ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
Национальный исследовательский университет
ресурсоэффективных технологий
«ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Ю.И. Галанов

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Издание второе, дополненное
Рекомендовано в качестве учебного пособия
Редакционно-издательским советом
Томского политехнического университета

Издательство ТПУ

Томск 2010

УДК 519.2, 681.3)
ББК 517.8
Б24

Галанов Ю.И.

Б24 Математическая статистика: Учебное пособие. — Томск: Изд–во ТПУ, 2010. — 80 с.

В пособии рассмотрены основные статистические процедуры для представления и предварительной обработки статистических данных, методы точечного и интервального оценивания неизвестных параметров распределений, принципы проверки статистических гипотез. В разделе «Исследование зависимостей» рассмотрены принципы проведения однофакторного, корреляционного и регрессионного анализа, условия их корректного применения для числовых и нечисловых данных. Пособие предназначено для студентов 2-го курса, изучающих раздел математики «Математическая статистика», составлено в соответствии с ГОС 3-го поколения для подготовки бакалавров. Электронная версия пособия может быть использована в системе дистанционного обучения.

УДК 519.21(0.75,8)
ББК 22 171 Я73

Рекомендовано к печати Редакционно-издательским советом
Томского политехнического университета

Рецензенты

Кандидат физико-математических наук, доцент СГТИ
И.Л. Фаустова

Кандидат технических наук, доцент ТГУ
И.Г. Устинова

Общие сведения

Теория производит тем большее впечатление, чем проще её предпосылки, чем разнообразнее предметы, которые она связывает, и чем шире область её приложения

Альберт Эйнштейн

Задачи математической статистики

Математическая статистика — это прикладная математическая дисциплина, базирующаяся на понятиях и методах *теории вероятностей*, имеющая, однако, свои задачи и методы.

Пусть исследуются исходы некоторого опыта, причем вероятности исходов известны. Задача теории вероятностей состоит в разработке методов нахождения вероятностей различных сложных событий, исходя из известных вероятностей более простых событий.

Теория вероятностей, другими словами, занимается разработкой и исследованием *вероятностных моделей* случайных экспериментов.

На практике вероятности элементарных событий (или законы распределения случайных величин) редко бывают известны. Часто известно лишь то, что опыт можно описать в рамках какой-либо *вероятностно-статистической модели*, имеющей некоторую неопределенность в задании вероятности P событий или закона распределения случайных величин.

Задача математической статистики состоит в том, чтобы уменьшить эту неопределенность (восстановить закон распределения исследуемой случайной величины), используя информацию, полученную из эксперимента (*статистические данные*).

В определенном смысле, математическая статистика решает задачи, *обратные теории вероятностей*: она уточняет структуру статистических моделей по результатам проводимых наблюдений.

Математическая статистика является также наукой о *статистических выводах*: зачастую на основании статистических данных нам приходится делать выбор одного из нескольких, противоречащих друг другу, предположений (*гипотез*) относительно законов распределения случайных величин или о значениях параметров распределений.

В силу своего прикладного характера, математическая статистика занимается также *разработкой методов получения, описания и обработки опытных данных для изучения закономерностей случайных массовых явлений.*

Особенность идей и методов математической статистики — универсальность, возможность использования в различных приложениях.

Рассмотрим некоторые конкретные задачи, решаемые математической статистикой.

- Оценка на основании измерений неизвестной функции распределения.

  Дано: множество значений случайной величины

  $$X = x_i,\ i\ =\ 1,\ \ldots,\ n.$$

  Найти функцию распределения случайной величины X.

- Оценка неизвестных параметров распределения.

  Дано: случайная величина X имеет функцию распределения вида

  $$F(x, \vartheta_1, \vartheta_2, \ldots, \vartheta_n),$$

  где ϑ_1, ϑ_2, $\ldots$, ϑ_n — неизвестные параметры.

  Найти оценки этих параметров.

- Проверка статистических гипотез.

  Например, гипотеза о виде распределения:

  Дано: Предполагаем, что функция распределения случайной величины есть $F(x)$. Имеем данные

  $$X:\ x_{1,}x_{2,}\ldots, x_N.$$

  Спрашивается: совместимы ли значения X с гипотезой о том, что случайная величина имеет распределение $F(x)$?

Глава 1.

Основные понятия математической статистики

1.1. Первичные данные и их представление

Пусть исследуется некоторая совокупность объектов, каждому из которых ставится в соответствие некоторая числовая функция — случайная величина X, распределенная по некоторому неизвестному закону $L(\xi)$. Множество (конечное или бесконечное) всех объектов (всех значений случайной величины X) называют *генеральной совокупностью*.

На практике мы имеем дело с конечным набором данных, полученным в результате проведения n измерений или наблюдений. Эту конечную совокупность экспериментальных данных

$$x : x_1, x_2, \ldots, x_n \tag{1.1.}$$

называют *выборкой* объема n из генеральной совокупности.

Выборка является первичной формой представления экспериментального материала.

Каждой реализации $x: \ x_1, x_2, \ldots, x_n$ можно поставить в соответствие упорядоченную последовательность [1]

$$x_{(1)} \leqslant x_{(2)} \leqslant \ \ldots \ \leqslant x_{(k)}, \ k \leqslant n \tag{1.2.}$$

которую называют *вариационным рядом* выборки.

$x_{(i)}$ — порядковые *статистики*, предполагается, что все они *различны*. [2]

$x_{(1)}$, $x_{(k)}$ — экстремальные значения выборки.

Если в выборке есть повторяющиеся значения ($k < n$), то выборка представляется в виде *статистического ряда*– это таблица, в которой указаны все различные значения (варианты) вариационного ряда и числа, показывающие их количество в выборке.

[1] То есть мы располагаем значения **x** выборки в порядке возрастания.

[2] Статистикой называют любую функцию от выборки не содержащую неизвестных параметров.

x_i	x_1	x_2	$\dots$	x_k
m_i	m_1	m_2	$\dots$	m_k

Таблица 1.1. Статистический ряд

Представление выборки в виде статистического ряда естественно для дискретных распределения генеральной совокупности. Если генеральная совокупность имеет непрерывный закон распределения, то представление выборок в виде статистического ряда обусловлено двумя причинами:

- округлением результатов – ограничением числа верных значащих цифр в представлении результатов измерений, в результате чего в выборке неизбежно появляются повторяющиеся значения;
- большим объемом выборок, что вызывает необходимость предварительной группировки данных.

При группировке данных область значений случайной величины (в выборке) разбивается на k непересекающихся интервалов, необязательно равной длины, подсчитывается число элементов выборки, попавших в каждый интервал. В качестве значения, представляющего каждый интервал берется либо (чаще всего) середина интервала, либо среднее арифметическое значений точек, принадлежащих данному интервалу. Количество точек, попавших в интервал служит *весом* представительской точки в полученном таким образом *статистическом ряде*. В таблице группированных данных часто указываются не сами представительные точки, а границы соответствующих интервалов.

1.2. Математическая модель выборки. Эмпирическая функция распределения

Если мы повторим еще раз серию из n экспериментов, то получим новый набор чисел $x_1^{\langle 2\rangle}, x_2^{\langle 2\rangle}, \dots, x_n^{\langle 2\rangle}$. Повторяя раз за разом эти серии экспериментов, мы всякий раз будем получать новые наборы чисел $\{x_i^{\langle k\rangle}\}$, т.е. «результат i-го измерения в $k-$й серии» рассматривается как реализация случайной величины с тем же законом распределения, что и исходная случайная величина.

Таким образом, выборку (до опыта!) можно интерпретировать как *систему n независимых, одинаково распределенных случайных величин.*

Плотность распределения системы случайных величин для непрерывного распределения имеет вид:

$$\begin{aligned} L(X,\Theta) =& f(X_1,\Theta)\cdot f(X_2,\Theta)\cdot\ldots\cdot f(X_n,\Theta) = \\ =& \prod_{i=1}^{n} f(X_i,\Theta). \end{aligned} \tag{1.3.}$$

Здесь Θ – вектор параметров распределения.

После опыта выборка рассматривается как *реализация* либо одной случайной величины, либо – как *реализацию* случайного вектора $X = X_1, X_2, \ldots, X_n$, где X_i— независимые, одинаково распределенные случайные величины. [3]

Определим для каждого действительного x случайную величину $\mu_n(x)$, равную числу элементов выборки, значения которых меньше x:

$$\mu_n(x) = \sum_{i=1}^{n} I(x_i < x), \tag{1.4.}$$

где $I(A)$ — индикатор события A [4] .

Положим

$$F_n(x) = \frac{\mu_n(x)}{n}. \tag{1.5.}$$

Функция (1.5.) называется *эмпирической* (опытной, статистической) *функцией распределения* (ЭФР), соответствующей выборке X.

По своему определению ЭФР — случайная величина, принимающая дискретные значения $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$.

Поскольку

$$P\left(F_n(x) = \frac{k}{n}\right) = P(\mu_n(x) = k),$$

то, как следует из определения $\mu_n(x)$, она подчиняется биномиальному распределению с параметром

$$p = \mathbf{P}(\xi < x) = F(x).$$

Итак, ЭФР (как и вариационный ряд) — некоторая сводная характеристика выборки. Для каждой реализации x выборки X функция $F_n(x)$ однозначно определена и обладает всеми свойствами функции распределения: изменяется от 0 до 1, не убывает и *непрерывна слева*. При этом она кусочно-постоянна и возрастает только в точках последовательности (1.5.). Если в вариационном ряду (1.5.) нет одинаковых значений, то

$$F_n(x) = \begin{cases} 0, & \text{при } x \leqslant x_{(1)}, \\ k/n & \text{при } x_{(k)} < x \leqslant x_{(k+1)},\ k = 1, \ldots, n-1, \\ 1 & \text{при } x > x_n \end{cases}$$

т. е. в этом случае величина всех скачков равна $1/n$ и типичный график функции $F_n(x)$ имеет вид, изображенный на рис. 1.1.

[3] Предполагается. что существует, по крайней мере гипотетически, возможность неограниченное число раз воспроизводить серии независимых испытаний.

[4] $I(A)$ равен единице, если событие A имеет место, и равен нулю в противном случае.

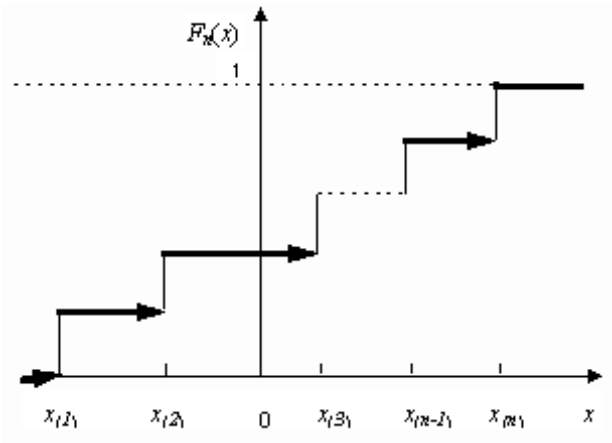


Рис. 1.1.

Согласно *теореме Бернулли*, ЭФР $F_n(x)$ при $n \to \infty$ сходится по вероятности к теоретической функции распределения $F(x)$.

Замечание. Процедуру получения выборки можно представить как выбор с возвращением из урны (*генеральной совокупности*) шаров (значений случайной величины). Чтобы выборка правильно представляла распределение генеральной совокупности, необходимо обеспечить случайность выборки, т.е., чтобы вероятность выбора любого элемента из генеральной совокупности была одинакова. В этом случае говорят. что выборка должна быть *репрезентативной*, т.е. представительной.

1.3. Гистограмма и полигон

Кроме эмпирической функции распределения существуют и другие способы наглядного представления статистических данных. Так, если наблюдаемая случайная величина принимает дискретные значения $a_1, a_2, \ldots,$ то более наглядное представление о законе распределения случайной величины ξ дадут частоты $\dfrac{\nu_r}{n}$, где ν_r — число элементов выборки $X = (X_1, \ldots, X_n)$, принявших значение a_r: $\nu_r = \sum\limits_{j=1}^{n} I\left(X_j = a_r\right)$. В этом случае, по теореме Бернулли, частоты $\dfrac{\nu_r}{n}$ сходятся по вероятности к вероятностям соответствующих событий $P(\xi) = a_r$.

Если случайная величина непрерывна, то данную методику приспосабливают для оценивания неизвестной плотности распределения следующим образом: область возможных значений ξ разбиваем точками на k непересекающихся интервалов; подсчитываем число точек m_r, попавших в каждый $r-$й интервал, вероятность попадания в некоторый $r-$й интервал оцениваем величиной $\dfrac{m_r}{n}$.

С другой стороны, эту же вероятность можно выразить через интеграл от плотности по данному интервалу ε_r:

$$\frac{m_r}{n} \approx \int\limits_{\varepsilon_r} f(x)\,dx \approx |\varepsilon_r| \cdot f(x_r),$$

где x_r — некоторая внутренняя точка интервала, [5] в качестве которой можно взять середину интервала.

Отсюда мы получаем оценку плотности распределения:

$$f_n(x_r) = \frac{m_r}{n \cdot |\varepsilon_r|}. \qquad (1.6.)$$

График кусочно-постоянной функции ((1.6.)) называется *гистограммой* (см. рис. 1.2). Если соединим точки $\mathrm{M}(x_r, f_n(x_r))$ отрезками прямых линий, то получим кусочно-линейный график, также являющийся статистическим аналогом плотности распределения, который называется *полигон частот*.

Более плавную кривую можно получить, используя *локальную аппроксимацию кривой полиномами* или *сглаживание результатов*.

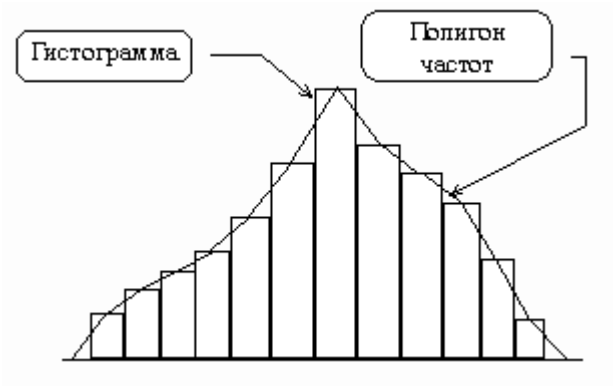


Рис. 1.2. Гистограмма и полигон частот

Замечание. При группировке данных приходится учитывать два взаимно исключающих обстоятельства. С одной стороны, число интервалов группировки должно быть достаточно велико, чтобы детально описать поведение плотности распределения, но, с другой стороны, число точек, попадающих в интервал, также должно быть достаточным, чтобы надежно представлять данный интервал. Если интервалов будет много. то некоторые из них могут оказаться пустыми, а плотность распределения – изрезанной, многолепестковой.

В литературе имеется много рекомендаций по выбору оптимального числа интервалов группировки. Приведем только две формулы:

[5] Здесь мы применили теорему о среднем.

Формула Старджеса:

$$K = 1 + \lfloor \log_2 n \rfloor, \tag{1.7.}$$

где $\lfloor x \rfloor$ есть целая часть, не превосходящая x.

Формула Брукса и Каррузера: $K = \lfloor \sqrt{n} \rfloor$.

При малых выборках ($n \leq 20$) некоторые интервалы могут оказаться пустыми. В таком случае надо уменьшить их количество. Считается приемлемым выбирать число интервалов так, чтобы выполнялось $n_j \leq 5$ (оптимисты допускают $n_j \leq 3$) для каждого интервала.

Глава 2.

Оценка неизвестных параметров распределений

2.1. Точечное оценивание неизвестных параметров

На практике вид функции распределения часто бывает известен с точностью до неизвестных параметров $F_\xi(x) = F(x, \theta)$. В этом случае определение функции распределения сводится к определению неизвестных параметров θ.

Например, если случайная величина ξ — результат прямых измерений некоторой физической величины a, то, при отсутствии систематических ошибок, распределение вероятностей случайной величины ξ будет описываться нормальным законом распределения с двумя параметрами: математическим ожиданием $M[\xi] = a$ и дисперсией $D[\xi] = \sigma^2$, которые нужно оценить по имеющейся выборке.

Определение 2.1. *Статистикой называют любую функцию от выборки, не содержащую неизвестных параметров.*

Всякая оценка неизвестного параметра по выборке (*статистика*) — является функцией выборочных значений: $\hat{\theta} = \hat{\theta}\,(x)$, следовательно, есть случайная величина со своим законом распределения.

