Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра информационных технологий
автоматизированных систем

В. С. Муха

Статистические методы обработки данных

Часть 1

Учебно-методическое пособия для студентов специальности "Автоматизированные системы обработки информации"

Минск 2007

УДК 519.21

Муха, В.С. Статистические методы обработки данных: Учебное пособие для студентов технических специальностей высших учебных заведений / В.С. Муха. – Мн.: БГУИР, 2007.

Излагается методы математической статистики для обработки результатов измерений. Рассматриваются основные понятия математической статистики, выборочный метод, точечные и интервальные оценки параметров, проверка гипотез, корреляционный анализ, регрессионный анализ, теория статистических решений, робастность статистических процедур, однофакторный дисперсионный анализ, статистика случайных процессов, стохастическая аппроксимация. Пособие предназначено для студентов, магистрантов и аспирантов технических специальностей высших учебных заведений.

Ил. – 23, табл. – 7, спис. лит. – 24 назв.

ПРЕДИСЛОВИЕ

Под статистическими методами обработки данных будем понимать методы обработки числовых данных, полученных в результате опытов (наблюдений) над объектами, случайными по своей природе, или содержащих ошибки наблюдений над детерминированными объектами. Теоретической основой статистических методов обработки данных является математическая статистика. *Математическая статистика* – это математическая наука, занимающаяся разработкой методов получения научно обоснованных выводов о случайных явлениях из экспериментальных (эмпирических) данных. Один из крупнейших статистиков современности, руководитель Индийского статистического института С. Р. Рао, оценивает математическую статистику как новый метод 20-го века [15]. Этот метод не потерял своей актуальности и сейчас, активно развивается, применяется на практике, и есть все основания считать, что он по праву остается также методом 21-го века. Статистические методы обработки данных представляют собой важную составляющую в системе подготовки специалистов технических специальностей высших учебных заведений. В совокупности с численными методами они образуют, пожалуй, подавляющую часть полного набора методов обработки числовой (математической) информации.

Задачи математической статистики являются задачами синтеза, так как в них речь идет о синтезе устройств обработки экспериментальных данных. Это значит, что статистические методы обработки данных практичны по своей сущности. В данном пособии наряду с обоснованием основных положений математической статистики автор стремился сохранить и усилить эту практическую направленность путем доведения изложенных методов до возможности их программирования и практического использования.

ОСНОВНЫЕ ОБОЗНАЧЕНИЯ

$X = (x_1, x_2, ..., x_n)$ – выборка объема n из генеральной совокупности.

$x_{(1)}, x_{(2)}, ..., x_{(n)}$ – вариационный ряд.

$x_{(m)}$ – m-я порядковая статистика.

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

$\bar{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – выборочная дисперсия.

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – исправленная (несмещенная) выборочная дисперсия.

$\bar{s}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$ – выборочная дисперсия при известном математическом ожидании.

$\bar{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ – выборочный начальный момент k-го порядка.

$\bar{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$ – выборочный центральный момент k-го порядка.

$\bar{s}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ – выборочный коэффициент ковариации.

$\bar{r}_{xy} = \frac{\bar{s}_{xy}}{\bar{s}_x \bar{s}_y}$ – выборочный коэффициент корреляции.

$N(a, \sigma^2)$ – одномерное нормальное распределение с математическим ожиданием a и дисперсией σ^2.

$H_1(n)$ – одномерное распределение хи-квадрат с n степенями свободы.

$T_1(n)$ – одномерное распределение Стьюдента с n степенями свободы.

$F_1(m, n)$ – одномерное распределение Фишера с m, n степенями свободы.

$v \in H_1(n)$ – случайная величина v имеет распределение $H_1(n)$.

$\xi_n \xrightarrow{p} \xi$ – ξ_n сходится к ξ по вероятности при $n \to \infty$.

$\xi_n \xrightarrow{п.н.} \xi$ – ξ_n сходится к ξ почти наверное, почти всюду или с вероятностью 1 при $n \to \infty$.

$\xi_n \xrightarrow{(2)} \xi$ – ξ_n сходится к ξ в среднем квадратичном при $n \to \infty$.

$\xi = \underset{n\to\infty}{l.i.m.}\xi_n$ – ξ есть предел в среднем квадратичном последовательности ξ_n.

$C_n^m = \dfrac{n!}{m!(n-m)!}$ – число сочетаний из n элементов по m элементов.

$\bar{\hat{\theta}} = (\hat{\theta}_1,...,\hat{\theta}_m)$ – точечная оценка векторного параметра $\bar{\theta} = (\theta_1,...,\theta_m)$.

$E(\xi)$ – математическое ожидание случайной величины ξ.

$D(\xi)$ – дисперсия случайной величины ξ.

$cov(\xi,\eta) = E\big((\xi - E(\xi))(\eta - E(\eta))\big)$ – коэффициент ковариации случайных величин ξ` и η.

