошибка первого рода

Нордин, Дж. А.: Определение размера выборки. Варенье. Стат. доц. 39, 497–506 (1944) Размер образца

OCDE: OCDE и международные индикаторы l'enseignement. Centre de la Recherche et l'innovation dans l'enseignement, OCDE, Париж (1992) Индикатор

OCDE: Принципы указателей науки и науки технология. OCDE, Париж (1988) Индикатор

Оливер, М.А., Вебстер, Р.: Кригинг: метод интерполяция для географической информационной системы.

Machine Translated by Google

Междунар. Дж. Геогр. Инф. Сист. 4 (3), 313–332 (1990)
Геостатистика, пространственная статистика Остле, Б.: Статистика в исследованиях: основные концепции и методы для научных работников. Издательство государственного колледжа Айовы, Эймс, Айова (1954) Хи-квадрат, контраст

Оуэн, Д.Б.: Справочник по статистическим таблицам. (3.3, 5.7, 5.8, 5.10). Аддисон-Уэсли, Рединг, Массачусетс (1962)
Подписанная таблица Уилкоксона

Пааше, Х.: Über die Preisentwicklung der letzten Jahre nach den Hamburger Borsen-notirungen.
Ярб. Натл. Стат. 23, 168–178 (1874 г.) Номер индекса, индекс Пааше

Palgrave, RHI: Валюта и стандарт стоимости в Англии, Франции, Индии и т. Д. (1886 г.) В: IUP (1969) Британские парламентские документы: сессия 1886 г.: Третий отчет и окончательный отчет Королевской комиссии по депрессии в торговле и промышленности, с протоколом свидетельских показаний и приложениями, Приложение В. Издательство Ирландского университета, индекс Шеннона.

Панатье, Ю.: Вариовин 2.2. Программное обеспечение для анализа пространственных данных в 2D. Спрингер, Берлин Гейдельберг, Нью-Йорк (1996) Геостатистика

Парзен, Э.: Об оценке функции и режима плотности вероятности. Анна. Мат. Стат. 33, 1065–1076 (1962)
Непараметрическая статистика

Паскаль, Б .: Месснард, Дж. (ред.) Œuvres complètes.
Том. 2. Дескле де Брауэр, Париж (1970)
Арифметический треугольник Паскаля, Б.: Трактат по арифметике треугольника (опубл. постум. в 1665 г.), Париж (1654 г.) Арифметика
Треугольник

Паскаль, Б.: Varia Opera Mathematica. Д. Петри де Ферма. Толосы (1679 г.) Отрицательное биномиальное распределение

Паскаль, Б .: Œuvres, тт. 1–14. Брунсвик, Л.,
Бутру, П., Газье, Ф. (ред.) Les Grands Ecrivains de France. Ашетт, Париж (1904–1925)
Арифметический треугольник Пирсон, Э.С., Хартли, Х.О.: Таблицы биометрии для статистиков, 1 (2-е изд.). Издательство Кембриджского университета, Кембридж (1948) Нормальная таблица

Пирсон, Э.С., Хартли, Х.О.: Биометрические таблицы для статистиков, тома. I и II. Издательство Кембриджского университета, Кембридж (1966, 1972) Коэффициент асимметрии

Пирсон, Э.С.: История Неймана-Пирсона: 1926–1934.
В: Дэвид, Ф. Н. (ред.) Исследовательские работы по статистике: Festschrift для Дж. Неймана. Уайли, Нью-Йорк (1966)
Проверка гипотезы

Пирсон, К., Филон, СПГ: математический
вклад в теорию эволюции. IV: О возможных ошибках констант частоты и о влиянии случайного выбора на вариацию и
корреляция. Филос. Транс. Рой. соц. Лонд. сер. А 191, 229–311 (1898 г.) Оценка

Пирсон, К.: Вклад в математическую теорию эволюции. I. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, Кембридж, стр. 1–40 (1948). Впервые опубликовано в 1894 году под названием: «О анализе асимметричных частотных кривых».
Филос. Транс. Рой. соц. Лонд. сер. А 185, 71–110 Коэффициент асимметрии, оценка, мера формы, метод моментов, стандартное отклонение