Один и тот же параметр можно оценивать с помощью различных статистик. Поэтому возникает вопрос о выборе наилучшей в некотором смысле оценочной функции $\hat{\theta}$.

2.1.1. Требования к оценкам

Принцип наименьших квадратов.

О качестве оценок неизвестных параметров будем судить по тому, насколько хорошо выполняется приближенное равенство:

$$\theta \approx \widehat{\theta}.$$

Рассмотрим ошибку Δ, возникающей при замене неизвестного точного значения параметра θ его приближенным значением $\widehat{\theta}$:

$$\Delta = \theta - \widehat{\theta} \tag{2.1.}$$

Ввиду случайности оценки, ошибка Δ также является случайной величиной со своим законом распределения. Найдем числовые характеристики ошибки:

Математическое ожидание.

$$M[\Delta] = M[\theta - \widehat{\theta}] = \theta - M[\widehat{\theta}] = b \tag{2.2.}$$

Дисперсия.

$$D[\Delta] = D[\theta - \widehat{\theta}] = D[\widehat{\theta}] = M[(\Delta)^2] - b^2 \tag{2.3.}$$

Величина b (2.2.) – называется *смещением* оценки.[1] Из (2.3.) найдем среднее квадратичное отклонение, которое примем за меру близости оценки и оцениваемого параметра:

$$\delta^2 = M[(\Delta)^2] = D[\widehat{\theta}] + b^2 \tag{2.4.}$$

Наилучшей в своем классе оценок будем считать такую оценку, которая имеет наименьшее среднее квадратичное отклонение δ^2 (2.4.).

Так как δ^2 складывается из двух частей: квадрата смещения и дисперсии оценки, то наилучшими оценками мы будем считать оценки с нулевым смещением и минимальной дисперсией.

Определение 2.2. ***Несмещенными*** *называют оценки с нулевым смещением, т.е. математическое ожидание несмещенной оценки равно оцениваемому параметру.*

Определение 2.3. *Если несмещенная оценка обладает минимальной в своем классе оценок дисперсией, то она называется* ***эффективной.***

Еще один подход к анализу качества оценок связан с поведением оценок с ростом объема выборки: чем больше объем выборки, чем точнее должна быть оценка.

Определение 2.4. *Оценка параметра называется* состоятельной *если она при $n \to \infty$ сходится по вероятности к оцениваемому параметру*

Если $\hat{\theta}$ — неизвестная числовая характеристика распределения, то оценочную функцию можно строить, например, следующим образом. Строим по имеющейся выборке статистический аналог нужной числовой характеристики и принимаем его за оценку неизвестного параметра.

[1] В теории ошибок она квалифицируется как *систематическая ошибка.* Величина $\Delta - b$ называется случайной ошибкой.

Обоснованием данного метода служит *асимптотическое* поведение статистических аналогов параметров распределений – сходимость по вероятности к теоретическим характеристикам.

При этом учитываем, что моделью выборки является дискретная случайная величина, для которой $p_i = \dfrac{1}{n}$.

Оценка математического ожидания

Оценкой математического ожидания является *выборочное среднее*:

$$\hat{m}_x = \bar{x} = \sum_{i=1}^{n} p_i x_i = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.5.}$$

Оценка дисперсии

Оценкой дисперсии будет *выборочная дисперсия*:

$$\hat{D}_x = S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2. \tag{2.6.}$$

Аналогично рассчитываются оценки и для других *числовых характеристик* распределения.

Рассмотренный выше способ оценки (с помощью *статистических аналогов*) пригоден не для всех параметрических функций распределения. Кроме того, он не всегда приводит к наилучшим оценкам. Возникает вопрос — какую оценочную функцию (*статистику*) считать наилучшей или «хорошей»?

2.1.2. Требования к статистикам

Несмещенность. *Оценка называется несмещенной, если при любом* θ $M\left[\widehat{\theta}\right] = \theta$, *т. е. нет систематической ошибки.*

Эффективность *Несмещенные оценки различаются своими дисперсиями. Оценка с наименьшей для оценок данного класса оценок дисперсией называется эффективной.*

Состоятельность *Оценка параметра называется состоятельной если она при* $n \to \infty$ *сходится по вероятности к оцениваемому параметру*:

$$\hat{\theta}\left(\mathbf{x}\right) \xrightarrow{p} \theta.$$

Пример 1.

Покажем, что выборочное среднее (1.8) является несмещенной и состоятельной оценкой математического ожидания.

$$M\left[\bar{X}\right] = \frac{1}{n}\sum_{i=1}^{n} M[X_i] = \frac{n \cdot m_x}{n} = m_x, \tag{2.7.}$$

так как $\{x_i\}$ — одинаково распределенные случайные величины.

Согласно закону *больших чисел* $\bar{X} \xrightarrow{p} M[\xi]$, т. е. оценка состоятельна. Доказывается, что в случае нормального распределения эта оценка к тому же и эффективна. Поскольку выборочное среднее — случайная величина, то она описывается некоторым законом распределения, который при $n \to \infty$ асимптотически приближается к нормальному с параметрами

$$\begin{aligned} m_{\bar{x}} &= m_x \\ D_{\bar{x}} &= D\left(\tfrac{1}{n}\sum_{i=1}^{n} x_i\right) = \tfrac{1}{n^2} n\sigma^2 = \tfrac{\sigma^2}{n}, \end{aligned} \tag{2.8.}$$

т. е. в пределе

$$L\left(\bar{X}\right) \xrightarrow{p} N\left(m_x, \frac{\sigma_x^2}{n}\right). \tag{2.9.}$$

Пример 2.

Рассмотрим оценку дисперсии (1.9) и покажем, что она является *смещенной* оценкой дисперсии. Преобразуем это выражение:

$$\begin{aligned} \hat{D}_x = S^2 &= \tfrac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \tfrac{1}{n}\sum_{i=1}^{n}\left(x_i - m_x - \bar{x} + m_x\right)^2 = \\ &= \tfrac{1}{n}\sum_{i=1}^{n}\left(x_i - m_x\right)^2 - \tfrac{2\cdot(\bar{x}-m_x)}{n}\sum_{i=1}^{n}\left(x_i - m_x\right) + \left(\bar{x} - m_x\right)^2 = \\ &= \tfrac{1}{n}\sum_{i=1}^{n}\left(x_i - m_x\right)^2 - \left(\bar{x} - m_x\right)^2. \end{aligned}$$

Теперь получим:

$$\begin{aligned} M\left[\hat{D}_x\right] &= \tfrac{1}{n}\sum_{i=1}^{n} M\left[\left(x_i - m_x\right)^2\right] - M\left[\left(\bar{x} - m_x\right)^2\right] = \\ &= \frac{1}{n}\cdot n \cdot \sigma^2 - D[\bar{x}] = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \cdot \frac{n-1}{n}. \end{aligned} \tag{2.10.}$$

Таким образом математическое ожидание оценки дисперсии (1.9) не равно оцениваемому параметру, т. е. оценка смещена. Доказывается, что эта оценка является состоятельной.

Из (1.13) легко можно получить несмещенную оценку дисперсии:

$$S^{*2} = \frac{n}{n-1} S^2 \tag{2.11.}$$

Асимптотическое поведение выборочных характеристик обосновывает их использование в качестве оценок неизвестных параметров, что дает хорошие результаты при оценке математического ожидания и дисперсии в случае выборок большого объема. Существуют и другие принципы построения оценок.

2.1.3. Метод моментов

Пусть $\theta = (\theta_1, \theta_2, \ldots, \theta_r)$ — вектор оцениваемых параметров распределения. Предположим, что у наблюдаемой случайной величины ξ существуют первые r моментов $\alpha_k = \mathrm{M}[\xi^k]$, $k = 1, \ldots, r$. Они являются функциями от неизвестных параметров θ: $\alpha_k = \alpha_k(\theta)$.

Суть *метода моментов* [2] состоит в том, что мы берем значения соответствующих выборочных моментов $a_k = A_{nk}(\mathrm{x})$ и приравниваем их теоретическим. В результате получаем систему уравнений относительно неизвестных параметров:

$\alpha_k(\theta) = a_k.,\ \ k = 1, \ldots, r.$ (1.(15)

Решая эти уравнения относительно $\theta_1, \theta_2, \ldots, \theta_r$, получаем значения оценок параметров по методу моментов. Обоснованием метода служит несмещенность и состоятельность выборочных моментов $A_{nk}(\mathrm{x})$.

Отметим, что полученные ранее оценки математического ожидания (1.8) и дисперсии (1.9) совпадают с оценками метода моментов.

Оценка математического ожидания

По определению, математическое ожидание — есть начальный момент первого порядка. Поэтому его оценкой по методу моментов является выборочное среднее (1.8).

Оценка дисперсии

По определению, дисперсия — есть центральный момент второго порядка. Поэтому его оценкой по методу моментов является выборочная дисперсия (1.9).

Достоинством метода моментов является относительная простота решения уравнений (1.15). Метод не применим, когда моменты нужного порядка не существуют. Кроме того, эти оценки часто неэффективны, поэтому обычно их используют в качестве первых приближений для нахождения последующих приближений с большей эффективностью.

2.1.4. Метод наибольшего правдоподобия

Пусть x: $x_1, \mathrm{x}_2, \ldots, x_n$ — реализация выборки из распределения $p(x, \theta)$, зависящего от параметра θ.

Рассмотрим совместную плотность распределения системы независимых случайных величин $\mathbf{X} = (X_1, \ldots, X_n)$ при *фиксированном* x как функцию параметра θ, называемую *функцией правдоподобия*:

[2] Метод моментов предложен К. Пирсоном в 1894 году.

$$L(x,\theta) = p(x_1,\theta)\cdot p(x_2,\theta)p(x_1,\theta)\cdots p(x_n,\theta). \qquad (2.12.)$$

Оценкой *наибольшего правдоподобия* называется оценка, которая обращает в максимум *функцию правдоподобия*:

$$L\left(\mathbf{x},\hat{\theta}\right) = \max_{\theta} L\left(\mathbf{x},\theta\right). \qquad (2.13.)$$

Если функция (1.17) дифференцируема, то оценку параметра можно найти, решив уравнение *правдоподобия*:

$$\frac{\partial \ln L\left(\mathbf{x},\theta\right)}{\partial\theta} = 0. \qquad (2.14.)$$

Оценки наибольшего правдоподобия являются состоятельными и эффективными, однако они могут оказаться смещенными. На практике, нахождение оценок методом наибольшего правдоподобия приводит к сложным системам уравнений.

Пример. Оценка значения измеряемой величины в случае неравноточных измерений.

Пусть имеем выборку: серию независимых измерений одной и той же величины, проведенных с различной точностью. Т. е. $\mathrm{M}[X_i] = m$, $\mathrm{D}[X_i] = \sigma_i^2$. Необходимо оценить неизвестное математическое ожидание m. Составим функцию правдоподобия для данного случая:

$$L\left(x,\theta\right) = \prod_i \frac{1}{\sqrt{2\pi}\cdot\sigma_i}\cdot e^{-\frac{1}{2}\sum\limits_i \frac{(x_i-m)^2}{\sigma_i^2}} \qquad (2.15.)$$

Приравнивая производную (1.19) по m нулю, получим:

$$\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2} - \hat{m}\sum_{i=1}^{n}\frac{1}{\sigma_i^2} = 0 \;\Rightarrow\; \hat{m} = \frac{\sum\limits_{i=1}^{n}\frac{x_i}{\sigma_i^2}}{\sum\limits_{i=1}^{n}\frac{1}{\sigma_i^2}} = \sum_{i=1}^{n} g_i x_i, \qquad (2.16.)$$

где $g_i = \frac{\frac{1}{\sigma_i^2}}{\sum\limits_{i=1}^{n}\frac{1}{\sigma_i^2}}$ — вес i-го измерения.

Смысл весового коэффициента состоит в том, что чем больше дисперсия i-го измерения, тем меньший вклад это измерение вносит в оценку измеряемой величины.

Покажем, что данная оценка несмещенная.

$$M[\hat{m}] = M\left[\sum_{i=1}^{n} g_i x_i\right] = \sum_{i=1}^{n} g_i M[x_i] = m\sum_{i=1}^{n} g_i = m\cdot 1 = m.$$

Докажем состоятельность оценки. Найдем дисперсию оценки:

$$D[\hat{m}] = \sum_{i=1}^{n} g_i^2 \sigma_i^2 = \frac{1}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_i^2}} \xrightarrow{n\to\infty} 0,$$

при условии, что $\sigma_i > 0$. Следовательно, оценка состоятельная. Легко доказать, что оценка (1.20) — эффективная в классе линейных оценок.

2.1.5. Оценка числовых характеристик при отклонении закона распределения от нормального

Экспериментальные результаты почти никогда не имеют "чисто" гауссовского распределения, в тоже время существует широкий класс распределений, очень близких к гауссовскому. [3] Для таких распределений характерна повышенная по сравнению с гауссовским вероятность больших отклонений от среднего, т.е. их распределение вероятности имеют утяжеленные "хвосты". Модель такого распределения может быть представлена в следующем виде:

$$W(x) = (1-\varepsilon) \cdot N(x, \mu_x, \sigma_x^2) + \varepsilon \cdot W(x, \mu_x, \theta), \tag{2.17.}$$

где ε – доля аномальных наблюдений.

Причины отклонения от нормальности

Основные причины «негауссовости» (не нормальности) измерений являются следующие.

- Внутренне присущие исследуемому явлению физические причины, формирующие сигналы с распределением, принципиально отличающимся от нормального.

  Например, шумы на выходе амплитудного детектора, при подаче на вход нормального шума распределены по закону Релея:

$$p(x) = \begin{cases} \frac{x}{\sigma^2} * \exp(-\frac{x^2}{2\sigma^2}), & x \geqslant 0 \\ 0, & x < 0 \end{cases}.$$

- Существование нерегулярных аномальных ошибок измерений.

[3] Даже когда мы, проверяя гипотезу о нормальности распределения, и принимаем нулевую гипотезу, вывод мы делаем о том, что экспериментальные данные *не противоречат* гипотезе. Эта проверка скорее всего может ответить на вопрос, какое распределение не *описывает* наблюдаемые величины, но не позволяет однозначно выбрать истинное распределение.

Последнее обстоятельство вообще делает невозможным использование параметрических оценок так как алгоритмы их получения основываются на конкретных свойствах законов распределений. Большинство оптимальных параметрических оценок резко теряют свои замечательные свойства — несмещенность и эффективность даже при незначительных отклонениях от стандартных условий.

Понятие о робастных оценках

Для повышения качества оценок в реальных условиях применяют так называемые *робастные* оценки.

Определение 1.1 Под робастностью понимается слабая чувствительность к отклонениям от стандартных условий и высокая эффективность для широкого класса распределений.

Оценки параметра сдвига

Выборочная *медиана*. Эта статистика более предпочтительна чем выборочное среднее при небольших засорениях распределений ($\omega \cong$ 0.1 - 0.9). выборочная медиана относится к классу взвешенных порядковых статистик, определяемых соотношением:

$$\bar{x}_n = \sum_{i=1}^{n} w_i x_i, \tag{2.18.}$$

где

w_i - вес i-го члена вариационного ряда.

Для выборочной медианы веса $\mathrm{w}_i = 0$ для всех членов, кроме одного среднего, когда n - нечетное и двух средних, когда n - четное:

$w_{n/2} = w_{n/2+1} = 1$, n - четное,

$w_{(n+1)/2} = 1$, n - нечетное..

Цензурированные оценки сдвига 1-й тип.

$$w_i = \begin{cases} w_0 > 0, & a \leqslant x_i \leqslant b \\ 0, & x_i < a,\ x_i > b \end{cases} \tag{2.19.}$$

w_0 – обратно пропорционально числу элементов вариационного ряда, попавших в заданный интервал значений $[a, b]$ – определяется после завершения эксперимента.

2-й тип.

$$w_i = \begin{cases} \frac{1}{k-m+1}, & (m \leqslant i \leqslant k) \\ 0, & (n-k < i < n) \end{cases} . \qquad (2.20.)$$

Частный случай - усеченное среднее, когда отбрасывается r наибольших и r наименьших членов вариационного ряда.

Есть и другие виды оценок. Как пример применения оценок данного типа - спортивное судейство.