R^n – n-мерное арифметическое пространство.

R^1 – действительная прямая (одномерное арифметическое пространство).

1 ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ

1.1 Законы больших чисел

Законы больших чисел теории вероятностей – это теоремы, в которых устанавливается сходимость некоторой случайной последовательности ξ_n при $n \to \infty$ (при большом числе n). Эти законы позволяют выяснять свойства точечных оценок характеристик и параметров распределений в математической статистике.

Теорема Бернулли (закон больших чисел в форме Бернулли). Если $p = P(A)$ – вероятность события A в одном испытании Бернулли, а m – количество появлений этого события в n независимых испытаниях, то для любого $\varepsilon > 0$ выполняется соотношение

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1.$$

В этой теореме утверждается, что относительная частота m/n случайного события A сходится по вероятности к вероятности p этого события при увеличении числа испытаний.

Теорема Чебышева (закон больших чисел в форме Чебышева). Если $\xi_1,...,\xi_n$ – независимые случайные величины с конечным математическим ожиданием $E(\xi_i) = a$ и конечными дисперсиями, ограниченными одной и той же константой c, то для любого $\varepsilon > 0$ выполняется соотношение

$$P(|\frac{1}{n}\sum_{i=1}^{n}\xi_i - a| > \varepsilon) \underset{n\to\infty}{\to} 0.$$

В этой теореме утверждается, что среднее арифметическое случайных величин $\xi_1,...,\xi_n$ с указанными свойствами сходится по вероятности к среднему арифметическому их математических ожиданий.

Теорема Хинчина (закон больших чисел в форме Хинчина). Если $\xi_1,...,\xi_n$ – независимые, одинаково распределенные случайные величины с конечным математическим ожиданием $E(\xi_i) = a$, то $\forall \varepsilon > 0$ выполняется соотношение

$$P(|\frac{1}{n}\sum_{i=1}^{n}\xi_i - a| > \varepsilon) \underset{n\to\infty}{\to} 0.$$

В этой теореме также утверждается, что среднее арифметическое случайных величин $\xi_1,...,\xi_n$ с указанными свойствами сходится по вероятности к среднему арифметическому их математических ожиданий. В отличие от теоремы Чебышева здесь не требуется существования дисперсии случайных величин $\xi_1,...,\xi_n$, однако эти величины должны быть одинаково распределенными.

Теорема Колмогорова (усиленный закон больших чисел). Если $\xi_1,...,\xi_n$ – независимые, одинаково распределенные случайные величины с конечным математическим ожиданием $E(\xi_i) = a$, то выполняется соотношение

$$P(\frac{1}{n}\sum_{i=1}^{n}\xi_i \underset{n\to\infty}{\to} a) = 1.$$

В данной теореме утверждает, что среднее арифметическое случайных величин $\xi_1,...,\xi_n$ с указанными свойствами сходится почти наверное к среднему арифметическому их математических ожиданий. Это более сильное утверждение по сравнению с теоремой Хинчина.

Доказательства теорем Бернулли и Чебышева сравнительно простые, их можно найти в [14]. Доказательства теорем Хинчина и Колмогорова несколько сложнее и здесь не приводятся. Их можно найти в [3].

1.2 Генеральная совокупность. Простой случайный выбор. Случайная выборка. Вариационный ряд

Будем изучать статистику случайных величин, случайных векторов и случайных процессов. Ограничимся сначала случайной величиной ξ, имеющей функцию распределения $F_\xi(x)$. Отдельный эксперимент (испытание) над случайной величиной ξ будет заключаться в том, что мы зафиксируем (измерим и запишем) значение x величины ξ. Выполнив n таких экспериментов, получим n значений $x_1, x_2, ..., x_n$ случайной величины ξ. Все множество возможных значений случайная величина ξ называется в математической статистике *генеральной совокупностью*. Значения x_i, полученные в результате экспериментов (в результате выбора), называются *выборочными значениями*, а вся их совокупность $X = (x_1, x_2, ..., x_n)$ – *выборкой* объема или размера n из генеральной совокупности $F_\xi(x)$ или просто из распределения $F_\xi(x)$.

Если эксперимент организован таким образом, что вероятность быть выбранным для каждого элемента генеральной совокупности одинакова, и эксперименты независимы один от другого, то такой выбор называется простым случайным, а полученная при этом выборка – случайной. В дальнейшем будем считать, что наша выборка случайная.