Пирсон, К.: Вклад в математическую теорию эволюции. II: Косая вариация в однородном материале. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, Кембридж, стр. 41–112 (1948). Впервые опубликовано в 1895 году в Philos. Транс. Рой. соц. Лонд.
сер. А 186, 343–414 Коэффициент асимметрии, частотная кривая, гистограмма, мера формы

Пирсон, К.: О критерии, что данная система отклонений от вероятного в случае коррелированной системы переменных такова, что можно разумно предположить, что она возникла в результате случайной выборки. В: Ранние статистические статьи Карла Пирсона. Издательство Кембриджского университета, стр. 339–357. Впервые опубликовано в 1900 году в Philos.
Маг. (5-я серия) 50, 157–175 (1948) Распределение хи-квадрат, критерий согласия хи-квадрат, критерий независимости хи-квадрат, логический вывод
Статистика

Пирсон, К.: О теории случайности и ее связи с ассоциацией и нормальной корреляцией.
Мемуары исследований компании Drapers ', биометрическая сер. I., стр. 1–35 (1904 г.) Анализ категорийных данных, таблица непредвиденных обстоятельств

Пирсон, К.: Исследования по истории статистики и вероятностей.
Биометрика 13, 25–45 (1920).
Перепечатано в: Пирсон, Э.С., Кендалл, М.Г. (ред.)
Исследования по истории статистики и вероятностей, том. И.
Griffin, Лондон Коэффициент корреляции

Пирсон, К.: Таблицы неполных -функций.
Офис канцелярских принадлежностей HM (Кембриджский университет

Machine Translated by Google

Press, Кембридж с 1934 г.), Лондон (1922 г.)
Таблица хи-квадрат

Пирс, Б.: Критерий отклонения сомнительных наблюдений.
Астрон. Журнал 2, 161–163 (1852 г.)
Выброс

Пенроуз, Р.: Обобщенная обратная для матриц. проц.
Камб. Филос. соц. 51, 406–413 (1955)
Обобщенный обратный

Персоны, WM: Индекс общего бизнеса
условия. Преподобный эконом. Стат. 1, 111–205 (1919)
Временная последовательность

Петти, В.: Еще одно эссе по политической арифметике,
касающееся роста лондонского Сити.
Келли, Нью-Йорк (1682 г.) (2-е изд. 1963 г.)
Биостатистика

Петти, В.: Наблюдения за дублинскими векселями
Смертность, 1681 год и состояние этого города. Келли, Нью-Йорк
(1683 г.) (2-е изд. 1963 г.) Биостатистика

Петти, В.: Политическая арифметика. В: Уильям Петти, Экономические
сочинения. . . , т. 1. Келли, Нью-Йорк, стр. 233–313 (1676).
Эконометрика

Питман, EJG: Критерии значимости, которые можно применять
к образцам из любых популяций.
Дополнение, Дж. Рой. Стат. соц. 4, 119–130 (1937)
Повторная выборка

Плакетт, Р.Л., Барнард, Джорджия (ред.): Студент:
Статистическая биография Уильяма Сили Госсета (на основе
произведений Э. С. Пирсона). Кларендон, Оксфорд (1990)
Госсет, Уильям Сили Плакетт,

Р.Л.: Исследования по истории вероятностей
и статистика. В: Кендалл, М., Плакетт, Р.Л. (ред.)
Открытие метода наименьших квадратов. об. II. Гриффин,
Лондон (1977) Лежандр,
Адриен Мари, Регрессионный анализ Плакетт,

Р.Л.: Исследования по истории вероятностей
и статистика. VII. Принцип среднего арифметического. Биометрика
45, 130–135 (1958).
Среднее арифметическое

Playfair, W .: Коммерческий и политический атлас.
Playfair, Лондон (1786 г.) Графический
Представление

Poisson, SD: Recherches sur la probabilité des jugements en
matière criminelle et en en matière civile. В: Procédés des Règles
Generales du Calcul des Probabilités. Башелье, Imprimeur-Libraire pour
les Mathématiques, Париж (1837 г.)