Оценка параметров масштаба распределений Наряду со среднеквадратичным отклонением σ_x используют среднее абсолютное отклонение

$$\delta_x = \int\limits_{-\infty}^{+\infty} |x - \mu_x| w(z) dx, \qquad (2.21.)$$

которое связано с σ_x соотношением:

$$\sigma_x = \sqrt{\pi/2} \cdot \delta_x \qquad (2.22.)$$

Оценка

$$S_x = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{n} \cdot \sum_i |x_i - m_x| \qquad (2.23.)$$

является более эффективной, чем выборочное среднеквадратичное отклонение.

Оценка параметров связи при наличии аномальных отклонений.

2.2. Интервальное оценивание неизвестных параметров распределений

Мы оценивали неизвестные параметры одним числом, т. е. *одной точкой* из области возможных значений оцениваемого параметра. В ряде задач требуется найти не только числовое значение параметра, но и оценить его *точность* и *надежность*. Т. е. надо знать, какая ошибка появится при замене неизвестного параметра θ его оценкой $\hat{\theta}$ и какова вероятность того, что эти ошибки не выйдут за известные пределы. [4]

Понятие доверительного интервала

Точность и *надежность* оценки задаются так называемыми *доверительными интервалами* и *доверительными вероятностями.*

Интервал l_γ, содержащий, с вероятностью γ, точное значение оцениваемого параметра, называется доверительным интервалом.

Вероятность γ того, что истинное значение θ лежит в интервале l_γ *называется* доверительной вероятностью *(*коэффициентом доверия*) или* надежностью, *соответствующей данному доверительному интервалу.*

Доверительный интервал и доверительная вероятность (см. рис. 1.3) связаны соотношением:

$$P\left(\left|\theta-\hat{\theta}\right|<\varepsilon\right)=\gamma \tag{2.24.}$$

или

$$P\left(\hat{\theta}-\varepsilon<\theta<\hat{\theta}+\varepsilon\right)=\gamma.$$

Отсюда

$$l_\gamma=\left(T_1\left(x\right),T_2\left(x\right)\right)=\left(\hat{\theta}-\varepsilon,\,\hat{\theta}+\varepsilon\right). \tag{2.25.}$$

Интервал (1.30) называют *γ-доверительным интервалом параметра θ.* $T_1(x)$, $T_2(x)$ — нижняя и верхняя доверительные границы.

Таким образом, диапазон возможных ошибок при замене параметра θ его оценкой $\hat{\theta}$ будет равен $\pm\varepsilon$; большие ошибки появляются с малой вероятностью $\alpha=1-\gamma$.

[4] Любая точечная оценка параметра есть функция выборки, т. е. является случайной величиной и содержит ошибки, которые становятся очень большими при малых объемах выборки.

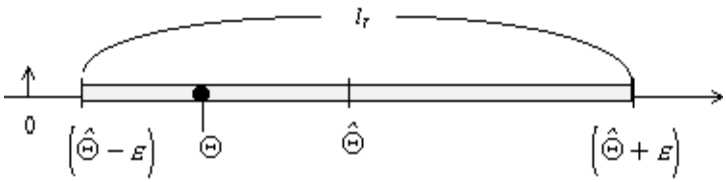


Рис. 2.1. К понятию доверительного интервала

2.2.1. Построение доверительных интервалов с помощью центральных статистик

Пусть имеем случайную величину, описываемую непрерывным распределением и существует функция от выборки (статистика) $\mathrm{G}(\mathbf{X};\theta)$, зависящая от θ такая, что: распределение $\mathrm{G}(\mathbf{X};\theta)$ не зависит от θ; при каждой реализации выборки x функция $\mathrm{G}(\mathbf{X};\theta)$ непрерывна и строго монотонна по θ (см. рис. 1.4).

Такую статистику называют *центральной статистикой* для θ.

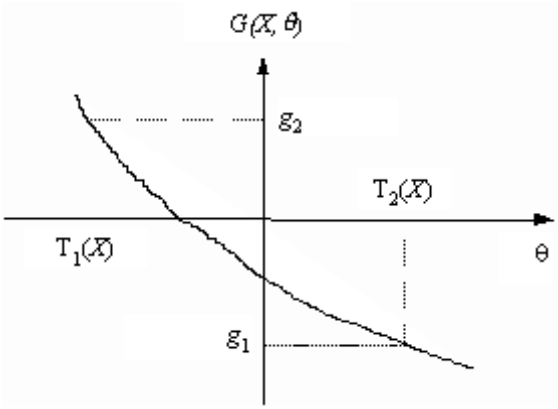


Рис. 2.2.

Пусть $f_G(g)$ — плотность распределения статистики G (напомним, что она не зависит от оцениваемого параметра!). Выберем величины $g_1 < g_2$ так, чтобы [5]

$$\mathbf{P}_0\left(g_1 < G\left(\mathbf{X};\vartheta\right) < g_2\right) = \int_{g_1}^{g_2} f_G\left(g\right) dg = \gamma. \qquad (2.26.)$$

Теперь мы можем найти соответствующие им числа $T_1 < T_2$. Тогда неравенства $g_1 < \mathrm{G}(\mathrm{x},\theta) < g_2$ будут эквивалентны неравенствам $T_1 < \theta < T_2$ (см. рис. 1.4). Уравнение (1.31) можно переписать в виде:

[5] Способ выбора этих величин уточним для конкретных случаев. Чаще всего используют *симметричные* интервалы.

$$\mathbf{P}_0\left(T_1\left(\mathbf{X}\right) < \theta < T_2\left(\mathbf{X}\right)\right) = \gamma, \qquad (2.27.)$$

при любом θ.

Построенный интервал $(T1; T2\ )$ является γ-доверительным интервалом для θ.

Мы рассмотрим ряд примеров построения доверительных интервалов для параметров нормальных моделей.

2.2.2. Распределение некоторых функций от выборки из нормального распределения

Особая трудность применения рассмотренной методики состоит в решении уравнений (1.31, 1.32). Для этого нам в первую очередь необходимо знать плотность распределения используемой статистики и значения *квантилей* этих распределений.

Квантили распределений

Квантиль — одна из числовых характеристик распределения вероятностей. Если функция распределения $F(x)$ случайной величины X непрерывна и строго монотонна, то для любого p, $0 < p < 1$,
квантиль *порядка* p *распределения случайной величины* X *определяется как корень* x_p *уравнения* $F(x_p) = p$ *или, иначе, как значение (при данном* p*) функции* $F^{--1}(p)$*, обратной к* $F(x)$*.*

Например, квантиль $x_{1/2}$ — есть *медиана* X (см. рис. 1.5).

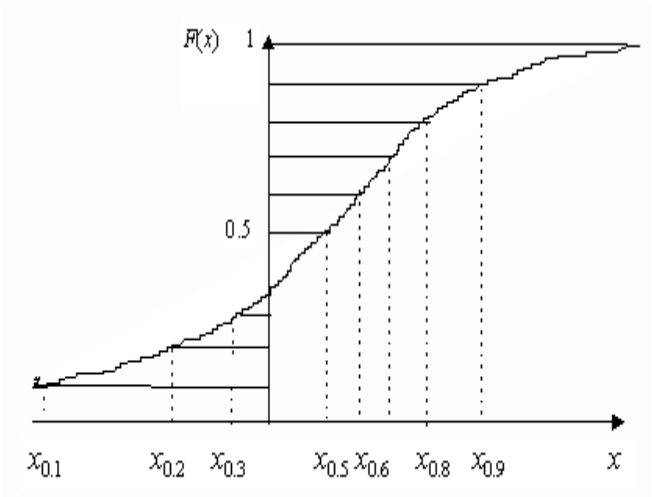


Рис. 2.3. Квантили распределения

Для выборочных распределений также вводят понятие квантилей. Для некоторых важных видов распределений существуют таблицы квантилей [6] . Ниже мы рассмотрим распределения некоторых *функций* от выборки из *стандартного* нормального распределения. Особенностью *распределений* этих функций является то, что они зависят не от неизвестных параметров, а лишь от *объема* выборки. поэтому данные функции выборки часто применяются в качестве статистик при доверительном оценивании параметров.

Стандартное распределение

Пусть случайная величина X подчиняется нормальному распределению с параметрами m, σ^2: $\mathrm{L}(X) = N(m,\ \sigma^2)$. Перейдем с помощью линейного преобразования к случайной величине

$$Y = \frac{X - m}{\sigma}, \tag{2.28.}$$

имеющей нормальное распределение с нулевым математическим ожиданием и единичной дисперсией: $\mathrm{L}(Y) = N(0,1)$. Такое распределение называется *стандартным*. Плотность стандартного распределения равна

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{2.29.}$$

Функция стандартного распределения называется *интеграл вероятности Гаусса*:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2} dt. \tag{2.30.}$$

Для данных функций составлены подробные таблицы.

Распределение хи-квадрат

Пусть имеем выборку $\xi = \{\xi_i,\ i = 1,2,\ \ldots,\ n\}$, где ξ_i — независимые случайные величины, имеющие *стандартное* распределенные. Введем функцию от выборки, называемую *хи-квадрат*:

$$\chi_n^2 = \sum_{i=1}^{n} \xi_i^2. \tag{2.31.}$$

[6] В среде MATHCAD, в версии 6.0 и выше, квантили важнейших распределений реализованы в виде функций.

Случайная величина (1.36) описывается *хи-квадрат распределением с n степенями свободы* [7] . Плотность хи-квадрат распределения обозначают через $k_n(x)$. Она имеет вид [8] :

$$k_n(x) = \frac{x^{n/2-1}}{2^{n/2}\Gamma\left(n/2\right)} e^{-x/2}, \ x > 0. \tag{2.32.}$$

Здесь $\Gamma(x) = \int\limits_0^\infty e^{-t} t^{x-1} dt$— гамма функция Эйлера. Первые два момента данного распределения равны: $\mathbf{M}[\chi^2] = n$, $\mathbf{D}[\chi^2] = 2n$ [9] .

Важным свойством распределения хи-квадрат является его *воспроизводимость* по параметру n, которое означает, что сумма независимых случайных величин, распределенных по закону хи-квадрат, распределена также по закону хи-квадрат с числом степеней свободы, равным сумме степеней свободы слагаемых.

Распределение Стьюдента (t распределение)

[10]

Распределением Стьюдента с n степенями свободы $S(n)$ называется распределение случайной величины

$$t = \frac{\xi}{\sqrt{\frac{\chi_n^2}{n}}}, \tag{2.33.}$$

где случайные величины ξ и χ_n^2*независимы* и при этом ξ имеет *стандартное* распределение. Плотность *t-распределения $s_n(x)$* имеет вид

$$s_n(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left((n+1)/2\right)}{\Gamma\left(n/2\right)} \frac{1}{\left(1 + x^2/n\right)^{(n+1)/2}}, \ -\infty < x < \infty. \tag{2.34.}$$

Распределение Фишера-Снедекора

Пусть случайные величины χ_n^2 и χ_m^2 независимы и

[7] Число степеней свободы — это число *линейно независимых* слагаемых в сумме (1.28).

[8] Здесь и далее, плотность распределения вычисляется по правилам вычисления плотности распределения *функции* от случайной величины с известным распределением.

[9] В соответствие с центральной предельной теоремой $L(\chi_n^2) \sim N(n, 2n)$ при $n \to \infty$.

[10] Стьюдент (Student) — псевдоним английского статистика В. Госсета, впервые использовавшего это распределение (1908).

$$F = \frac{\chi_n^2}{n} : \frac{\chi_m^2}{m} = \frac{m}{n}\frac{\chi_n^2}{\chi_m^2}. \qquad (2.35.)$$

Распределение случайной величины F называют *распределением Снедекора с n и m степенями свободы* и обозначают $S(n,\ m)$. Иногда это распределение называют *F-распределением* или *распределением дисперсионного отношения Фишера.*

Построение доверительного интервала для математического ожидания

Дисперсия известна.

Пусть по выборке $\mathbf{X} = (\ X_1,\ X_2,\ \ldots,\ X_n)$ из *нормального распределения* $N(\theta,\ \sigma^2)$требуется построить доверительный интервал для неизвестного среднего θ при известной дисперсии σ^2.

В данном случае центральной статистикой для θ будет

$$G\left(\mathbf{X};\theta\right) = \sqrt{n}\frac{\bar{X}-\theta}{\sigma}, \qquad (2.36.)$$

имеющая, как легко убедиться, стандартное распределение. Границы доверительного интервала определятся выражениями:

$$T_1\left(\mathbf{x}\right) = \bar{X} - \frac{\sigma}{\sqrt{n}}g_2;\ T_2\left(\mathbf{x}\right) = \bar{X} - \frac{\sigma}{\sqrt{n}}g_1. \qquad (2.37.)$$

Значения чисел g_1 и g_2 определены не однозначно. Известно только, что они должны удовлетворять соотношению $\Phi(g_1) - \Phi(g_2) = \gamma$. Доказывается, что доверительный интервал будет кратчайшим, если доверительные границы расположить *симметрично* относительно оцениваемого параметра, т. е. положить $g_1 = -\ g_2$. Учитывая, что $\Phi(-x) = 1 - \Phi(x)$, получаем

$$\Phi\left(g_2\right) = \frac{1+\gamma}{2},$$

т. е. g_2 есть $\frac{1+\gamma}{2}-$ квантиль стандартного распределения:

$$g_2 = c_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right).$$

И окончательно получим доверительный интервал:

$$\Delta_\gamma\left(\mathbf{X}\right) = \left(\bar{X} - \frac{\sigma}{\sqrt{n}}c_\gamma, \bar{X} + \frac{\sigma}{\sqrt{n}}c_\gamma\right). \qquad (2.38.)$$

Дисперсия неизвестна.

Пусть имеем общую нормальную модель с неизвестными параметрами: $N(\theta_1,\theta_2)$.

Центральной статистикой для оценивания среднего является

$$G(\mathbf{X};\theta_1) = \sqrt{n-1}\frac{\bar{X}-\theta_1}{S(\mathbf{X})}, \tag{2.39.}$$

где $S^2(\boldsymbol{X})$ — выборочная дисперсия. Данная статистика описывается распределением Стьюдента $S(n\text{-}1)$ с n -1 степенью свободы [11] Распределение Стьюдента симметрично относительно своей средней точки, поэтому расчет доверительного интервала производится так же, как в предыдущем примере.

γ — доверительный интервал для среднего получим в виде:

$$\left(\bar{X} - \frac{S(\mathbf{X})}{\sqrt{n-1}}t'_{\gamma,n-1},\ \bar{X} + \frac{S(\mathbf{X})}{\sqrt{n-1}}t'_{\gamma,n-1}\right), \tag{2.40.}$$

где $t'_{\gamma,n-1}$— $(1+\gamma)/2$-квантиль распределения $S(n$ - 1).

Построение доверительного интервала для дисперсии

Математическое ожидание известно.

Пусть имеем выборку из нормального распределения $N(m,\ \theta^{\ 2})$. Находим центральную статистику для $\tau = \tau(\theta) = \theta^{\ 2}$:

$$G(\mathbf{X};\tau) = \frac{1}{\tau}\sum_{i-1}^{n}(X_i - m)^2, \tag{2.41.}$$

которая подчиняется хи-квадрат распределению с n степенями свободы.

Если g_1 и g_2— граничные значения для статистики, то граничные значения γ-доверительного интервала для дисперсии запишутся как

$$T_1(\mathbf{x}) = \frac{1}{g_2}\sum_{i=1}^{n}(x_i - m)^2,\ \ T_2(\mathbf{x}) = \frac{1}{g_1}\sum_{i=1}^{n}(x_i - m)^2. \tag{2.42.}$$

Рассмотрим подробнее, как выбираются g_1 и g_2. Необходимо, чтобы выполнялись условия: $g_1 < g_2$ и $\int_{g_1}^{g_2} k_n(x)\,dx = \gamma$. На практике довольствуются *симметричным*[12] интервалом, т.е. выбирают симметричное положение интервала на шкале вероятностей:

$$\int\limits_0^{g_1} k_n(x)\,dx = \frac{1-\gamma}{2},\ \int\limits_{g_2}^{\infty} k_n(x)\,dx = \frac{1-\gamma}{2}. \tag{2.43.}$$

Отсюда видно, что

[11] Данная функция содержит другую статистику — выборочное среднее, что уменьшает число независимых слагаемых на единицу.

[12] Симметричный интервал не является наикратчайшим, но с ним удобнее работать.

$g_1 = \chi^2_{(1-\gamma)/2,n}$, $g_2 = \chi^2_{(1+\gamma)/2,n}$, $\chi^2_{p,n}$ — p–квантили распределения хи-квадрат.

Математическое ожидание неизвестно.