Например, в коробке имеется 5000 деталей, и мы желаем сделать статистические выводы о размере этих деталей. Эти 5000 деталей и составляют генеральную совокупность. Отобрав и измерив часть деталей, мы получим выборку. Если каждая деталь выбирается наугад, причем после измерения она возвращается обратно в коробку, то наш выбор будет простым случайным. Это так называемый *выбор с возвратом*. Возврат в случае конечной генеральной совокупности необходим для того, чтобы обеспечить равную вероятность выбора для каждого элемента совокупности. Предположив, что в коробке бесконечно большое число

деталей, мы можем представить себе простой случайный выбор из бесконечной генеральной совокупности. Однако в этом случае выбранную деталь возвращать обратно не обязательно. Такой выбор называется *выбором без возврата*.

Простейшая обработка выборки заключается в ее сортировке, то есть в расположении выборочных значений в порядке их возрастания: $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$. Выборка, расположенная в порядке возрастания значений, называется *вариационным* или *простым статистическим рядом*. Элемент $x_{(i)}$ вариационного ряда называется i-й порядковой статистикой. Например, минимальное выборочное значений называется первой порядковой статистикой,

$$\min(x_1, x_2, ..., x_n) = x_{(1)},$$

максимальное – n-й порядковой статистикой,

$$\max(x_1, x_2, ..., x_n) = x_{(n)}.$$

Пример 1.1. Для изучения роста мужчин некоторого коллектива наугад отобраны 5 мужчин, рост которых в сантиметрах оказался равным $x_1 = 175$, $x_2 = 190$, $x_3 = 180$, $x_4 = 173$, $x_5 = 178$. Эти данные составляют выборку объемом $n = 5$. Вариационный ряд для данной выборки имеет вид $x_{(1)} = 173$, $x_{(2)} = 175$, $x_{(3)} = 178$, $x_{(4)} = 180$, $x_{(5)} = 190$.

1.3 Выборка как дискретная случайная величина и как случайный вектор. Статистика

Простейший взгляд на выборку $x_1, x_2, ..., x_n$ из распределения $F_\xi(x)$ состоит в том, что числа $x_1, x_2, ..., x_n$ считают возможными значениями некоторой дискретной случайной величины ξ^*, имеющими одинаковые вероятности $1/n$. Ряд распределения этой дискретной случайной величины ξ^* имеет вид таблицы 1.1.

Таблица 1.1

x_i	x_1	x_2	...	x_n
p_i	$1/n$	$1/n$	...	$1/n$

Другой взгляд на выборку состоит в следующем. Выборка $x_1, x_2, ..., x_n$ как набор n чисел представляет собой точку в n-мерном пространстве. Осуществив повторный выбор, мы получим новую точку, отличную от первой. Осуществив выбор большое число раз, получим множество точек в n-мерном пространстве. Эти соображения дают нам основание считать вектор $X = (x_1, x_2, ..., x_n)$ случайным. Каждая конкретная выборка является реализацией этого случайного вектора X. Для простого случайного выбора компоненты $x_1, x_2, ..., x_n$ вектора X независимы. Каждая компонента x_i вектора X имеет то же распределение, что и генеральная совокупность. Выборка X, рассматриваемая как случайный вектор, имеет распределение, определяемое формулами

$$f(X) = \prod_{i=1}^{n} f_\xi(x_i),\ F(X) = \prod_{i=1}^{n} F_\xi(x_i),$$

где $f_\xi(x)$, $F_\xi(x)$ – плотность вероятности и функция распределения генеральной совокупности. Обычно плотность вероятности генеральной совокупности зависит от одного или нескольких параметров $\bar{\theta} = (\theta_1, \theta_2, ..., \theta_m)$ и записывается в виде $f_\xi(x, \bar{\theta})$. Плотность вероятности выборки X, рассматриваемая как функция векторного параметра $\bar{\theta} = (\theta_1, \theta_2, ..., \theta_m)$, называется функцией правдоподобия и записывается в виде

$$L(X, \bar{\theta}) = \prod_{i=1}^{n} f_\xi(x_i, \bar{\theta}).$$

Любая функция выборочных значений $g(x_1, x_2, ..., x_n)$ называется статистикой. Если выборку считать случайным вектором, то статистика как функция случайного вектора будет случайной величиной. В этом случае можно говорить

о законе распределения статистики или о числовых характеристиках статистики. В частности, математическое ожидание любой статистики определяется выражением

$$E(g(x_1,...,x_n)) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} g(x_1...x_n) \prod_{i=1}^{n} f_{\xi}(x_i) \prod_{i=1}^{n} dx_i .$$

Примером статистики является среднее арифметическое выборочных значений $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, так что можно ставить вопрос о нахождении закона распределения этой статистики или ее математического ожидания, дисперсии, других числовых характеристик.

1.4 Свойства точечных оценок характеристик и параметров распределений

Характеристики распределений – это функция распределения, плотность вероятности, математическое ожидание, дисперсия, моменты, и др.