Распределение Пуассона

Poisson, SD: Sur la probabilité des résultats moyens desObservations.
Connaissance des temps pour
l'an 1827, стр. 273–302 (1824 г.) Центральный предел
Теорема

Полиа, Г.: Ueber den zentralen Grenzwertsatz der
Wahrscheinlichkeitsrechnung und das
Momentproblem. Mathematische Zeitschrift 8, 171–181 (1920)
Центральная предельная теорема

Портер, ТМ: Расцвет статистического мышления,
1820–1900 гг. Издательство Принстонского университета, Принстон, Нью-
Джерси (1986) Коэффициент корреляции

Preece, DA: латинские квадраты, латинские кубы, латинские
прямоугольники и т. д. В: Коц, С., Джонсон, Н.Л. (ред.)
Энциклопедия статистических наук, том. 4. Уайли, Нью-Йорк (1983)
Latin Square Designs Ру, В .:

Описательная статистика. Экономика, Париж
(1987) Ковариация

Кенуй, М. Х.: Приблизительные тесты корреляции во временных рядах.
Дж. Рой. Стат. соц. сер. В 11, 68–84 Метод складного ножа
(1949)

Кенуй, М. Х.: Заметки о предвзятости в оценке.
Биометрика 43, 353–360 (1956). складной нож
Метод

Кетле, А.: Письма, адресованные Его Королевскому Высочеству Великому
Герцог Саксен-Кобург-Готский о теории вероятностей в применении
к моральным и политическим наукам. (Французский перевод
Даунса, Олинтуса Грегори). Чарльз и Эдвин Лейтон, Лондон (1849
г.)
Критерий независимости хи-квадрат Рюэгг,

А.: Вероятности и статистика. Прессы
Polytechniques Romandes, Лозанна, Швейцария (1985)
Эксперимент

Реньи, А.: Теория вероятностей. Северная Голландия,
Амстердам (1970) Теорема Бернулли Raktoe,

BL Hedayat, A., Federer, WT: Factorial
Дизайн. Уайли, Нью-Йорк (1981) Фактор

Рао, Ч.Р., Митра, С.К., Маттай, А., Рамамурти,
К.Г.: Формулы и таблицы для статистических работ.
Статистическая издательская компания, Калькутта, стр. 34–37
(1966). Биномиальный
Распределение

Rao, CR, Mitra, SK: обобщенные инверсии
Матрицы и их приложения. Уайли, Нью-Йорк (1971)
Обобщенный обратный

Рао, Ч.Р.: Передовые статистические методы в
Биометрические исследования. Уайли, Нью-Йорк (1952)
Анализ дисперсии Рао, Ч.Р.:

Заметка об обобщенной обратной матрице с приложениями к
задачам математической статистики. Дж. Рой. Стат. соц. сер.
Б 24, 152–158 (1962)
Обобщенный обратный

Machine Translated by Google

Рао, ЧР: Расчет обобщенного обратного
матрицы. I. Общая теория. Санкхья А29, 317–342 (1967)
Обобщенный обратный

Рао, Ч.Р.: Линейный статистический вывод и его
Приложения, 2-е изд. Уайли, Нью-Йорк (1973)
Теорема Гаусса-Маркова, Обобщенная линейная регрессия

Рашед, Р.: La
naissance de l'algèbre. В: Ноэль, Э. (ред.) Le Matin des
Mathématiciens. Белин-Радио Франция, Париж (1985)
Алгоритм, Арифметический
Треугольник

Рид, К.: Нейман — из жизни. Спрингер, Берлин Гейдельберг,
Нью-Йорк (1982) Проверка
гипотезы

Рипли, Б.Д., Венейблс, В.Н. (2000) S-программирование.
Спрингер, Берлин Гейдельберг Нью-Йорк
Статистическое программное обеспечение

Рипли, Б.Д.: Мысли о генераторах псевдослучайных чисел. Дж.
Вычисл. заявл. Мат. 31, 153–163 (1990)
Генерация случайных чисел

Роберт, СР: Le choix Bayesien. Springer, Paris Bayesian Statistics
(2006) Rose, D.: Панельные
исследования домохозяйств: обзор.
Инновации 8 (1), 7–24 (1995) Панель

Росс, С.М.: Введение в вероятностные модели, 8-е изд. Джон
Уайли, Нью-Йорк (2006 г.) Испытание
Бернулли, совместная функция распределения, предельная
функция плотности

Ротманн, Дж. К.: Эпидемиология. Введение.
Издательство Оксфордского университета (2002) Причина и
Эффект в эпидемиологии, эпидемиологии