Центральной статистикой для дисперсии в этом случае является

$$G\left(\mathbf{X};\theta_2^2\right) = \frac{nS^2\left(\mathbf{X}\right)}{\theta_2^2}, \tag{2.44.}$$

здесь $S^2\left(\mathbf{X}\right)$— выборочная дисперсия. Доверительный интервал для дисперсии находим по рассмотренной выше схеме. Получаем *центральный* доверительный интервал для неизвестной дисперсии при неизвестном математическом ожидании:

$$\left(\frac{nS^2\left(\mathbf{X}\right)}{\chi^2_{(1+\gamma)/2,n-1}},\ \frac{nS^2\left(\mathbf{X}\right)}{\chi^2_{(1-\gamma)/2,n-1}}\right). \tag{2.45.}$$

Глава 3.

Проверка статистических гипотез

3.1. Основные понятия и терминология

Определение 3.1. *Статистической гипотезой называют любое утверждение о виде или свойствах распределения наблюдаемых случайных величин.*

Если для исследуемого явления сформулирована та или иная гипотеза (ее называют *основной* или *нулевой* и обозначают H_0), то задача проверки гипотезы состоит в том, чтобы сформулировать и обосновать такое *правило*, которое позволило бы по результатам наблюдений принять или отклонить эту гипотезу.

Определение 3.2. *Правило, согласно которому проверяемая гипотеза принимается или отвергается, называется статистическим критерием проверяемой гипотезы H_0.*

Примеры формулировок статистических гипотез

Пример 1.

Гипотеза о виде распределения. Пусть производится n независимых наблюдений над случайной величиной ξ с неизвестной функцией распределения $F_\xi(x)$. Гипотезой, подлежащей проверке может быть утверждение типа

$\boldsymbol{H}_0$: $F_\xi(x) = F(x)$, где функция $F(x)$ полностью задана, либо типа

$\boldsymbol{H}_0$: $F_\xi(x) \in \Phi_0$, где Φ_0— заданное семейство функций распределения.

Пример 2.

Гипотеза однородности. Произведено k серий независимых наблюдений:

$$(x_1, \ldots, x_{n1})_1, (x_1, \ldots, x_{n2})_2, \ldots, (x_1, \ldots, x_{nk})_k.$$

Можно ли с достаточной надежностью считать, что закон распределения наблюдений от серии к серии не менялся? Если это так, то говорят, что статистические данные однородны.

Пусть $F_i(x)$ — функция распределения наблюдений i-й серии, i=1, 2, ..., k. Тогда задача стоит в проверке гипотезы однородности $\boldsymbol{H}_0$: $F_1(x) \equiv \ldots \equiv F_k(x)$.

Пример 3.

Гипотеза независимости. В эксперименте наблюдается двухмерная случайная величина $\xi = (\xi_1, \xi_2)$ с неизвестной функцией распределения $F_\xi(x, y)$ и есть основания предполагать, что компоненты ξ_1 и ξ_2независимы. В этом случае надо проверить *гипотезу независимости* $\boldsymbol{H}_0$: $F_\xi(x, y) = F_{\xi 1}(x) \cdot F_{\xi 2}(y)$, где $F_{\xi 1}(x)$ и $F_{\xi 2}(y)$ — некоторые одномерные функции распределения. В общем случае рассматривают k-мерную случайную величину ξ и проверяют гипотезу независимости ее компонент.

Пример 4.

Гипотеза случайности. Результат эксперимента описывается $n-$мерной случайной величиной $X = (X_1, \ldots, X_n)$ с неизвестной функцией распределения $F_x(x)$, $x = (x_1, \ldots, x_n)$. Можно ли рассматривать X как случайную выборку из распределения некоторой случайной величины ξ (т. е. являются ли компоненты X_i независимыми и одинаково распределенными)? В данном случае требуется проверить *гипотезу случайности* $H_0 : F_x(x) = F_\xi(x_1) \cdot \ldots \cdot F_\xi(x_n)$, где $F_\xi(x)$ — некоторая одномерная функция распределения.

В рассмотренных примерах нулевые гипотезы формулируются как утверждения о принадлежности функций распределения некоторой случайной величины определенному классу распределений $\Phi_0 \in \Phi$: H_0: $F_x \in \Phi_0$. Распределения из дополнительного класса $\Phi_1 = \Phi \backslash \Phi_0$ называются *альтернативными распределениями.* А гипотезы вида $H_1 : \ F_x \in \Phi_1$, конкурирующие с основными гипотезами, называются *альтернативными гипотезами.*

В этих терминах задача формулируется как задача проверки гипотезы H_0 против альтернативы H_1. Любая гипотеза называется *простой,* если соответствующий класс распределений содержит лишь одно распределение, в противном случае гипотезы называются *сложными.*

Параметрическими называют гипотезы о значениях параметров распределений.

3.2. Общие принципы построения статистических критериев

На практике, для проверки гипотезы выбирается какая-либо функция выборки (статистика) $T(\mathbf{X})$. Значение этой статистики для заданной выборки и служит основанием для принятия или отклонения основной гипотезы.

Функция $T(\mathbf{X}) \in \tau$ называется *статистикой критерия,* а правило, которое для каждой реализации $\mathbf{x}$ выборки $\mathbf{X}$ должно приводить к одному из двух решений: принять гипотезу $\boldsymbol{H}_0$ или отклонить ее (принять $\boldsymbol{H}_1$) — называется *статистическим критерием.*

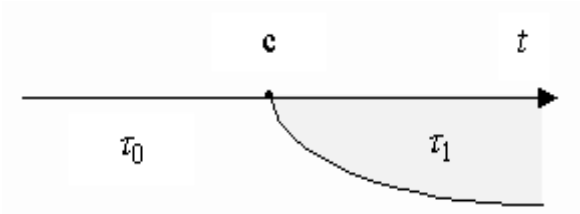


Рис. 3.1. К понятию критической области.

Каждому критерию отвечает разбиение области значений t статистики τ на две непересекающихся части τ_0 и τ_1 (рис.3.1).

Если значение статистики попадает в область τ_0, то принимается нулевая гипотеза H_0, в противном случае H_0 отвергается (принимается H_1). Множество значений статистики τ_0 называется *областью принятия нулевой гипотезы,* множество τ_1 — это область отклонения нулевой гипотезы или *критическая область.*

Нулевая гипотеза H_0 может конкурировать с несколькими альтернативными гипотезами. Каждой паре H_0 — H_1 соответствует свое разбиение области значений статистики критерия на τ_0 и τ_1.

Различают односторонние и двухсторонние критические области. На рисунке (3.1) показана *правосторонняя* критическая область.

Таким образом, задать критерий это значит

- задать статистику критерия;
- задать критическую область.

В ходе проверки гипотезы H_0 можно прийти к правильному выводу либо совершить два рода ошибок.

Ошибка первого рода — отклонить $H_{0,}$ когда она верна.

Ошибка второго рода — принять H_0, когда она не верна.

Вероятность ошибки первого рода выражается через функцию мощности критерия — вероятность попадания в критическую область при условии, что F — истинное распределение: $W(F) = P(t \in \tau_1|F)$:

$$\alpha = W(F_0)$$

Вероятность ошибки второго рода — вероятность попадания в область принятия нулевой гипотезы при условии, что F_1 — *истинное распределение (верна альтернативная гипотеза):*

$$\beta = 1 - W(F_1).$$

Проверку гипотезы желательно провести так, чтобы свести к минимуму ошибки обоего рода. Однако одновременно минимизировать обе ошибки невозможно.

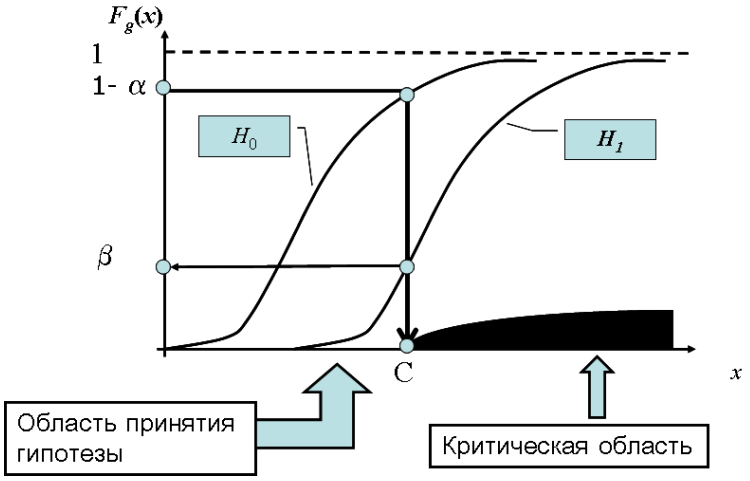


Рис. 3.2. Распределение статистики критерия для нулевой и альтернативной гипотез (односторонний критерий)

Рациональный принцип выбора критической обрасти формулируется следующим образом:

при заданном числе испытаний n устанавливается граница (α) для вероятности ошибки первого рода и при этом выбирается та критическая область, для которой вероятность ошибки второго рода минимальна.

Граничное значение α называется *уровнем значимости критерия.* События с такими вероятностями считаются практически невозможными. Допустимая величина уровня значимости определяется теми последствиями, которые наступают после совершения ошибки.

На практике для α выбирают одно из следующих стандартных значений: $0.05, 0.01, 0.005, \ldots$, для этих значений рассчитаны таблицы квантилей основных распределений, используемые при проведении расчетов доверительных интервалов и поверке гипотез.

Среди критериев выделяют такие, которые улавливают любые отклонения от нулевой гипотезы, без конкретизации вида отклонения. Такие универсальные критерии часто называют *критериями согласия.* Мы рассмотрим два наиболее известных: критерий Колмогорова и критерий хи-квадрат (χ^2).

Пять шагов проверки гипотезы

Сформулированные выше принципы проверки статистических гипотез вполне можно уложить в следующие «пять шагов»:

1. Сформулировать нулевую H_0 и альтернативную H_1 гипотезы.

2. Выбрать статистику критерия $T(X)$ и уяснить её закон распределения.

3. Задать уровень значимости критерия.

4. По таблицам квантилей распределения статистики найти критические точки и указать критическую область.

5. Подсчитать значение статистики критерия и проверить условие попадания в критическую область.

6. Сделать вывод о принятии нулевой или альтернативной гипотезы.

3.3. Критерий согласия Колмогорова

Критерий применяется для проверки гипотезы о виде распределения при условии, что теоретическая функция распределения непрерывная и полностью определена.

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из неизвестного распределения $F_\xi(x)$, о котором выдвинута простая гипотеза $H_0: F_\xi(x) = F(x)$. *Статистикой критерия* является величина

$$D_n = D_n(\mathbf{X}) = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|, \tag{3.1.}$$

представляющее собой максимальное отклонение эмпирической функции распределения $F_n(x)$ от гипотетической функции распределения $F(x)$.

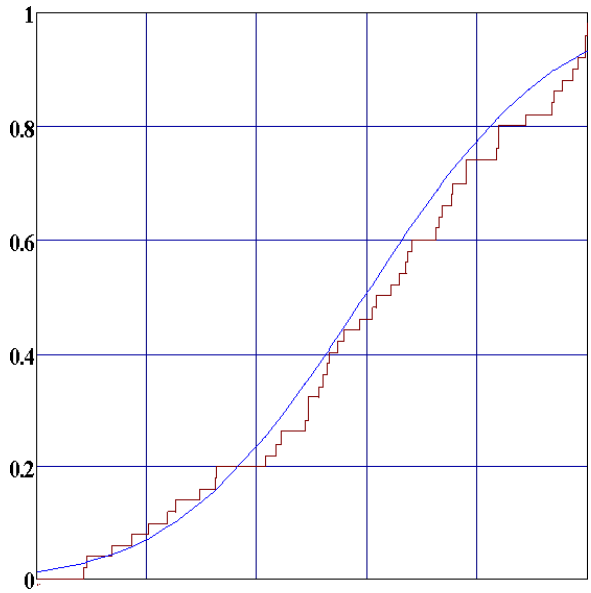


Рис. 3.3. Теоретическая и эмпирическая функции распределения.

Особенностью статистики D_n является то, что ее распределение при гипотезе H_0 не зависит от вида функции $F(x)$ и при достаточно больших n (уже при $n > 20$) практически от n не зависит и принимает вид:

$$\lim_{n\to\infty} \mathbf{P}\left(\sqrt{n}D_n \leqslant t\right) = K(t) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2t^2}. \qquad (3.2.)$$

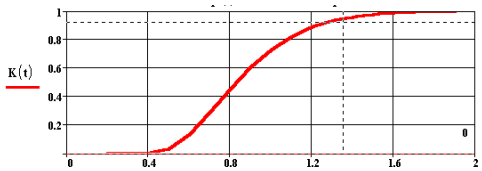


Рис. 3.4. Расчетные значения функции Колмогорова

Критическую границу при заданном уровне значимости α рассчитываем следующим образом:
из соотношения:

$$\mathbf{P}\left(\sqrt{n}D_n \geqslant \lambda_\alpha | H_0\right) \approx 1 - K(\lambda_\alpha) = \alpha \qquad (3.3.)$$

находим значения λ_α, которые протабулированы (см. табл. 2.1).

Таблица 2.1

α	0.05	0.01
λ_α	1.3581	1.6276

Значение статистики (3.1.) $t = D_n$ должно удовлетворять неравенству:

$$\sqrt{n}t \geqslant \lambda_\alpha. \qquad (3.4.)$$

Отсюда находим критическую область $t \geqslant \frac{\lambda_\alpha}{\sqrt{n}}$.

Теперь сформулируем правило проверки гипотезы — критерий согласия Колмогорова:

если наблюдавшееся значение $t = D_n$ статистики 3.1. удовлетворяет неравенству 3.4., то гипотезу H_0 отвергают; в противном случае делают вывод о том, что статистические данные не противоречат гипотезе.

Следуя этому правилу, можно ошибочно отклонить гипотезу H_0 с вероятностью, приблизительно равной α.

3.4. Критерий согласия хи-квадрат К. Пирсона

Данный критерий первоначально разработан для случая, когда наблюдаемая дискретная случайная величина, принимала N различных значений, причем закон распределения был полностью задан (простая гипотеза). Однако его можно применять и для непрерывных случайных величин, если предварительно сгруппировать исходные данные и перейти к их частотному представлению (Фишер).

Пусть $v = v_1, \ldots, v_N$ — наблюдаемый вектор попадания выборочных точек в интервалы группировки, а $p = p_1, \ldots, p_N$ — вектор теоретических (ожидаемых) значений вероятностей попадания в соответствующие интервалы группировки, здесь $v_1 + \ldots + v_N = n$. В этом случае вектор v имеет *полиномиальное* распределение и гипотеза H_0 сводится к проверке гипотезы о том, что вероятности полиномиального распределения вектора частот v имеют заданное значение $p_j,\ j = 1, \ldots, N$.

В качестве статистики, характеризующей отклонение выборочных (наблюдаемых) частот ν_j от соответствующих гипотетических (ожидаемых) значений принимается величина

$$X_n^2 = X_n^2(v) = \sum_{j=1}^{N} \frac{(v_j - np_j)^2}{np_j}, \tag{3.5.}$$

имеющая при больших n распределение хи-квадрат с (N –1) степенью свободы, когда закон распределения полностью задан и равно $N - 1 - r$, где r – число неизвестных параметров, дополнительно оцениваемых по выборке.

Критерий согласия хи-квадрат формулируется следующим образом:

пусть заданы уровень значимости α и объем выборки n и наблюдавшиеся значения вектора частот $v_1, \ldots, v_N$ удовлетворяют условиям $n \geqslant 50$, $v_j \geqslant 5$, $j = 1, \ldots, N$; тогда, если наблюдавшееся значение $t = X_n^2(v)$ статистики (3.5.) удовлетворяет неравенству $t \geqslant \chi^2_{1-\alpha, N-1-r}$, то гипотезу H_0 отвергают; в противном случае гипотеза H_0 не противоречит результатам испытаний.

Достоинством данного критерия является его универсальность и то, что его можно применять и в случае, когда данные носят *нечисловой* характер.

На практике, проверка статистической гипотезы, как правило, является составной частью решения более общей задачи исследования разного рода зависимостей. В следующих разделах мы рассмотрим многочисленные примеры проверки гипотез в рамках проведения факторного, корреляционного и регрессионного анализа.

Глава 4.

Однофакторный анализ

4.1. Постановка задачи.

При исследовании зависимостей между случайными величинами одной из наиболее простых является ситуация, когда можно указать только один фактор (величину), влияющую на конечный результат, и этот фактор может принимать лишь конечное число значений (уровней). Такие задачи и являются задачами однофакторного анализа.