Параметры распределений – это величины, от которых зависит функция распределения или плотность вероятности и которые остаются постоянными для всех выборочных значений. Например, нормальное распределение $N(a,\sigma^2)$ имеет два параметра a и σ^2, экспоненциальное $E(\lambda)$ – один параметр λ, равномерное $u(a,b)$ – два параметра a и b.

В общем случае параметр распределения будем обозначать θ. Одной из задач математической статистики является получение оценок характеристик или параметров распределений по выборкам из этих распределений. Поскольку здесь отыскивается точки $\hat{\theta}$ в пространстве характеристик или параметров, наилучшим образом характеризующая истинное значение характеристики или параметра, то такая задача называется задачей получения точечных оценок.

Любая точечная оценка является функцией выборочных значений: $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$.

Оценка считается хорошей, если она обладает свойствами несмещенности, состоятельности, эффективности или хотя бы частью этих свойств.

Оценка $\hat{\theta}$ параметров θ называется *несмещенной,* если ее математическое ожидание совпадает с оцениваемым параметром:

$$E(\hat{\theta}) = \theta.$$

Несмещенная оценка – это оценка, точная в среднем, то есть не содержащая систематической (постоянной) ошибки.

Величина $b(\theta) = E(\hat{\theta}) - \theta$ называется смещением оценки $\hat{\theta}$. Смещение является функцией параметра θ.

Если смещение является линейной функцией параметра, $b(\theta) = \alpha + \beta\theta$, то оно может быть устранено. Новая оценка

$$\hat{\theta}_1 = \frac{\hat{\theta} - \alpha}{\beta + 1},$$

которая называется исправленной оценкой, является несмещенной. Действительно,

$$E(\hat{\theta}_1) = E(\frac{\hat{\theta} - \alpha}{\beta + 1}) = \frac{E(\hat{\theta}) - \alpha}{\beta + 1} =$$

$$= \frac{b(\hat{\theta}) + \theta - \alpha}{\beta + 1} = \frac{\alpha + \beta\theta - \alpha + \theta}{\beta + 1} = \frac{\theta(\beta + 1)}{\beta + 1} = \theta.$$

Если смещение некоторой оценки $b(\theta) \to 0$ при $n \to \infty$, то такая оценка называется асимптотически несмещенной.

Оценка $\hat{\theta}$ параметра θ называется состоятельной, если она сходится к параметру по вероятности, то есть если

$$P(|\hat{\theta} - \theta| > \varepsilon) \underset{n \to \infty}{\to} 0.$$

Состоятельная оценка – это оценка, которая становится все более точной по мере увеличения объема выборки n.

Теорема 1.1. Несмещенная оценка $\hat{\theta}$ является состоятельной, если ее дисперсия $D(\hat{\theta}) \underset{n\to\infty}{\to} 0$.

Доказательство основано на неравенстве Чебышева. Считая оценку случайной величиной, получим

$$P(|\hat{\theta} - \theta| > \varepsilon) \le \frac{D(\hat{\theta})}{\varepsilon^2}, \ \varepsilon > 0.$$

Так как наша оценка несмещенная, то есть $E(\hat{\theta}) = \theta$, то

$$P(|\hat{\theta} - \theta| > \varepsilon) \le \frac{D(\hat{\theta})}{\varepsilon^2} \underset{D(\hat{\theta})\to 0}{\to} 0.$$

Оценка $\hat{\theta}$ параметра θ называется строго состоятельной, если она сходится к параметру почти наверное: $P(\hat{\theta} \xrightarrow{п.н.} \theta) = 1$.

Оценка $\hat{\theta}$ параметра θ называется эффективной, если ее вариация

$$V(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

минимальна в сравнении с любыми другими оценками. Вариация оценки $V(\hat{\theta})$ – это математическое ожидание квадрата отклонения оценки от параметра. Ее не следует смешивать с дисперсией оценки $D(\hat{\theta}) = E((\hat{\theta} - E(\hat{\theta}))^2)$. Вариация совпадает с дисперсией только для несмещенных оценок. Для других оценок $V(\hat{\theta}) \ge D(\hat{\theta})$. Действительно,

$$V(\hat{\theta}) = E((\hat{\theta} - \theta)^2) = E((\hat{\theta} - E(\hat{\theta}) + b(\theta))^2) = D(\hat{\theta}) + b^2(\theta) \ge D(\hat{\theta}).$$

Итак, эффективная оценка имеет минимальный средний разброс относительно параметра по сравнению с любыми другими оценками.

Величина $e(\hat{\theta}) = \dfrac{E((\hat{\theta}_{эф} - \theta)^2)}{E((\hat{\theta} - \theta)^2)}$, где $\hat{\theta}_{эф}$ – эффективная оценка, называется эффективностью оценки $\hat{\theta}$. Всегда $0 \le e(\hat{\theta}) \le 1$. Оценка называется эффективной, если ее эффективность $e(\hat{\theta})$ равна 1.