Ротшильд, В., Логотетис, Н.: Распределения
вероятностей. Уайли, Нью-Йорк (1986)
Распределение вероятностей

Ройстон, Э.: Заметка об истории графического
представление данных. В: Пирсон, Э.С., Кендалл, М. (ред.)
Исследования по истории статистики и вероятностей,
том. И. Гриффин, Лондон (1970)
Графическое представление

Рубин, Д.Б.: Неитерационный алгоритм для оценки
отсутствующих значений по методу наименьших
квадратов в любом дисперсионном анализе. заявл.
Стат. 21, 136–141 (1972) Отсутствующие
данные Сапорта, Г.: Вероятность, анализ данных и
статистика. Технип, Париж (1990)
Анализ соответствия Savage, IR:

Библиография непараметрических
Статистика. Издательство Гарвардского университета, Кембридж,
Массачусетс (1962) Непараметрический тест

Сэвидж, Р.: Библиография непараметрической статистики и
смежных тем: введение. Варенье. Стат.
доц. 48, 844–849 (1953) Непараметрический
Статистика

Шеффе, Х.: Метод оценки всех контрастов в дисперсионном анализе.
Биометрика 40, 87–104 (1953).
Контраст

Шеффе, Х.: Дисперсионный анализ. Уайли, Нью-Йорк (1959)
Анализ дисперсии Schmid,

CF, Schmid, SE: Справочник по графическому представлению, 2-
е изд. Уайли, Нью-Йорк (1979)
Количественный график

Шмид, CF: Справочник по графическому представлению.
Рональд Пресс, Нью-Йорк (1954) Графический
Представление

Серл, С.Р.: Матричная алгебра полезна для статистики.
Уайли, Нью-Йорк (1982) Матрица

Себер, А.Ф., Уилт, С.К.: Нелинейная регрессия. Wiley, Нелинейная
Нью-Йорк (2003) регрессия Seber, GAF (1977)

Линейный регрессионный анализ.
Wiley, Нью-Йорк, простая линейная регрессия, взвешенный
метод наименьших квадратов

Сенн, Ст.: Перекрестные испытания в клинических исследованиях.
Уайли (1993) Уход

Шао, Дж., Ту, Д.: Складной нож и бутстрап.
Спрингер, Берлин Гейдельберг, Нью-Йорк (1995)
Bootstrap

Sheffé, H.: Дисперсионный анализ. C'n'H (1959)
Однофакторный дисперсионный анализ

Шеппард, В.Ф.: Новые таблицы интеграла вероятности.
Биометрика 2, 174–190 (1903).
Нормальная таблица

Шеппард, В.Ф.: О расчете наиболее
вероятные значения постоянных частот для данных,
расположенных по равноотстоящим делениям шкалы.
Лондонская математика. соц. проц. 29, 353–380
(1898) Момент

Шеппард, В.Ф.: Таблица отклонений от нормальной кривой.
Биометрика 5, 404–406 (1907). Нормальная таблица

Шейнин О.Б.: К истории некоторых статистических законов
распределения. В: Кендалл, М., Плакетт, Р.Л. (ред.)
Исследования по истории статистики и вероятностей,
том. II. Гриффин, Лондон (1977)
Распределение хи-квадрат

Шейнин О.Б.: Теория ошибок П.С. Лапласа,
Архив истории точных наук 17, 1–61.
(1977) Теорема де Муавра – Лапласа.

Machine Translated by Google

Шейнин О.Б.: Работа Лапласа П.С. о вероятности,
Архив истории точных наук 16, 137–187.
(1976) Теорема де Муавра – Лапласа.

Симпсон, Т.: Попытка показать преимущество, возникающее
при использовании среднего значения ряда
наблюдений в практической астрономии. В:
Разные трактаты по некоторым любопытным и очень
интересным предметам механики,
физической астрономии и спекулятивной
математики. Норс, Лондон (1757 г.). стр. 64–75
Среднее арифметическое

Симпсон, Т.: Письмо достопочтенному Джорджу графу Маклсфилду,
президенту Королевского общества, о преимуществе
использования среднего значения ряда наблюдений в
практической астрономии.
Филос. Транс. Рой. соц. Лонд. 49, 82–93 (1755)
Среднее арифметическое, ошибка