Пример. Сравнение результатов нескольких различных способов действия, направленных на достижение одной цели. (Нескольких школьных учебников, нескольких лекарств, нескольких способов обработки и т.д.).

Терминология факторного анализа

То, что, как мы считаем, должно оказывать влияние на конечный результат, называют *фактором* (или *факторами,* если их несколько).

Конкретную реализацию фактора (определенный учебник, лекарство) называют *уровнем* фактора, или *способом обработки.*

Значение измеряемого параметра (т.е. величину результата) называют *откликом.*

Экспериментальные данные.

Для сравнения влияния факторов на результат необходим статистический материал. Его получают следующим образом: каждый из k способов обработки применяют несколько раз (не обязательно одно и то же) к исследуемому объекту и регистрируют результаты. В итоге имеем k выборок не обязательно одинакового объема. Результаты представляются в таблице.

Уровни факторов	1	2	...	k
Результаты измерений	$X_{1,1}$	$X_{1,2}$	...	$X_{1,k}$
	$X_{2,1}$	$X_{2,2}$	...	$X_{2,k}$
	...	...	...	...
	$X_{n1,1}$	$X_{n2,2}$	...	$X_{nk,k}$

Таблица 4.1. Данные: таблица с одним входом. $n1, n2 \ldots nk$ — объемы выборок, $N = n_1 + n_2 + \ldots + n_k$— общее число наблюдений.

4.2. Статистические предположения и стратегия факторного анализа

При организации исследований и последующем статистическом анализе надо учитывать такие характеристики статистического материала как *шкала измерений* и *характер случайной изменчивости* наблюдений – т.е. зависимость закона распределения от способов обработки.

Различают следующие виды шкал:

КОЛИЧЕСТВЕННЫЕ шкалы подразделяют на шкалы отношений и интервальные:

- интервальной шкалой называют такую шкалу с непрерывным множеством значений, в которой о двух сопоставляемых объектах можно сказать не только, одинаковы они или различны, не только в каком из них признак более выражен, но и *на сколько* более этот признак выражен;
- шкалой отношений называют такую шкалу с непрерывным множеством значений, в которой о двух сопоставляемых объектах можно сказать не только, одинаковы они или различны, не только в каком из них признак более выражен, но и *во сколько раз* более этот признак выражен.

ПОРЯДКОВЫЕ (ординальные). В данных шкалах существенен лишь порядок следования элементов, т.е. в каком из них признак более выражен.

НОМИНАЛЬНЫЕ. В них числа служат лишь для различения отдельных возможностей, заменяя названия и имена. Т.е. о двух сопоставляемых объектах можно сказать только, одинаковы они или различны.

Опыт показывает, что изменчивости больше подвержено положение случайной величины *медиана* или *среднее*.

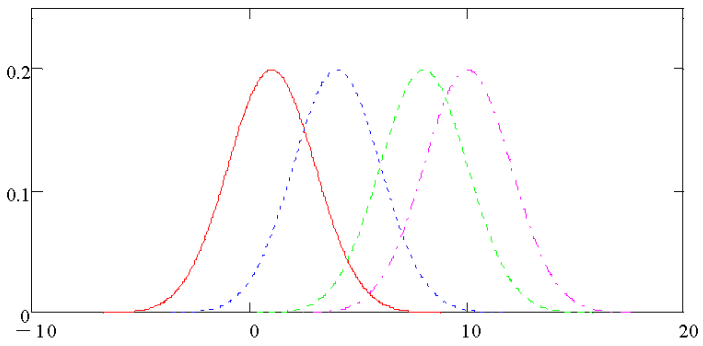


Рис. 4.1. Сдвиговое семейство распределений

Т.е. предполагается часто, что распределение выборок – одного вида (*сдвиговое семейство выборок*).

Для описания данных, представленных в таблице в большинстве случае приемлемой оказывается *аддитивная линейная* модель:

$$x_{ij} = a_j + e_{ij}\,,\ j = 1...k,\ i = 1\,...\,n. \tag{4.1.}$$

Здесь a_j –неслучайные неизвестные величины, являющиеся результатом действия соответствующих обработок.
e_{ij} – неизвестные, независимые, одинаково непрерывно распределенные случайные величины, отражающие изменчивость, внутренне присущую наблюдениям.

Если распределение e_{ij} нормальное, то в дальнейшем используют методы *дисперсионного анализа,* в противном случае применяются непараметрические методы – например, *ранговые* методы анализа.

Стратегия факторного анализа предполагает две стадии.

- Выяснение наличия влияния фактора.
- Количественная оценка эффекта обработки.

В начале следует ответить на вопрос: существует ли влияние фактора на измеряемый признак?

Для этого проводится сравнение изменчивости признака внутри каждой выборки и изменчивость между выборками.

На статистическом языке это формулируется в виде гипотезы *однородности* H_0: все данные принадлежат одному и тому же распределению.

Если гипотеза принимается, то анализ на этом и заканчивается (нет влияния факторов), если H_0 отвергается, то возникает следующая задача количественной оценки величины эффектов обработки и выяснение качества полученных оценок.

Примерная схема проведения однофакторного анализа приведена на рисунке (4.2). На этой схеме особо отмечены условия применимости дисперсионного факторного анализа:

1. Количественная шкала измерений.
2. Нормальный закон распределения исследуемых случайных величин.
3. Однородность дисперсий.

Поскольку эти три обязательных условия часто нарушаются, необходимо обратить внимание на применение и других, альтернативных методов проверки однородности статистического материала: непараметрических и ранговых методов.

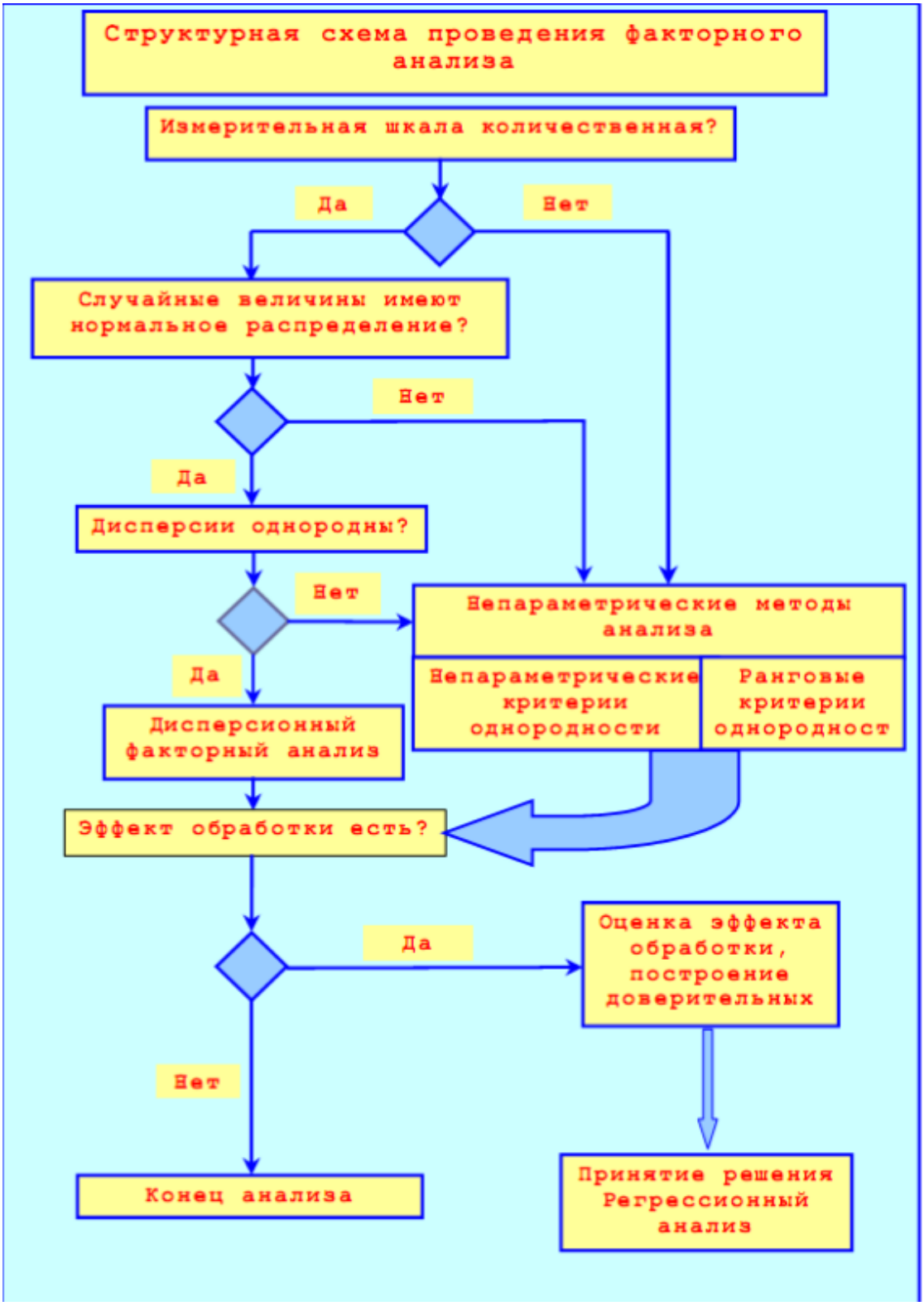


Рис. 4.2. Схема факторного анализа

4.3. Дисперсионный факторный анализ

В аддитивной модели $x_{ij} = a_j + e_{ij}$ мы предполагали, что e_{ij}– непрерывные, одинаково распределенные случайные величины. Часто о распределении e_{ij} можно сказать больше $L(e_{i,j} \sim N(0, \sigma^2$, где $\sigma^2)$- неизвестная дисперсия.

Дополнительная информация о законе распределения e_{ij} позволяет использовать более сильные методы в модели однофакторного анализа как для проверки гипотез, так и для оценки параметров. Совокупность этих методов носит название *однофакторного дисперсионного анализа*

Суть этого метода состоит в том, что сопоставляются две независимые оценки дисперсии σ^2. Одна оценка действует вне зависимости от какой-либо гипотезы относительно параметра a_j. Другая оценка σ^2 основывается на нулевой гипотезе: $H_0 : a_1 = a_2 = \ldots = a_k$.

Если две оценки заметно (значимо) различны, то H_0 следует отвергнуть.

Внутригрупповая дисперсия.

Будем рассматривать каждый столбец в (табл. 4.1) как отдельную выборку (группу данных). Для каждого столбца (выборки) найдем выборочное среднее $\overline{x}_j$ и выборочную дисперсию S_j^2.

Введем выборочное среднее

$$\overline{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \tag{4.2.}$$

статистика, построенная на выборочной дисперсии

$$\frac{n_j S_j^2}{\sigma^2} = \sum_{i=1}^{n_j} \left(\frac{x_{ij} - \overline{x}_j}{\sigma} \right)^2 \tag{4.3.}$$

имеет хи-квадрат распределение с $n_j - 1$ степенями свободы.

Объединенная (для всей таблицы) статистика

$$\sum_{j=1}^{K} \frac{n_j S_j^2}{\sigma^2} \tag{4.4.}$$

также будет иметь распределение хи-квадрат с числом степеней свободы R[1]

$$R = \sum_{j=1}^{K} \big(n_j - 1\big) = N - K \tag{4.5.}$$

[1]Воспользуемся свойством воспроизводимости по параметру хи-квадрат распределения.

Используя статистику (4.4.) с числом степеней свободы $N-K$ построим *несмещенную* оценку дисперсии:[2]

$$\sigma_1^2 = \frac{1}{N-k}\sum_{j}^{K} n_j S_j^2 \tag{4.6.}$$

которая не зависит от гипотез о распределении (от H_0). Эту оценку ещё называют *внутригрупповой* дисперсией.

Межгрупповая дисперсия

Рассмотрим совокупность выборочных средних $\{\overline{x}_j\}$. Согласно нулевой гипотезе они имеют нормальное распределение с параметрами $\left\{a, \dfrac{\sigma^2}{n_j}\right\}$. Введем выборочное среднее для всей совокупности данных

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{K} n_j \overline{x}_j. \tag{4.7.}$$

Используя совокупность групповых средних построим статистику, имеющую хи-квадрат распределение с $K-1$ степенью свободы:

$$\sum_{j=1}^{K}\left(\frac{\overline{x}_j - \overline{x}}{\sigma_{\overline{x_j}}}\right)^2 = \sum_{j=1}^{K} n_j \cdot \left(\frac{\overline{x_j} - \overline{x}}{\sigma}\right)^2. \tag{4.8.}$$

Используя (4.8.) введём еще одну несмещенную оценку дисперсии

$$\sigma_2^2 = \frac{1}{K-1}\sum_{j=1}^{k} n_j \cdot (\overline{x}_j - \overline{x})^2. \tag{4.9.}$$

Ее называют *межгрупповой* дисперсией. Основанная на гипотезе об однородности совокупности, данная оценка резко увеличивается при нарушении нулевой гипотезы.

Дисперсионное отношение Фишера (F - отношение)

Так как мы имеем две независимые несмещенные оценки дисперсии, то их отношение

$$F = \frac{\sigma_2^2}{\sigma_1^2} \tag{4.10.}$$

имеет распределение Фишера-Снедекора с $K-1$ и $N-K$ степенями свободы. Эта статистика уже не зависит от неизвестной дисперсии σ^2 и используется как

[2] Это следствие другого свойства хи-квадрат распределения: математическое ожидание статистики (4.4.) равно числу степеней свободы $N-K$

статистика критерия для проверки гипотезы об однородности всей совокупности данных, представленных в таблице.

Чем сильнее воздействие фактора на групповые средние, тем больше межгрупповая дисперсия, тем больше значение статистики F. Если значение статистики критерия становится больше критической точки для заданного уровня значимости α, то нулевая гипотеза отвергается.

Доверительные интервалы

Если нулевая гипотеза отвергается, то имеет смысл говорить о различных значениях a_j. Их оценками могут служить внутригрупповые средние $\overline{x}_j$, а оценкой дисперсии — несмещенная оценка внутригрупповой дисперсии σ_1^2

Построение доверительного интервала для a_j проводят по стандартной методике с помощью статистики

$$t = \frac{x_j - a_j}{\sigma_1}\sqrt{n_j},$$

имеющей распределение Стьюдента с $N - k$ степенью свободы.

Проверка гипотезы однородности дисперсий

Необходимость такой проверки возникает, как уже отмечалось, на начальной стадии проведения дисперсионного факторного анализа. Пусть $S_1^2,\ S_2^2, \ldots\ S_K^2$ — взаимнонезависимые несмещенные оценки дисперсий $\sigma_1^2,\ \sigma_2^2, \ldots\ \sigma_K^2$ соответственно и пусть $\dfrac{v_j S_j^2}{\sigma_j^2}$ подчиняются хи-квадрат распределению с v_j степенями свободы и предполагается, что $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2$, то для проверки такой гипотезы воспользоваться критерием Бартлетта, основанного на статистике:

$$\begin{aligned} M =& N \ln\left(\frac{1}{N}\sum_{j=1}^{K} v_j S_j^2\right) - \sum_{j=1}^{K} v_i \ln\ S_j^2 \\ N =& \sum_{j=1}^{K} v_j \end{aligned} \tag{4.11.}$$

Если гипотеза $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2$ верна и все $v_j > 3$, то отношение

$$M\left[1 + \frac{1}{3(K-1)}\left(\sum_{j=1}^{K}\frac{1}{v_j} - \frac{1}{N}\right)\right]^{-1} \tag{4.12.}$$

распределено приближенно как хи-квадрат с $K - 1$ степенями свободы. М-критерий Бартлетта очень чувствителен к отклонениям от нормальности исходных величин.

Еще один критерий (Кокрена) предназначен для проверки однородности дисперсий для случая, когда объемы всех выборок одинаковы: $v_1 = v_2 = \ldots = v_k = v$. Статистика критерия Кокрена G выражается формулой

$$G = \frac{S^2_{max}}{S^2_1 + S^2_2 + \cdots + S^2_K} \qquad (4.13.)$$
$$S^2_{max} = \max\{S^2_1, S^2_2, \cdots, S^2_K\}$$

Значения процентных точек приведены в таблицах [7].