Оценка $\hat{\theta}$ параметра θ называется асимптотически эффективной, если ее эффективность $e(\hat{\theta}) \underset{n\to\infty}{\to} 1$.

1.5 Неравенство Рао-Крамера

Для повышения эффективности точечной оценки необходимо уменьшать ее вариацию. Однако вариацию (и дисперсию) оценки нельзя уменьшить до нуля. Существует некоторая нижняя граница вариации оценки, которая определяется неравенством Рао-Крамера. Получим это неравенство.

Предположим, что плотность вероятности $f_\xi(x,\theta)$, параметр θ которой оценивается, дифференцируема по параметру. Предположим также, что границы области аргумента x, где функция плотности вероятности $f_\xi(x,\theta)$ отлична от нуля, не зависят от θ. Это условие выполняется, например, если $f_\xi(x,\theta) \ne 0$ на всей действительной прямой или для $x > 0$. Примером распределения, которое не удовлетворяет этому условию, является равномерное распределение вида

$$f_\xi(x,\theta) = \frac{1}{\theta},\ 0 < x < \theta.$$

Пусть $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$ – некоторая оценка параметра θ, полученная по выборке объема n из распределения $f_\xi(x,\theta)$. По определению смещения $b(\theta)$ оценки $\hat{\theta}$ имеем

$$E(\hat{\theta} - \theta - b(\theta)) = 0,$$

или

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\hat{\theta}-\theta-b(\theta))L(X,\theta)dX=0,$$

где $L(X,\theta)$ – функция правдоподобия, $X=(x_1,x_2,...,x_n)$ – выборка объема n (см. раздел 1.3). Дифференцируя обе части по θ (левую часть под интегралом), получим

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(-1-b'(\theta))L(X,\theta)dX+\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\hat{\theta}-\theta-b(\theta))\frac{\partial L(X,\theta)}{\partial\theta}dX=0.$$

Это равенство можно переписать в виде

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\hat{\theta}-\theta-b(\theta))\frac{\partial \ln L(X,\theta)}{\partial\theta}L(X,\theta)dX=1+b'(\theta),$$

что означает выполнение равенства

$$E\left((\hat{\theta}-\theta-b(\theta))\frac{\partial \ln L(X,\theta)}{\partial\theta}\right)=1+b'(\theta),$$

а также равенства

$$E^2\left((\hat{\theta}-\theta-b(\theta))\frac{\partial \ln L(X,\theta)}{\partial\theta}\right)=(1+b'(\theta))^2.$$

Применяя к левой части неравенство Шварца [14], получим

$$E\left((\hat{\theta}-\theta-b(\theta))^2\right)E\left(\left(\frac{\partial \ln L(X,\theta)}{\partial\theta}\right)^2\right)\geq(1+b'(\theta))^2.$$

Так как

$$E\left((\hat{\theta}-\theta-b(\theta))^2\right)=E\left((\hat{\theta}-E(\hat{\theta}))^2\right)=D(\hat{\theta}),$$

то мы получили неравенство Рао-Крамера

$$D(\hat{\theta})\geq\frac{(1+b'(\theta))^2}{E\left(\left(\frac{\partial \ln L(X,\theta)}{\partial\theta}\right)^2\right)}.$$

Учитывая существующее неравенство между дисперсией и вариацией оценки, неравенство Рао-Крамера можно записать в виде

$$V(\hat{\theta}) \geq D(\hat{\theta}) \geq \frac{(1+b'(\theta))^2}{E\left(\left(\frac{\partial \ln L(X,\theta)}{\partial \theta}\right)^2\right)}.$$

Неравенство Рао-Крамера дает нижнюю границу для дисперсии и вариации оценки, определяемую правой частью неравенства.

Неотрицательная величина

$$I(\theta) = E\left(\left(\frac{\partial \ln L(X,\theta)}{\partial \theta}\right)^2\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial \ln L(X,\theta)}{\partial \theta}\right)^2 L(X,\theta) dX$$

называется информацией о параметре θ (по Фишеру), содержащейся в выборке объема n. С учетом введенного понятия информации неравенство Рао-Крамера можно записать в виде

$$V(\hat{\theta}) \geq D(\hat{\theta}) \geq \frac{(1+b'(\theta))^2}{I(\theta)}.$$

Для несмещенной оценки $V(\hat{\theta}) = D(\hat{\theta})$, $b(\theta) = 0$, и неравенство Рао-Крамера приобретает простой вид

$$D(\hat{\theta}) \geq \frac{1}{I(\theta)},\ V(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