Симпсон, Т .: Разные трактаты по некоторым любопытным и очень
интересным предметам механики, физической
астрономии и спекулятивной математики.
Норс, Лондон (1757 г.) Ошибка

Sittig, J.: Экономичный выбор системы отбора проб при приемочном
отборе. Бык. Междунар. Стат. Инст. 33(V), 51–84 (1952) Размер
выборки

Смирнов Н.В.: Оценка отклонения между
эмпирические функции распределения в двух независимых
выборках. (Русский). Бык. Московский ун-т. 2 (2), 3–16 (6.1, 6.2)
(1939) Критерий Колмогорова-Смирнова,
непараметрический критерий.

Смирнов Н.В.: Таблица для оценки согласия эмпирических
распределений. Анна. Мат. Стат. 19, 279–281 (6.1) (1948)
Колмогоров–Смирнов
Тест

Смит, CAB: Статистика генетики человека. В:
Коц, С., Джонсон, Н.Л. (ред.) Энциклопедия статистических
наук, том. 3. Уайли, Нью-Йорк (1983)
Генетическая статистика

Снит, РНА, Сокаль, Р.Р.: Численная таксономия: принципы и
практика числовой классификации (серия книг по
биологии).
WH Freeman, Сан-Франциско, Калифорния (1973)
Дендрограмма

Snedecor, GW: Расчет и интерпретация
Дисперсионный и ковариационный анализ. Коллегиальный,
Эймс, Айова (1934) Распределение Фишера

Соболь И.М. Метод Монте-Карло. Мир
Издательство, Москва (1975) Монте-Карло
Метод

Сокаль, Р. Р., Снит, П. Х.: Принципы числовой таксономии.
Фриман, Сан-Франциско (1963)
Дендрограмма

Спирмен, К.: Доказательство и измерение
Ассоциация между двумя вещами. Являюсь. Дж. Психол.
15, 72–101 (1904) Коэффициент ранговой корреляции Спирмена.

Статистические науки: Статистический анализ в S-Plus,
Версия 3.1. Сиэтл: StatSci, подразделение MathSoft.
(1993) Статистическое программное обеспечение

Staudte, RG, Sheater, SJ: Надежная оценка и сходимость Stein, C.:
Тестирование. Уайли, Нью-Йорк (1990) Двухвыборочный
тест для линейной гипотезы, мощность которой не зависит от
дисперсии. Анна.
Мат. Стат. 16, 243–258 (1945) Размер образца

Стивенс, Массачусетс: статистика EDF для соответствия и некоторые
сравнения. Варенье. Стат. доц. 69, 730–737 (1974)
Тест Андерсона-Дарлинга

Стиглер, С.: Отчет Фрэнсиса Гальтона об изобретении корреляции.
Стат. науч. 4, 73–79 (1989)
Коэффициент корреляции

Стиглер, С.: Саймон Ньюкомб, Перси Дэниелл и история надежных
оценок 1885–1920 гг. Варенье.
Стат. доц. 68, 872–879 (1973) Выброс

Стиглер, С.: Закон эпонимии Стиглера, Труды Нью-Йоркской
академии наук, 2-я серия 39, 147–157 (1980) Нормальное
распределение Стиглер, С.: История статистики,
измерение неопределенности до 1900 года. Белкнап , Лондон (1986)

Семья Бернулли, комбинаторный анализ,
официальная статистика, вероятность, регрессионный
анализ, статистика Стюарт, А., Орд, Дж. К.:
продвинутая теория статистики Кендалла. Том. I. Теория
распределения. Уайли, Нью-Йорк (1943)
Мера центральной тенденции

Сукхатме, П.В.: Теория выборки обследований с
Приложения. Издательство государственного университета Айовы,
Эймс, Айова (1954) Выборка квот Takács,
L.: Комбинаторика. В: Коц, С.,
Джонсон, Н.Л. (ред.) Энциклопедия статистических наук,
том. 2. Уайли, Нью-Йорк (1982)
Комбинационный анализ

Тасси, доктор философии: De l'exhaustif au partiel: un peu d'histoire
sur le développement des sondages. Оценка и зондаж, несколько
вкладов в историю статистики. Экономика, Париж (1988)
Выборка