Проверку гипотезы о равенстве дисперсий для двух выборок проще всего провести с помощью критерия Фишера. Пусть $S^2_{0,1}$ и$S^2_{0,2}$ – исправленные дисперсии для первой и второй выборок соответственно. Тогда их отношение $F = \frac{S^2_{0,1}}{S^2_{0,2}}$ будет иметь распределение Фишера-Снедекора со степенями свободы $R_1 = n_1 - 1$ и $R_2 = n_2 - 1$, где n_1, n_2 – объемы первой и второй выборок. При использовании данного критерия рекомендуется применять односторонний критерий и в качестве статистики критерия брать отношение большей дисперсии к меньшей (должно выполняться условие $F > 1$).

Критическую точку находим как $(1-\alpha)$-квантиль F-распределения:

$$C = qF(1-\alpha, R_1, R_2). \qquad (4.14.)$$

Нулевая гипотеза принимается, если $F < C$, в противном случае гипотеза о равенстве дисперсий отвергается на уровне значимости α.

4.4. Ранговый однофакторный анализ

Данный вид анализа применяется, когда нарушаются условия применимости дисперсионного анализа (см. схему на рисунке 4.2), если распределение случайной величины не известно или измерения сделаны в порядковой шкале (тестовые баллы, экспертные оценки).

Разберемся, в чем состоит суть ранговых методов. Согласно нулевой гипотезе, данные представленные в таблице 4.1 получены из одной и той же генеральной совокупности. Представим эти данные как одну большую упорядоченную выборку (вариационный ряд). В этом ряду каждый элемент получит свой порядковый номер (ранг). Заменим величины, представленные в таблице 4.1 их рангами. В результате получим следующую таблицу рангов.

Уровни факторов	1	2	...	k
Ранги результатов измерений	$r_{1,1}$	$r_{1,2}$	...	$r_{1,k}$
	$r_{2,1}$	$r_{2,2}$	...	$r_{2,k}$
	...	...	...	...
	$r_{n1,1}$	$r_{n2,2}$	...	$r_{nk,k}$

Таблица 4.2. Таблица рангов.

При выполнении нулевой гипотезы H_0 любые возможные распределения рангов по местам в таблице 4.2 являются равновероятными.

Нулевая гипотеза H_0 формулируется следующим образом: все K выборок однородны, т.е. являются выборками из одного и того же распределения.

Согласно общей методике проверки статистических гипотез, необходимо сконструировать такую статистику (в данном случае функцию от рангов – *ранговую статистику*) которая имела бы распределение, заметно отличающееся при H_0 и альтернативных гипотезах[3] и сформулировать правило (ранговый критерий), согласно которому по значению статистики критерия можно было бы либо принять, либо отклонить нулевую гипотезу.

Критерий Краскела-Уоллиса

Если ничего нельзя сказать об альтернативах, то можно воспользоваться данным критерием, т.к. он не зависит от распределения альтернатив.

При обработке таблицы рангов (табл. 4.2) получают следующие статистики:

Средний ранг, приходящийся на один элемент j-го столбца

$$\overline{R}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} r_{ij}, \tag{4.15.}$$

Средний ранг для всей совокупности $\{r_{ij}\}$

$$\overline{R} = \frac{1+2+\ldots+N}{N} = \frac{N+1}{2}. \tag{4.16.}$$

Если между столбцами нет систематических различий, то средние ранги $\overline{R}_j$ не должны отличаться от среднего ранга $\overline{R}$, рассчитанного по всей совокупности $\{r_{ij}\}$.

Статистика Краскела-Уоллиса строится следующим образом:

$$\begin{aligned} H =& \sum_{j=1}^{K}\left(\frac{\overline{R}_j - \overline{R}}{\sigma_{\overline{R}_j}}\right)^2 = \frac{12}{N(N+1)}\sum_{j=1}^{K} n_j\left(\overline{R}_j - \frac{N+1}{2}\right)^2 = \\ =& \frac{12}{N(N+1)}\sum n_j \cdot \overline{R}_j^2 - 3(N+1) \end{aligned} \tag{4.17.}$$

[3]Статистический критерий должен быть направлен против определенной совокупности альтернатив

Таблицы и асимптотика

Для малых n_j есть таблицы *квантилей,* за пределами таблиц можно воспользоваться асимптотическим распределением:

$$L(H) \approx \chi^2_{K-1}, \tag{4.18.}$$

если $H > \chi^2_{1-\alpha}$, то гипотеза отвергается с вероятностью ошибки α.

4.5. Проверка гипотезы однородности для двух выборок

Проверка гипотезы о равенстве мат. ож. для двух выборок из норм. распределения (независимые выборки)

Проверка гипотезы о равенстве мат. ож. для двух выборок из норм. распределения (зависимые выборки)

Критерий знаков

Задача проверки однородности статистического материала состоит в следующем.

Пусть имеются две независимые выборки $\mathbf{X} =(X_1, \ldots, X_n)$ и $\mathbf{Y} = (Y_1, \ldots, Y_m)$, описывающие один и тот же процесс, явление и т. д., но полученные в разное время или в разных условиях. Требуется установить, являются ли они выборками из одного и того же распределения или же закон распределения от выборки к выборке изменялся.

Такая задача может возникнуть, например, при контроле качества некоторой продукции, когда по контрольным выборкам из различных партий надо установить, не изменилось ли качество продукции от смены к смене в результате нарушений в технологическом процессе.

В общем виде задача формулируется следующим образом. Пусть имеются две независимые выборки

$$\mathbf{X} =(X_1, \ldots, X_n) \text{ и } \mathbf{Y} = (Y_1, \ldots, Y_m)$$

с неизвестными функциями распределения $F_1(x)$ и $F_2(y)$.

Требуется проверить *гипотезу однородности* H_0: $F_1(x) \equiv F_2(x)$.

Критерий однородности Смирнова.

Данный критерий применяется в случае непрерывных распределений. Он основан на статистике

$$D_{nm} = \sup_{-\infty<x<\infty} |F_{1n}(x) - F_{2m}(x)|, \tag{4.19.}$$

где $F_{1n}(x)$ и $F_{2m}(x)$— эмпирические функции распределения, построенные по выборкам X и Y соответственно. Если справедлива нулевая гипотеза, то эти

функции оценивают одну и ту же функцию распределения, поэтому статистика (4.19.) не должна существенно отклоняться от нуля (по крайней мере при больших объемах выборок). При расчете критических значений статистики $D_{n\ m}$ используют ее предельное при $n, m \to \infty$ распределение. В этом случае

$$\mathbf{P}\left(\sqrt{\frac{nm}{n+m}}D_{n\,m} \geqslant \lambda_\alpha | H_0\right) \approx 1 - K\left(\lambda_\alpha\right) = \alpha, \tag{4.20.}$$

где $K(x)$ – функция распределения Колмогорова (3.2.).

Сформулируем критерий однородности Смирнова:

если объемы выборок достаточно велики, то, вычислив по выборочным данным значение t статистики $D_{n\ m}$, принимают решение отвергнуть гипотезу H_0 в том и только в том случае, когда

$$t \geqslant \sqrt{\frac{n+m}{nm}}\lambda_\alpha. \tag{4.21.}$$

Вероятность ошибочно отвергнуть при этом истинную гипотезу примерно равна α.

4.5.1. Ранговые критерии

Иногда исходная статистическая информация может быть задана не числовыми значениями наблюдений, а отношением порядка между ними (типа «больше — меньше»). Обычно это имеет место в социологических исследованиях. В таких случаях наблюдения *ранжируются*, т. е. упорядочиваются по степени их предпочтения. Номер места, которое занимает наблюдение в таком упорядоченном ряду называют *рангом* соответствующего наблюдения.

Статистическая информация с самого начала может быть представлена рангами наблюдений. Статистические методы, которые применяют в таких случаях, называют *ранговыми методами,* а используемые статистики — *ранговыми статистиками;* критерии, основанные на таких статистиках — *ранговыми критериями.*

Ранговые методы можно применять и в тех случаях, когда заданы числовые значения наблюдений., так как при этом всегда можно упорядочить элементы выборки, построив соответствующий *вариационный ряд.* В этом случае рангом i-го наблюдения (X_i) является номер места R_i, которое занимает X_i в вариационном ряду$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$.

Рассмотрим некоторые примеры использования ранговых методов.

Глава 5.

Корреляционный анализ

Во многих практических задачах исследуются объекты, обладающие несколькими признаками, и необходимо выяснить, насколько эти признаки связаны между собой. Например, у каждого человека есть возраст и место рождения, уровень образования и годовой доход, пол и социальная принадлежность, и т.п.

Интерес состоит в том, можно ли по степени выраженности одного признака судить о выраженности другого, или же эти признаки следует считать проявляющимися независимо (с вероятностной точки зрения).

Практическую ценность таких исследований можно пояснить таким примером: если мы установим, что признаки «профессия» и «политические убеждения» независимы, то социологические опросы по предсказанию результатов парламентских выборов можно проводить без учета профессии опрашиваемых, что позволит уменьшить размер представительной выборки.

Как и в случае факторного анализа, на выбор методов обработки результатов измерений существенное влияние оказывает шкала измерений.

Вопрос о независимости признаков, представленных в разных измерительных шкалах (номинальной, порядковой и количественной) решаются своими методами.

5.1. Виды связей между случайными величинами

Можно выделить два вида зависимостей между наблюдаемыми величинами: функциональные и статистические (стохастические, случайные, корреляционные).

Функциональные связи являются результатом причинно-следственных отношений между величинами и, обычно, выражаются в форме физических законов, кроме того, на их формирование влияют случайные факторы;

Корреляционные связи возникают, когда среди случайных факторов, влияющих на формирование величин X и Y есть общие.

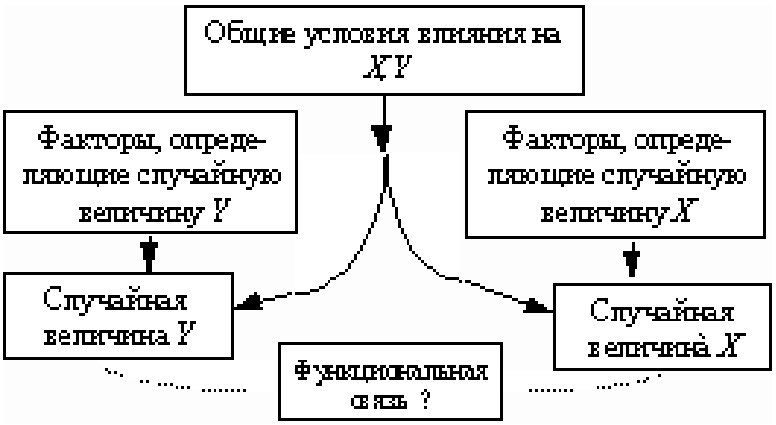


Рис. 5.1. Причины корреляции в отсутствие функциональной зависимости между величинами

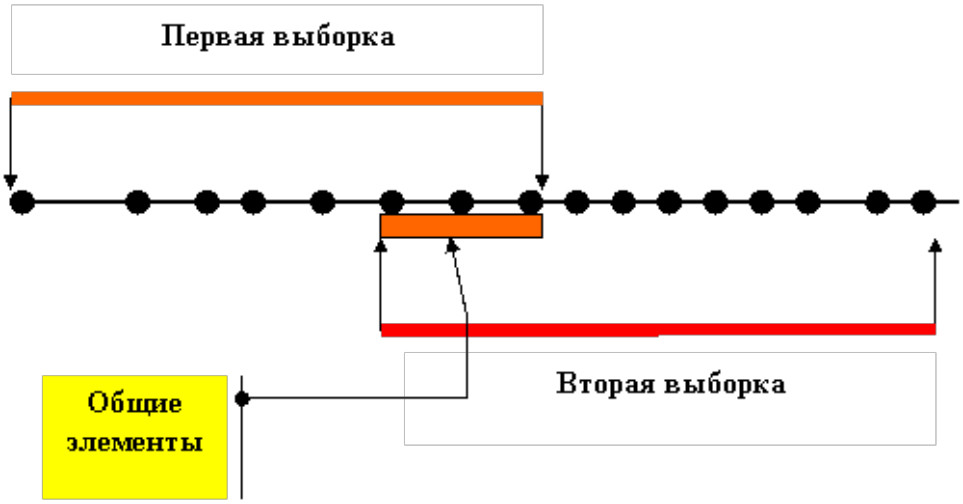


Рис. 5.2. Пример формирования статистической связи между двумя выборками

Определение 5.1. *Корреляция — это статистическая (вероятностная) зависимость между величинами, не имеющими строго функционального характера, проявляющаяся в том, что изменение значений одной величины влияет на закон распределения другой величины.*

Стратегия исследования зависимостей, таким образом, должна содержать два этапа.

- На первом этапе необходимо выявить наличие и силу связи между величинами. Эти задачи решаются корреляционным анализом.
- На втором этапе исследования необходимо выяснить конкретный вид функциональной зависимости между величинами. Эти задачи решаются регрессионным анализом.

В случае линейной функциональной связи корреляционный анализ полностью заменяет регрессионный анализ.

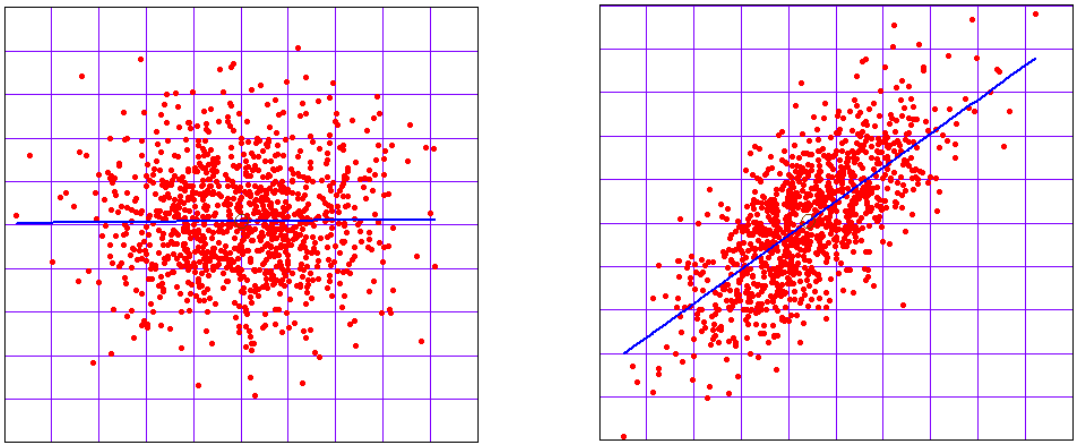
5.2. Приемы корреляционного анализа

Экспериментальные данные. Исходные данные представляют выборку из многомерной генеральной совокупности. В простейшем случае – это множество пар: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Графическое представление этих данных называется корреляционным полем (рис. 5.3).

По характеру расположения точек поля можно составить предварительное мнение о наличии и форме зависимости случайных величин.

Для численной обработки результаты обычно группируют и представляют в форме *корреляционной таблицы*. Группировка данных производится следующим образом.

- Корреляционное поле покрывается квадратной сеткой со стороной, равной выбранному интервалу группировки (см. рис.5.3).
- В качестве новых координат выбирают центры группировки: середины ячеек (x_i, y_i).
- Подсчитывается число точек v_{ij}, попавших в в ij-тую ячейку корреляционного поля, и это число заносится в ij-тую клетку корреляционной таблицы.



а: Отсутствие корреляции б: Наличие линейной корреляции

Рис. 5.3. Примеры корреляционных полей в отсутствии и при наличии линейной корреляции

X	Y			
	y_1	y_2	$\dots$	y_m
x_1	v_{11}	v_{12}	$\dots$	v_{1m}
x_2	v_{21}	v_{22}	$\dots$	v_{2m}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_n	v_{n1}	v_{n2}	$\dots$	v_{nm}

Таблица 5.1. Корреляционная таблица

Таким образом, после группировки число точек, подлежащих дальнейшей обработке значительно уменьшается, что сокращает объем вычислений без заметной потери точности.

Практически, проведение корреляционного анализа разбивается на этапы со своими специфическими приемами и методами обработки первичных данных.

Этапы корреляционного анализа. Можно предложить следующий алгоритм проведения корреляционного анализа экспериментальных данных для двух случайных величин:

- Построение корреляционного поля и корреляционной таблицы
- Вычисление выборочных коэффициентов корреляции и (или) корреляционных отношений.
- Проверка статистических гипотез о значимости связи.

Установление конкретного вида связи, как уже отмечалось, — это задача дальнейшего исследования методами *регрессионного анализа* и его разновидностей.

Расчет выборочных числовых характеристик системы случайных величин (оценка силы корреляционной связи).