или

$$D(\hat{\theta}) I(\theta) \geq 1,\ V(\hat{\theta}) I(\theta) \geq 1.$$

Если вариация оценки $\hat{\theta}$ совпадает с нижней границей, определяемой неравенством Рао-Крамера, то эта оценка является эффективной, т.е. для эффективной оценки неравенство Рао-Крамера превращается в равенство. Из неравенства Рао-Крамера следует, что эффективность несмещенной оценки определяется выражением

$$e(\hat{\theta}) = \frac{1}{V(\hat{\theta}) I(\theta)}.$$

Обозначим

$$i(\theta) = E\left(\left(\frac{\partial \ln f_\xi(x,\theta)}{\partial\theta}\right)^2\right) = \int_{-\infty}^{\infty}\left(\frac{\partial \ln f_\xi(x,\theta)}{\partial\theta}\right)^2 f_\xi(x,\theta)dx,$$

где $f_\xi(x,\theta)$ – плотность вероятности генеральной совокупности, и назовем эту величину информацией о параметре θ (по Фишеру), содержащейся в одном выборочном значении. Учитывая, что $L(X,\theta) = \prod_{i=1}^{n} f_\xi(x_i,\theta)$, легко показать, что информация о параметре, содержащаяся в выборке объема n, в n раз больше информации, содержащейся в одном выборочном значении, т.е.

$$I(\theta) = ni(\theta).$$

Обобщением неравенства Рао-Крамера на случай векторного параметра $\bar{\theta} = (\theta_1, \theta_2, ..., \theta_m)$ и несмещенной оценки $\hat{\bar{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m)$ является следующее утверждение: для несмещенной оценки $\hat{\bar{\theta}}$ матрица $(D(\hat{\bar{\theta}}) - I^{-1}(\bar{\theta}))$ неотрицательно определенная, т.е.

$$(D(\hat{\bar{\theta}}) - I^{-1}(\bar{\theta})) \geq 0, \qquad (1.1)$$

где

$$D(\hat{\bar{\theta}}) = \left(d_{i,j}\right) = E\left((\hat{\theta}_i - E(\hat{\theta}_i))(\hat{\theta}_j - E(\hat{\theta}_j))\right),\ i, j = \overline{1,m},$$

– дисперсионная матрица вектора оценки $\hat{\bar{\theta}}$,

$$I(\bar{\theta}) = (I_{i,j}) = E\left(\left(\frac{\partial \ln L(X,\bar{\theta})}{\partial\theta_i}\right)\left(\frac{\partial \ln L(X,\bar{\theta})}{\partial\theta_j}\right)\right) = -E\left(\frac{\partial^2 \ln L(X,\bar{\theta})}{\partial\theta_i\partial\theta_j}\right),\ i, j = \overline{1,m},$$

– информационная матрица Фишера.

Неравенство Рао-Крамера (1.1) можно записать также в виде

$$\det D(\hat{\bar{\theta}})\, \det I(\bar{\theta})) \geq 1,$$

где *det* означает определитель матрицы. Величину

$$e(\hat{\bar{\theta}}) = \left(\det D(\hat{\bar{\theta}})\, \det I(\bar{\theta})\right)^{-1} \leq 1$$

можно назвать эффективностью векторной оценки $\hat{\hat{\theta}}$, а оценку, для которой $e(\hat{\hat{\theta}}) = 1$ – эффективной.

1.6 Эмпирическая функция распределения

Определение. Эмпирической или выборочной функцией распределения $F_{\xi}^{*}(x)$ называется функцией распределения дискретной случайной величины ξ^{*}, определенной в разделе 1.3, то есть функция распределения выборки, рассматриваемой как дискретная случайная величина с равновозможными значениями:

$$F_{\xi}^{*}(x) = F_{\xi^{*}}(x).$$

В соответствии с этим определением эмпирическая функция распределения определяется формулой

$$F_{\xi}^{*}(x) = \frac{m}{n},$$

где m – количество выборочных значений, удовлетворяющих условию $x_i < x$.

Эмпирическую функцию распределения удобно строить с использованием порядковых статистик $x_{(i)}$, $i = \overline{1,n}$. В этом случае она определяется формулой

$$F_{\xi}^{*}(x) = \begin{cases} 0, & \text{если } x < x_{(1)}, \\ \dfrac{i}{n}, & \text{если } x_{(i)} \le x < x_{(i+1)},\ i = \overline{1,n-1}. \\ 1, & \text{если } x \ge x_{(n)}. \end{cases} \qquad (1.2)$$

Эмпирическая функция распределения $F_{\xi}^{*}(x)$ является ступенчатой функцией (рис. 1.1), поскольку это функция распределения дискретной случайной величины. Она является оценкой генеральной функции распределения $F_{\xi}(x)$.