Чебычева, П.Л. (1890–1891). Sur deux théorèmes relatifs aux
probabilités. Акта Математика. 14, 305–315
Центральная предельная теорема

Machine Translated by Google

Чебычев, П.Л.: Des valeurs moyennes. Дж. Матем.
Чистый Appl. 12 (2), 177–184 (1867). Publié simultanément en
Russe, dans Mat. Сборник 2(2), 1–9 Закон больших чисел Тененхаус,
М.: La régression PLS. Теория и практика.
Технип, Париж (1998)
Частичный
Регрессия наименьших квадратов

Тиле, Теннесси: Теория наблюдений. С. и Э.
Лейтон, Лондон; Анна. Мат. Стат. 2, 165–308 (1903) Метод
моментов Торндайк, Р.Л.: Кто принадлежит к
семье?
Психометрика 18, 267–276 (1953).
Классификация, анализ данных Тодхантер,
И.: История математической теории вероятностей. Издательство Челси,
Нью-Йорк (1949)
Вероятность

Томассон Р., Даудин Дж. Дж., Данзарт М.,
Массон, Дж. П.: Дискриминация и классификация.
Массон, Париж (1988) Кластерный анализ

Tryon, RC: Кластерный анализ. Макгроу-Хилл, Нью-Йорк
Йорк (1939) Классификация

Тьюки, Дж. В.: Предвзятость и уверенность в не очень больших
выборках. Анна. Мат. Стат. 29, 614 (1958)
Метод складного ножа

Тьюки, Дж. В.: Сравнение индивидуальных средств в
дисперсионный анализ. Биометрия 5, 99–114 (1949).
Контраст

Tukey, JW: Объяснительный анализ данных, ограниченный
предварительное издание. Аддисон-Уэсли, Рединг, Массачусетс
(1970–1971) Анализ данных,
диаграмма «стебель-листья» Тьюки,
Дж. В.: Исследовательский анализ данных.
Аддисон-Уэсли, Рединг, Массачусетс (1977) Данные
Анализ, исследовательский анализ данных Тьюки, Дж.
В.: Быстрые и грязные методы в статистике.
Часть II: Простой анализ стандартных проектов.
Документы конференции по контролю качества 1951.
Американское общество контроля качества, Нью-Йорк, стр. 189–197
(1951). Контраст

Tukey, JW: Некоторые графические и полуграфические
дисплеи. В: Бэнкрофт, Т.А. (ред.) Статистические документы в честь
Джорджа У. Снедекора. Издательство государственного
университета Айовы, Эймс, Айова, стр. 293–316 (1972).
Блочная диаграмма, Исследовательский анализ данных, Шляпная
матрица

Twisk, JWR: Прикладной продольный анализ данных для эпидемиологии.
Практическое руководство. Издательство Кембриджского
университета, Кембридж (2003 г.)
Продольные данные

Статистическая комиссия Организации Объединенных Наций Основы
Принципы официальной статистики Официальный
Статистика

Венейблс, В. Н., Рипли, Б. Д.: Современная прикладная статистика
с S-PLUS. Спрингер, Берлин Гейдельберг, Нью-Йорк (1994)
Статистическое
программное обеспечение

Фон Нейман, Дж.: Различные методы, используемые в
Связь со случайными цифрами. Симпозиум Национального бюро
стандартов, NBS, Серия прикладной математики 12, стр.
36–38, Национальное бюро стандартов, Вашингтон, округ
Колумбия (1951)
Метод Монте-Карло

Фон Нейман, Дж.: Über ein ökonomisches
Gleichungssystem und eine Verallgemeinerung des Brouwerschen
Fixpunktsatzes. Эргебн. математика.
Коллоку. Вена 8, 73–83 (1937) Линейный
Программирование

Вальд, А.: Асимптотически самые мощные тесты
Статистические гипотезы. Анна. Мат. Стат. 12, 1–19 (1941)
Тест отношения правдоподобия

Уолд, А.: Некоторые примеры асимптотически наиболее мощных тестов.
Анна. Мат. Стат. 12, 396–408 (1941).
Тест отношения правдоподобия

Вегман, Э.Дж., Хейс, А.Р.: Статистическое программное обеспечение.
Энцикл. Стат. науч. 8, 667–674 (1988) Статистическое
программное обеспечение

Вайзеберг, С.: Прикладная линейная регрессия. Уайли, Новая множественная
Йорк (2006) линейная регрессия. Уайт, К.:
Использование рангов в тесте значимости.
для сравнения двух методов лечения. Биометрия 8, 33–41 (1952) Тест
Манна – Уитни.