Более точную информацию о характере и силе связи дают оценки *коэффициента корреляции* и *корреляционного отношения*.

Выборочный коэффициент корреляции r оценивается по формуле:

$$r = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}(x_i-\bar{x})(y_j-\bar{y})\,v_{ij}}{\sqrt{\sum\limits_{i=1}^{n} n_{i\bullet}(x_i-\bar{x})^2}\sqrt{\sum\limits_{j=1}^{m} m_{\bullet j}(y_j-\bar{y})^2}}, \qquad (5.1.)$$

где

$$n_{i\bullet} = \sum_{j=1}^{m} v_{ij},\ m_{\bullet j} = \sum_{i=1}^{n} v_{ij},$$
$$\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^{n} n_{i\bullet} x_i,\ \ \bar{y} = \frac{1}{N} \cdot \sum_{j=1}^{m} m_{\bullet j} y_j,\ N = \sum_{i=1}^{n} \sum_{j=1}^{m} v_{ij}\ .$$

Использование выборочного коэффициента корреляции в качестве меры связи между случайными величинами имеет четко определенный смысл только для *нормальных* распределений и распределений, близких к ним.

Во всех других случаях в качестве характеристики силы связи рекомендуется использовать *корреляционное отношение* η, интерпретация которого не зависит от вида исследуемой зависимости. Выборочное корреляционное отношение $\hat{\eta}_{Y|X}$ вычисляется по данным корреляционных таблиц:

$$\hat{\eta}^2_{Y|X} = \frac{\frac{1}{N} \cdot \sum\limits_{i=1}^{n} n_{i\bullet} \left(\bar{y}_i - \bar{y}\right)^2}{\frac{1}{N} \sum\limits_{j=1}^{m} m_{\bullet j} \left(\bar{y}_j - \bar{y}\right)^2}, \tag{5.2.}$$

где числитель характеризует рассеяние условных средних около безусловного среднего.

Аналогично определяется выборочное значение $\hat{\eta}_{X|Y}$. Величина $\hat{\eta}^2_{Y|X} - \hat{\rho}^2$ используется в качестве меры отклонения зависимости от линейной.

Проверка гипотезы о значимости связи

Основывается на знании законов распределений выборочных корреляционных характеристик. В случае нормального распределения величина выборочного коэффициента корреляции считается значимо отличной от нуля, если выполняется неравенство

$$\left(\hat{\rho}\right)^2 > \left[1 + (n-2)/t^2_\alpha\right]^{-1}, \tag{5.3.}$$

где t_α есть критическое значение квантили t-распределения Стьюдента с $n-2$ степенями свободы, соответствующее выбранному уровню значимости α.

При малых выборках используют $z-$преобразование Фишера: $z = \frac{1}{2} \ln \frac{1+\hat{\rho}}{1-\hat{\rho}}$, в результате которого получается статистика, не зависящая от ρ и n и имеющая распределение, близкое к нормальному. Исходя их приближенной нормальности z можно определить доверительные интервалы для истинного коэффициента корреляции ρ.

Ранговые критерии проверки гипотезы независимости

На практике для проверки гипотезы независимости часто используют ранговые критерии. Наиболее известным среди них является критерий Спирмена.

Критерий Спирмена.

Обозначим через R_i, ранг X_i, среди элементов соответствующего вариационного ряда; аналогично, введем S_i — ранг Y_i среди элементов вариационного ряда $Y_{(1)} \leqslant Y_{(2)} \leqslant \ldots \leqslant Y_{(n)}$.

Таким образом, из исходной выборки мы получаем множество пар рангов $(R_1,\ S_1), \ldots, (R_n,\ S_n)$.

Переставим теперь эти пары в порядке возрастания первой компоненты. Переобозначим полученное множество пар через $(1, T_1), \ldots, (n, T_n)$.

Введем ранговую статистику

$$\rho = \frac{\sum\limits_{i=1}^{n} \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(R_i - \bar{R}\right)^2 \sum\limits_{i=1}^{n} \left(S_i - \bar{S}\right)^2}}, \tag{5.4.}$$

представляющую собой коэффициент корреляции двух множеств рангов $(R_1, \ldots, R_n)$ и $(S_1, \ldots, S_n)$. Часто применяют следующую форму представления статистики:

$$\rho = 1 - \frac{6}{n\left(n^2 - 1\right)} \sum_{i=1}^{n} (i - T_i)^2. \tag{5.5.}$$

Коэффициент корреляции по модулю не превышает единицы. Если его значения близки к единице, то это свидетельствует против нулевой гипотезы, поэтому критическую область критерия Спирмена задают в виде $\tau_{1\,\alpha} = \{|\rho| \geqslant t_\alpha(n)\}$.

Для определения численного значения границы критической области при заданных объеме выборки n и уровне значимости α используют таблицы табулированного распределения статистики ρ, рассчитанные для $n = 2, \ldots, 30$. При больших n статистика $\sqrt{n}\rho$ стремится к *стандартному распределению.*

Глава 6.

Основные положения классического регрессионного анализа

Постановка задачи

Главная задача регрессионного анализа – создание математической модели объекта или явления на основе экспериментов или наблюдений.

Под математическими моделями мы понимаем, в данном случае, определенные математические отношения между показателями работы объекта $y_1, y_2, \ldots y_l$ и обусловливающими их величинами $x_1, x_2, \ldots, x_m$.

В специальной литературе совокупность величин y_i называют по-разному: зависимые переменные, выходные характеристики, отклики объектов и т.д. Совокупность величин x_i называют: входные переменные, независимые переменные, факторы.

Любая модель отражает только некоторые характерные черты объекта и никогда не бывает его точной копией. В зависимости от целей исследования один и тот же объект может описываться различными моделями.

Приближенность моделей обусловлена рядом причин: не учетом некоторых неизвестных факторов или случайным изменением некоторых факторов под действием случайных возмущений.

Следовательно, нельзя говорить об «истинной» модели в полном смысле слова. Тем не менее модели с успехом используются на практике.

Обычно под истинным значением понимают условное математическое ожидание зависимой переменной при заданном значении факторов (независимых переменных)

$$\eta\,(x_1, x_2,\, \ldots,\, x_m) = M[y|x_1, x_2,\, \ldots,\, x_m]. \qquad (6.1.)$$

Фактически, измеряемая выходная характеристика представляется как сумма двух составляющих:

$$y\,(x_1, x_2, \ldots, x_m) = \eta\,(x_1, x_2, \ldots, x_m) + \varepsilon \quad , \qquad (6.2.)$$

где ε - случайное возмущение с нормальным распределением и дисперсией σ^2, заменяющая действие на объект множества случайных возмущений.

Основанием такой замены служит *центральная предельная теорема теории вероятностей.*

6.1. Классификация регрессионных моделей

Конкретный вид функции $\eta\,(x_1, x_2, \ldots, x_m)$, т.е. вид модели, определяется характером решаемой задачи. Эти функции полностью определяются набором параметров $\beta_1, \beta_2, \ldots, \beta_k$, которые надо найти по экспериментальным данным.

В зависимости от того, как эти коэффициенты входят в уравнение регрессии, модели делятся на линейные и нелинейные (по параметрам).

Примеры.

$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ – линейная по параметрам.

$\eta = \beta_1 e^{\beta_2 x_1} + \beta_3 e^{\beta_4 x_2}$ – нелинейная по параметрам.

Мы будем рассматривать общую линейную модель:

$$y = \eta + \varepsilon = \sum_{i=1}^{n} \beta_i f_i\,(x_1, x_2, \ldots, x_m) + \varepsilon \qquad (6.3.)$$

Здесь, f_i – набор произвольных известных функций факторов, не содержащих неизвестных параметров. Функции f_i называют *регрессорами.*

Выбор того или иного набора регрессоров зависит от того, какой информацией мы располагаем о поведении исследуемой функции.

Так, если исследуется периодический процесс, то его наилучшим приближением служит разложение в тригонометрический ряд Фурье, если функция дифференцируема, то её можно представить отрезком ряда Тейлора и т.п.

6.2. Основные предположения классического регрессионного анализа

Стандартная процедура регрессионного анализа получила широкое практическое применение, поскольку она справедлива при некоторых, достаточно часто выполняемых предположениях.

Будем рассматривать модели линейного класса, для которых выходная характеристика представима в виде:

$$y_u = \sum_{i=1}^{k} \beta_i f_{i\,u} + \varepsilon \quad , \qquad (6.4.)$$

где $u = 1, 2, \ldots, N$ – номер наблюдения.

Представим совокупность всех значений f_i в форме матрицы F :

$$F = \begin{pmatrix} f_{1\,1} & f_{2\,1} & \cdots & f_{k\,1} \\ f_{1\,2} & f_{2\,2} & \cdots & f_{k\,2} \\ \cdot & \cdot & \cdot & \cdot \\ f_{1\,N} & f_{2\,N} & \cdots & f_{k\,N} \end{pmatrix}. \qquad (6.5.)$$

Данную матрицу называют матрицей *регрессоров*. При планировании эксперимента она еще называется расширенной матрицей плана эксперимента. В каждой строке F записаны значения f_i из модели при данном наблюдении. Для N наблюдения исходные данные представляют в виде следующей таблицы:

№ u	$x_1 x_2 \ldots x_u \ldots x_m$	y
1	$x_{11}\ x_{21} \ldots x_{i1} \ldots x_{m1}$	y_1
2	$x_{12}\ x_{22} \ldots x_{i2} \ldots x_{m2}$	y_2
$\vdots$	$\vdots\ \vdots\ \vdots\ \vdots$	$\vdots$
u	$x_{1u} x_{2u \ldots} x_{iu} \ldots x_{mu}$	y_u
$\vdots$	$\vdots\ \vdots\ \vdots\ \vdots$	$\vdots$
N	$x_{1N} x_{2N \ldots} x_{iN} \ldots x_{mN}$	y_N

Таблица 6.1.

В матричной форме уравнение регрессии можно записать в виде:

$$y = h + e = F \cdot b + e. \tag{6.6.}$$

Здесь

$y = (y_1, y_2, \ldots, y_N)^T$ – N- мерный вектор-столбец измеренных значений отклика.

β – вектор-столбец коэффициентов модели.

В классическом регрессионном анализе делают следующие основные предположения.

1. Величина $\varepsilon_u, u = 1, 2, \ldots, N$ есть случайная величина.

2. Случайная величина ε имеет нулевое математическое ожидание.

3. Значения случайной величины $\varepsilon_u, u = 1, 2, \ldots, N$ не коррелированы и имеют одинаковые дисперсии σ^2.

4. Случайная величина $\varepsilon_u, u = 1, 2, \ldots, N$ имеет нормальное распределение.

5. Матрица F не случайна.

6. На значения параметров модели β_i в модели не налагается никаких ограничений.

7. Ранг матрицы F равен числу коэффициентов в модели k.

Классическим регрессионным анализом называют процедуру оценивания регрессионных коэффициентов и статистический анализ модели, когда выполняются все семь предположений.

6.3. МНК-оценки коэффициентов регрессии

Поскольку результаты наблюдений есть случайные величины, то получить "истинные" значения коэффициентов β_1,β_2, ..., β модели нельзя. Вместо этого на основе данных (Таблица 4.1) можно получить их *оценки* b_1, b_2, ..., b_k.

Оценкой регрессии служит величина

$$\hat{y}_u = \sum_{i=1}^{k} b_i f_{iu}. \tag{6.7.}$$

Величина $\hat{y}_u$ называется *предсказанным значением* отклика.

Из-за действия случайных возмущений предсказанное значение $\hat{y}_u$ будет отличаться от результата измерений y_u.

Разности $e_u = y_u - \hat{y}_u$, u=1, 2, ..., N называют остатками.

В дальнейшем изложении мы будем использовать в основном матричную (векторную) форму записи.

Оценки коэффициентов регрессии естественно искать так, чтобы обеспечить наименьшие возможные остатки. Чтобы учесть все остатки, вводят функцию потерь:

$$Q = \sum_{u=1}^{N} (y_u - \hat{y}_u)^2 \tag{6.8.}$$

Или в векторной форме:

$$Q = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}})$$

Оценки коэффициентов регрессии находим из условия минимума суммы Q. Минимум суммы Q находим, приравняв нулю ее частные производные по неизвестным оценкам b_1, b_2, ..., b_k.

В результате мы получим систему линейных уравнений:

$$\mathbf{F}^T\mathbf{F}\mathbf{b} = \mathbf{F}^T\mathbf{y}. \tag{6.9.}$$

Решением данной системы является $\mathbf{b}$ - вектор оценок коэффициентов модели:

$$\mathbf{b} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{y}. \qquad (6.10.)$$

Введем обозначения:

$\mathbf{G} = \mathbf{F}^T\mathbf{F}$ – информационная матрица.

$\mathbf{C} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}$ – дисперсионная матрица.

Оценки, полученные методом наименьших квадратов, сокращенно называют МНК-оценками.

6.4. Основные свойства МНК-оценок

Оценки регрессионных коэффициентов, полученные с помощью МНК, — случайные величины, поскольку они основаны на случайных наблюдениях.

Свойства, не зависящие от вида распределения

МНК-оценки не смещены, т.е. их математические ожидания равны истинным значениям: $M[b] = \beta$.

Дисперсии и ковариации оценок регрессионных коэффициентов определяются по формулам:

$$D[b_i] = c_{ii}\sigma^2, \qquad (6.11.)$$

$$cov(b_i b_j) = c_{ij}\sigma^2 \qquad (6.12.)$$

где c_{ii}, c_{ij} — элементы дисперсионной матрицы $\mathbf{C}$.

Оценки, полученные с помощью МНК, эффективны, т.е. имеют наименьшие дисперсии среди всех возможных линейных несмещенных оценок.

МНК-оценки состоятельны.

Дисперсия предсказанного значения отклика $\hat{y}$ определяется по формуле:

$$D[\hat{\mathbf{y}}\,(\mathbf{x})] = \mathbf{f}_x^T\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{f}_x\sigma^2, \qquad (6.13.)$$

где

$$f_x = [f_1 x, f_2(x), \ldots, f_k(x)]^T$$

— вектор функций $f_i(x)$, вычисленный для совокупности факторов, заданной вектором X.

Если регрессионная модель выбрана правильно, то несмещенная оценка дисперсии задается выражением:

$$s^2 = \frac{1}{N-k}\sum_{u=1}^{N}\left(y_u - \hat{y}_u\right)^2. \qquad (6.14.)$$

Свойства, связанные с предположением о нормальном распределении.

Если случайные возмущения $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ – независимые нормально распределенные случайные величины с нулевым математическим ожиданием и одинаковыми дисперсиями σ^2, то вектор оценок регрессионных коэффициентов $\mathbf{b}$ имеет многомерное нормальное распределение с математическим ожиданием β и матрицей дисперсий-ковариаций $C\sigma^2$.

Отношение $\chi^2 = Q/\sigma^2$ имеет распределение χ^2 с числом степеней свободы, равным $N - k$.

Вектор оценок b и оценка дисперсии s^2 распределены независимо друг от друга.

6.5. Статистический анализ качества регрессионной модели

Полученные оценки параметров регрессионной модели обеспечивают высокое качество оценки регрессии только при условии, что ее структура $\mathbf{F}\beta$соответствует истинной структуре $\mathbf{F}_0\beta_0$. Но на практике истинная модель заранее не известна. Поэтому приходится перебирать различные модели, находя ту, которая лучше всего согласуется с экспериментальными данными. Такую модель называют *адекватной* . Она должна удовлетворять условию:

$$M\{y\} = F \cdot \beta S.$$

Адекватная модель не обязательно должна совпадать с истинной. Более того, адекватная модель не единственная. Общим для всех адекватных моделей является то, что для каждой из них существует неособенное линейное преобразование, приводящее ее к истинной модели.

При выборе структуры модели возможны два вида ошибок.

Модель содержит меньше параметров, чем истинная — так называемый *недобор* параметров.

Число параметров в модели больше, чем в истинной — *перебор* параметров.

При недоборе параметров оценки параметров регрессии оказываются смещенными (появляется систематическая ошибка) и несостоятельными.

При переборе параметров оценки являются несмещенные и состоятельные, но при этом теряется точность.

Таким образом, недобор параметров — более серьезный недостаток регрессионной модели, чем перебор, но чрезмерное усложнение модели снижает эффективность (точность) оценивания, увеличивая дисперсию.