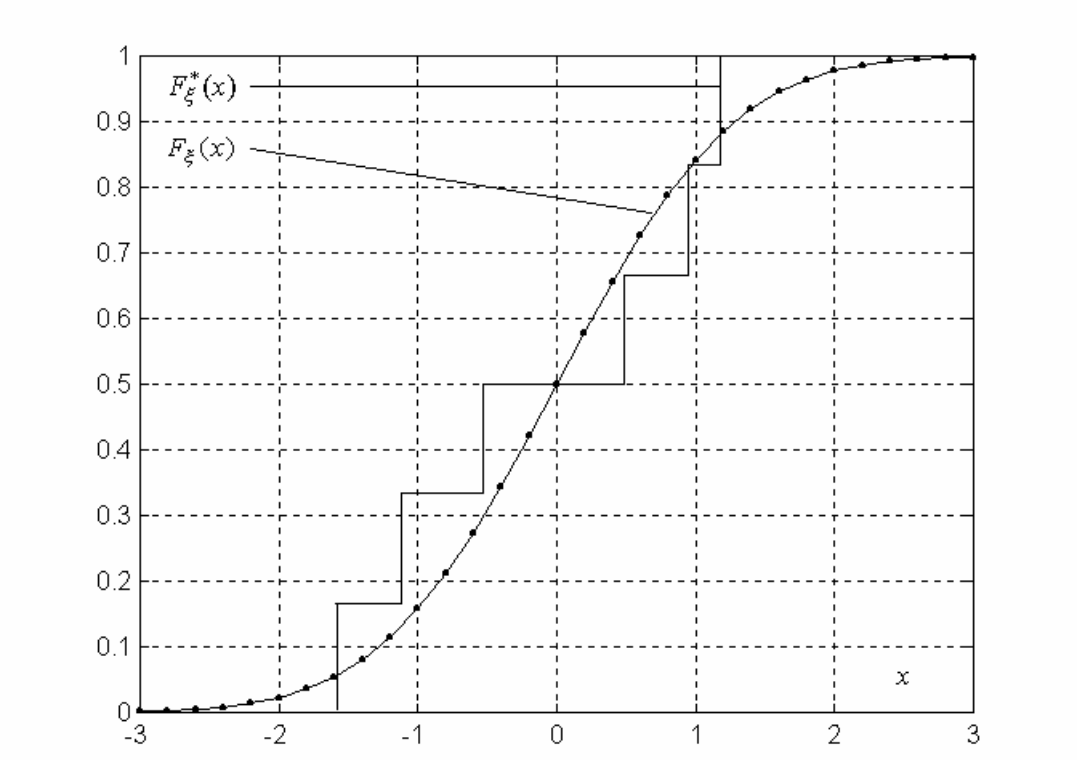

Рис. 1.1. Эмпирическая $F_\xi^*(x)$ и генеральная $F_\xi(x)$ функции распределения