Уайт, Ланселот Лоу: Роджер Джозеф Боскович,
Исследования его жизни и творчества на 250-м
Годовщина его Рождения. Аллен и Анвин, Лондон (1961)
Боскович, Роджер Дж.

Уилкоксон, Ф.: Индивидуальные сравнения методами ранжирования.
Биометрия 1, 80–83 (1945).
критерий Манна-Уитни, непараметрический критерий,
Таблица с подписью Уилкоксона, подпись Уилкоксона
тест, тест Уилкоксона

Уилкоксон, Ф.: Таблицы вероятностей для отдельных
сравнения методом ранжирования. Биометрия 3, 119–122 (1947).
Таблица Уилкоксона

Уилкоксон, Ф.: Некоторые быстрые приблизительные статистические
процедуры. Американский цианамид, Стэмфордские
исследовательские лаборатории, Стэмфорд, Коннектикут (1957 г.)
Непараметрический критерий, знаковый критерий Уилкоксона

Machine Translated by Google

Вильдт, А.Р., Ахтола, О.: Анализ ковариации (Серия статей Университета Сейдж о количественных приложениях в социальных науках, статья 12). Мудрец, Таузенд-Оукс, Калифорния (1978)

Ковариационный анализ Уилк, М.Б., Ганадесикан, Р.: Методы построения вероятностных графиков для анализа данных. Биометрика 55, 1–17 (1968) График QQ (график от квантиля к квантилю)

Уилкс, С.С.: Математическая статистика. Уайли, Нью Йорк (1962) Тест отношения правдоподобия

Уолд, Х.О. (ред.): Библиография по временным рядам и случайным процессам. Оливер и Бойд, Эдинбург (1965) Циклические колебания

Wonnacott, RJ, Wonnacott, TH: Эконометрика. Уайли, Нью-Йорк (1970) Эконометрика

Йейтс, Ф.: Дизайн и анализ факториала. эксперименты, техн. комм. 35, Имперское бюро почвоведения, Харпенден (1937 г.) Дробный Факторный дизайн

Йейтс, Ф.: Сложные эксперименты. Доп. Дж. Рой. Стат. соц. 2, 181–247 (1935) Двусторонний дисперсионный анализ.

Йейтс, Ф.: Методы выборки для переписей и обследований. Гриффин, Лондон (1949)

Стратифицированная выборка Йейтс, Ф.: Анализ повторных экспериментов, когда результаты поля неполные. Империя Дж. Эксп. Сельскохозяйственный. 1, 129–142 (1933) Недостающие данные

Йейтс, Ф.: Дизайн и анализ факториала. эксперименты. Техническое сообщение Бюро почв Содружества 35, Сельскохозяйственные бюро Содружества, Фарнхэм Роял (1937) Двусторонний анализ дисперсии, Алгоритм Йейтса

Юшкевич, А. П.: Арабские математические науки. (VIIIème-XVème веков). Частичный перевод Казенав, М., Жауиш, К. Врин, Париж (1976) Арифметический треугольник

Юл, Г.У. (1926) Почему мы иногда получаем бессмысленные корреляции между временными рядами? Исследование выборки и природы временных рядов. Дж. Рой. Стат. соц. (2) 89, 1–64 Коэффициент корреляции, Частичная корреляция Юле, ГУ: Об ассоциации признаков в статистика: с иллюстрацией из материала общество детства. Филос. Транс. Рой. соц. Лонд. сер. А 194, 257–319 (1900) Критерий независимости Хи-квадрат Юл Г.Ю.: К теории корреляции. Дж. Рой. Стат. соц. 60, 812–854 (1897 г.) Коэффициент корреляции

Юл, Г.В., Кендалл, М.Г.: Введение в теорию статистики, 14-е изд. Гриффин, Лондон (1968) Коэффициент Юла и Кендалла

Зубин, Дж.: Метод измерения сходства. Дж. Ненормальный. Социальная психология. 33, 508–516 (1938) Классификация, анализ данных