При анализе качества регрессионной модели последовательно проверяется два вида гипотез: гипотеза об адекватности модели, т.е. проверяется соответствие выбранного класса функций регрессии истинной функции регрессии;

процедура проверки использует методы дисперсионного анализа. гипотеза о *значимости* параметров модели, т. е. о равенстве нулю соответствующего параметра; процедура проверки позволяет убрать из модели “лишние” параметры, что приводит к повышению точности оценок остальных.

6.6. Нелинейная регрессия

На практике очень часто вид функции регрессии бывает известен, например, при исследовании зависимостей, выражаемых физическими законами. При этом вид функции регрессии полностью определяется набором параметров, часть из которых входит в уравнение регрессии не линейно. Рассмотрим нелинейную модель регрессии вида

$$y = \eta + \varepsilon = f\left(\beta_1, \beta_2, \ldots, \beta_r, x_1, x_2, \ldots, x_m\right) + \varepsilon. \qquad (6.15.)$$

Для определения параметров модели (6.15.) нелинейную функцию регрессии заменяют линейной (линеализация функции), с последующим применением аппарата классического регрессионного анализа.

Способы линеализации

Линеализацию функций проводят двумя способами:

1. введением новых переменных;
2. заменой функции линейной частью отрезка ряда Тейлора.

Замена переменных

Функция	Линейная форма
$\eta(x) = \beta_0 x^{\beta_1}$	$\ln \eta(x) = \ln \beta_0 + \beta_1 \ln x$
$\eta(x) = \beta_0 e^{\beta_1 x}$	$\ln \eta(x) = \ln \beta_0 + \beta_1 x$
$\eta(x) = \beta_0 e^{\beta_1 x} + \beta_2$	$\ln(\eta(x) - \beta_2) = \ln \beta_0 + \beta_1 x$

Таблица 6.2. Примеры нелинейных функций и их линейных форм

Итерационные методы, использующие разложение в ряд Тейлора

Полагаем, что функция регрессии есть дифференцируемая функция неизвестных параметров. Тогда, ограничиваясь линейной частью разложения функции в ряд Тейлора по неизвестным параметрам, мы получим линейную по параметрам функцию:

$$\eta(x,b) = \eta(x,b_0) + \left.\frac{\partial\eta(x,b)}{\partial b}\right|_{b_0}(b_1 - b_0),$$

где

b_0— начальное приближение параметра;

b_1 — следующее приближение параметра.

Задавая начальное приближение, находим последующее приближение, применяя процедуру классического регрессионного анализа. Итерационный процесс можно продолжать до тех пор пока не выполнится некоторое условие остановки.

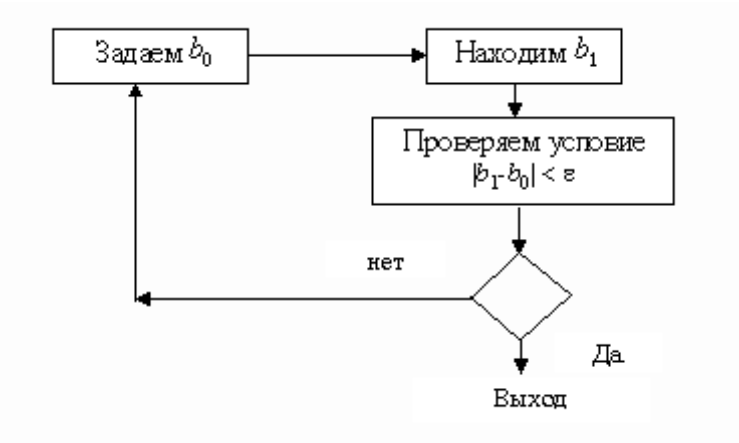


Рис. 6.1. Блок-схема организации итерационного цикла

6.7. Непараметрический регрессионный анализ

При анализе зависимостей часто представляет интерес не сама функция регрессии, а ее производные или другие функционалы или некоторые ее свойства: монотонность, наличие экстремумов, расположение нулей или величина экстремальных значений.

При большом числе пар данных картина корреляционного поля не всегда позволяет выявить какую-либо регрессионную зависимость. Возможен просто обман зрения из-за большого числа точек или из-за нечетких структур.

Цель регрессионного анализа состоит в разумной аппроксимации неизвестной функции отклика. За счет уменьшения ошибок наблюдений становится возможным сосредоточить внимание на важных деталях средней зависимости Y

от X при ее интерпретации. Эта процедура аппроксимации обычно называется «сглаживанием».

Аппроксимация регрессии может быть выполнена двумя способами.

Применением параметрических моделей, для которых важен только аналитический вид и не имеет значения физический смысл параметров или модели, представляющие собой физические законы. Параметры модели имеют определенный физический смысл [1]

Непараметрическая регрессия. В данной модели не вводится никаких параметрических (аналитических) зависимостей между входными и выходными характеристиками. Функция регрессии ищется в виде таблицы данными.

6.7.1. Цели непараметрического регрессионного анализа

Непараметрический подход к оцениванию регрессионной кривой преследует 4 цели:

- он представляет гибкий метод исследования соотношения между двумя переменными;
- он позволяет предсказывать наблюдения, которые еще только должны быть сделаны без привязки к фиксированной параметрической модели;
- он дает нам средство нахождения ложных наблюдений путем изучения влияния изолированных точек;
- он порождает гибкий способ подстановки пропущенных значений или интерполяции между соседними значениями переменной x.

Гибкость метода полезна при предварительном исследовании для определения параметрической зависимости.

6.7.2. Основная идея сглаживания

Уже при визуальном изучении множества точек на корреляционном поле, мы мысленно пытаемся заменить его плавной кривой, проходящей по центрам групп скопления точек переменного размера. В качестве таких центров естественно брать усредненные по некоторой окрестности x значения Y.

Эта процедура *«локального усреднения»* рассматривается как основная идея сглаживания.

Формально, процедура сглаживания определяется следующим выражением:

[1] Пример параметрического подхода — рассмотренный нами классический регрессионный анализ.

$$m(x) = M[Y|X] = n^{-1}\sum_{i=1}^{n} W_{n\,i}\,(x)\,Y_i, \qquad (6.16.)$$

где $\{W_{n\,i}\,(x)\}_{i=1}^{n}$ означает последовательность весов, которые могут зависеть от всего вектора $\{X_i\}_{i=1}^{n}$; $m(x)$ — МНК-оценка регрессии.

Основная идея локального усреднения эквивалентна процедуре нахождения оценки локально взвешенных наименьших квадратов.

6.7.3. Методы сглаживания

Простейшая процедура сглаживания состоит в представлении данных в виде «регрессограммы»[2] . В этом случае веса W_{ni}(x) берутся постоянными на блоках постоянной длины.

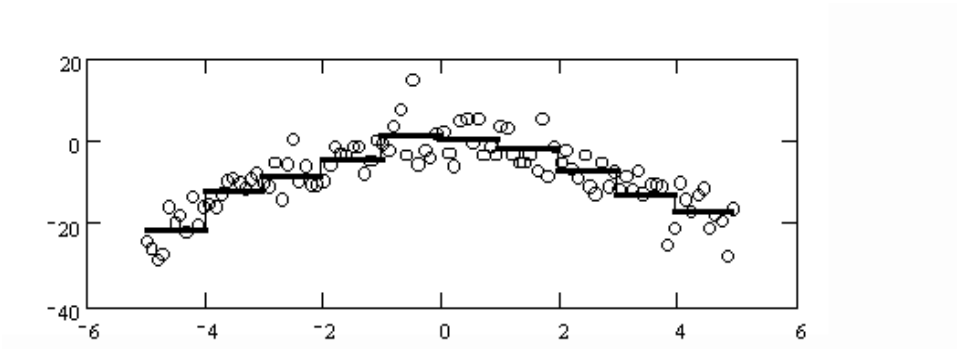


Рис. 6.2. Сглаживание результатов с помощью «регрессограммы»

Мы рассмотрим далее два метода: ядерное сглаживание и сглаживание по $k-$ ближайшим соседям.

Ядерное сглаживание

Идея метода: последовательность весов $W_{ni}(x)$ описывается функцией плотности со скалярным параметром, который регулирует размер и форму весов около x. Эта функция формы называется ядром K. Ядро — это непрерывная, ограниченная функция K с единичным интегралом

$$\int\limits_X K\,(u)\,du = 1 \quad .$$

Веса определяются по формуле

$$W_{n\,i}\,(x) = \frac{\frac{1}{h_n}K_{h_n}\left(\frac{x-x_i}{h_n}\right)}{\frac{1}{h_n}\sum\limits_{i=1}^{n} K_{h_n}\left(\frac{x-x_i}{h_n}\right)}, \qquad (6.17.)$$

[2] Построение «регрессограммы» похоже на построение гистограммы, отсюда и название.

где h_n — параметр масштаба (ширина окна).

Обычно, ядром берут такие функции, которые равны нулю вне окна. В качестве примера приведем формулу так называемого *ядра Епанечникова*

$$K(u) = 0.75\left(1 - u^2\right), |u| \leqslant 1. \qquad (6.18.)$$

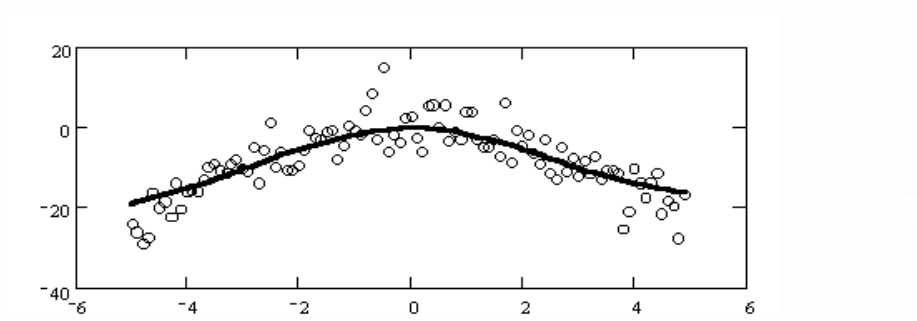


Рис. 6.3. Ядерное сглаживание с ядром Епанечникова

Оценки k-ближайших соседей

Эти оценки дают среднее взвешенное значение в изменяющейся окрестности. Эта окрестность определяется только теми значениями переменной X, которые являются k ближайшими к x по *евклидовому расстоянию.*

$$m_k(x) = \frac{1}{n}\sum_{i=1}^{n} W_{k\,i} Y_i. \qquad (6.19.)$$

Введем множество индексов $J_x = \{i : X_i$ одно из ближайших k наблюдений x $\}$. Тогда для весов получим:

$$W_{ki}(x) = \begin{cases} n/k, & i \in J_x \\ 0, & i \notin J_x \end{cases}$$

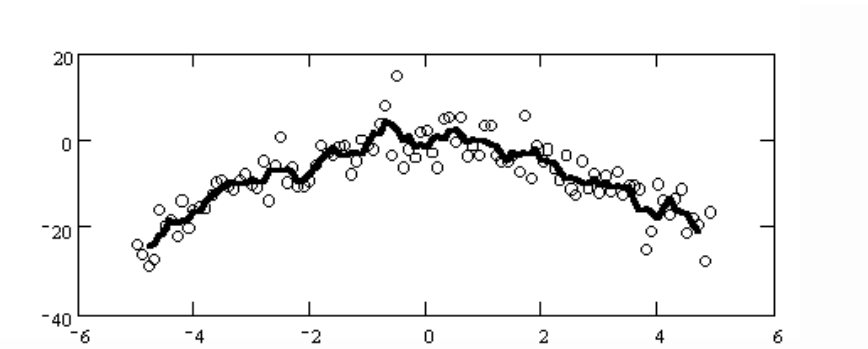


Рис. 6.4. Сглаживание по 5 ближайшим соседям

Приложение А

Приложения

А1. Статистические функции, встроенные в MathCad

cvar(A, B) Returns the covariance of the elements of A and B given by
where the overbar indicates the complex conjugate.

corr(A, B) Returns the Pearson's r correlation coefficient of the elements in A and B.

Arguments: A and B are real or complex arrays of the same size.

Содержание

Общие сведения **3**

1. Основные понятия математической статистики **5**
1.1. Первичные данные и их представление 5
1.2. Математическая модель выборки 6
1.3. Гистограмма и полигон 8

2. Оценка неизвестных параметров распределений **11**
2.1. Точечное оценивание неизвестных параметров 11
2.1.1. Требования к оценкам 11
Принцип наименьших квадратов. 11
Оценка математического ожидания 13
Оценка дисперсии 13
2.1.2. Требования к статистикам 13
2.1.3. Метод моментов 15
2.1.4. Метод наибольшего правдоподобия 15
2.1.5. Робастные оценки 17
Причины отклонения от нормальности 17
Понятие о робастных оценках 18
Оценки параметра сдвига 18
Оценка параметров масштаба распределений 19
2.2. Интервальное оценивание неизвестных параметров 20
Понятие доверительного интервала 20
2.2.1. Центральные статистики 21
2.2.2. Распределение некоторых функций от выборки 22
Квантили распределений 22
Стандартное распределение 23
Распределение хи-квадрат 23
Распределение Стьюдента (t распределение) 24
Распределение Фишера-Снедекора 24
Доверительные интервалы для математического ожидания 25
Построение доверительного интервала для дисперсии . . . 26

3. Проверка статистических гипотез **28**
3.1. Основные понятия и терминология 28
Примеры формулировок статистических гипотез 28
3.2. Общие принципы построения статистических критериев 29

Пять шагов проверки гипотезы 31
3.3. Критерий согласия Колмогорова 32
3.4. Критерий согласия хи-квадрат К. Пирсона 34

4. Однофакторный анализ **35**
4.1. Постановка задачи. 35
Терминология факторного анализа 35
Экспериментальные данные. 35
4.2. Статистические предположения и стратегия факторного анализа 36
4.3. Дисперсионный факторный анализ 39
Внутригрупповая дисперсия. 39
Межгрупповая дисперсия 40
Дисперсионное отношение Фишера (F - отношение) 40
Доверительные интервалы 41
Проверка гипотезы однородности дисперсий 41
4.4. Ранговый однофакторный анализ 42
Таблицы и асимптотика 44
4.5. Проверка гипотезы однородности для двух выборок 44
4.5.1. Ранговые критерии 45

5. Корреляционный анализ **46**
5.1. Виды связей между случайными величинами 46
5.2. Приемы корреляционного анализа 48
Экспериментальные данные. 48
Этапы корреляционного анализа. 49

6. Основные положения классического регрессионного анализа **52**
6.1. Классификация регрессионных моделей 53
6.2. Основные предположения классического регрессионного анализа 53
6.3. МНК-оценки коэффициентов регрессии 55
6.4. Основные свойства МНК-оценок 56
6.5. Статистический анализ качества регрессионной модели 57
6.6. Нелинейная регрессия 58
6.7. Непараметрический регрессионный анализ 59
6.7.1. Цели непараметрического регрессионного анализа 60
6.7.2. Основная идея сглаживания 60
6.7.3. Методы сглаживания 61

A Приложения **63**
A1. Статистические функции, встроенные в MathCad 63

Список литературы

1. Ивченко Г.И., Медведев Ю.И. Математическая статистика — М.: Высш. Шк., 1992.—304 с.

2. Гмурман В.Е. Теория вероятностей и математическая статистика. М.— 1972, 1998.

3. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М. — 1975.

4. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ.—М.: “Финансы и статистика”, 1987 г.—239 с.

5. Хардле В. Прикладная непараметрическая регрессия.— М.: Мир, 1993.—347 с.

6. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере.—М.: ИНФРА-М, “Финансы и статистика”, 1995 — 384 с.

7. Большев Л.Н., Смирнов Н.В.–М.: Наука. Главная редакция физико-математической литературы, 1983.–416 с.

Учебное издание

ГАЛАНОВ Юрий Иванович

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

Научный редактор
доктор физ.-мат. наук,
профессор К.П. Арефьев

Редактор Н.Я. Горбунова

Верстка в системе *MikTeX* 2.7
с использованием шрифтов
из пакета LHFONTS 3.5. Ю.И. Галанов

Дизайн обложки И.О. Фамилия

Подписано к печати 00.00.2008. Формат 60x84/8.
Бумага «Снегурочка». Печать XEROX.
Усл.печ.л. 000. Уч.-изд.л. 000. Заказ XXX. Тираж XXX экз.

Томский политехнический университет
Система менеджмента качества
Томского политехнического университета
сертифицирована
NATIONAL QUALITY ASSURANCE по стандарту
ISO 9001:2000

ИЗДАТЕЛЬСТВО ТПУ 634050, г. Томск, пр. Ленина, 30.