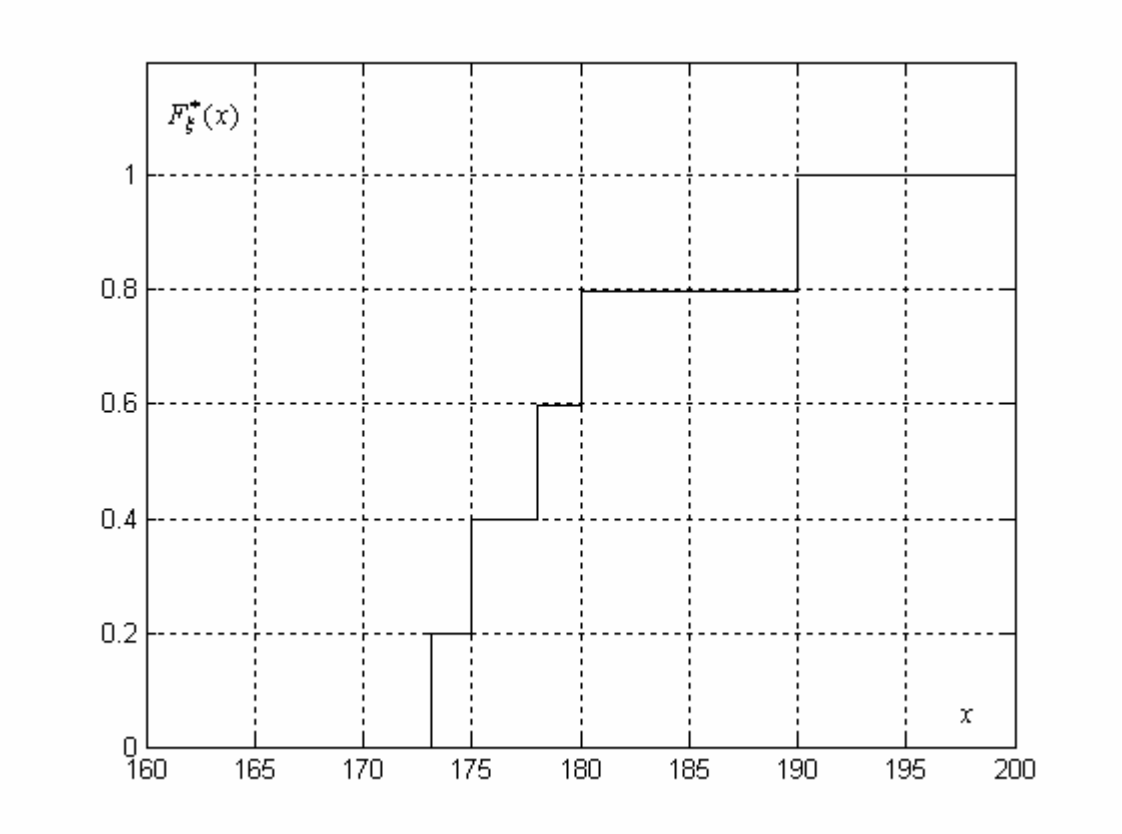

Рис. 1.2. Эмпирическая функция распределения для примера 1.2 раздела 1.6

Пример 1.2. Воспользуемся данными о росте мужчин из примера 1.1 раздела 1.1 для построения эмпирической функции распределения. Поскольку вариационный ряд имеет вид $x_{(1)} = 173$, $x_{(2)} = 175$, $x_{(3)} = 178$, $x_{(4)} = 180$, $x_{(5)} = 190$, то по формуле (1.2) получим функцию, представленную на рис. 1.2. Эта функция является оценкой функции распределении роста мужчин в рассматриваемом коллективе.

Выясним свойства эмпирической функции распределения $F_\xi^*(x)$ как оценки теоретической (генеральной) функции распределения $F_\xi(x)$.

Теорема 1.2. Эмпирическая функция распределения $F_\xi^*(x)$ является состоятельной оценкой теоретической функции распределения $F_\xi(x)$, то есть для всех x

$$P(|F_\xi^*(x) - F_\xi(x)| > \varepsilon) \underset{n\to\infty}{\to} 0.$$

Доказательство. Рассмотрим событие $A = (\xi < x)$. Тогда $P(A) = p = F_\xi(x)$. Поскольку $F_\xi^*(x) = m/n$, где m – количество выборочных значений, которые меньше x, m/n – частота события A, то по теореме Бернулли получаем

$$P(|\frac{m}{n} - p| > \varepsilon) \underset{n\to\infty}{\to} 0,$$

или

$$P(|F_\xi^*(x) - F_\xi(x)| > \varepsilon) \underset{n\to\infty}{\to} 0.$$

*Теорема 1.*3 (*Гливенко-Кантелли).*

$$P\left(\sup_x |F_\xi^*(x) - F_\xi(x)| \underset{n\to\infty}{\to} 0\right) = 1,$$

где sup означает верхнюю грань множества. Эта теорема представляет собой более сильное утверждение по сравнению с предыдущей теоремой о состоятельности. Доказательство этой теоремы выходит за рамки данного пособия.

Из приведенных теорем заключаем, что эмпирическая функция распределения является достаточно хорошей оценкой генеральной функции распределения.

1.7 Гистограмма

Гистограмма – это фигура $f_{\xi}^{*}(x)$, аппроксимирующая генеральную плотность вероятности $f_{\xi}(x)$. Для ее построения интервал выборочных значений $[x_{(1)}, x_{(n)}]$ делится на l непересекающихся интервалов длиной Δ_i, $i = \overline{1, l}$, и подсчитывается количество выборочных значений m_i, попавших в i-й интервал. Если на каждом частичном интервале как на основании построить прямоугольник высотой $h_i = \dfrac{m_i}{n\Delta_i}$, то мы получим фигуру, которая называется гистограммой (рис. 1.3).

Существуют два способа построения гистрограммы.

1. Равноинтервальный способ. Выбирают количество интервалов l, а длину Δ каждого интервала берут одной и той же, вычисляемой по формуле

$$\Delta = \frac{x_{(n)} - x_{(1)}}{l}.$$

2. Равновероятный способ. Выбирают количество выборочных значений m, попавших в каждый интервал. Объем выборки должен быть кратен m. Тогда число интервалов определяется по формуле $l = n / m$, и интервалы будут следующими: $[x_{(1)}, x_{(m)}]$, $[x_{(m)}, x_{(2m)}]$, …, $[x_{((l-1)m)}, x_{(lm)}]$, где $x_{(i)}$ – i-я порядковая статистика. При этом способе интервалы имеют различную длину, а границы интервалов попадают на выборочные значения. Принято считать, что граничное значение делится поровну между двумя интервалами, то есть $0{,}5$ значения попадает в левый интервал и $0{,}5$ – в правый. Понятно, что при этом в крайний ле-

вый интервал попадает $m - 0{,}5$ значений, в крайний правый – $m + 0{,}5$ значений, а в средние интервалы – по m значений.

При объеме выборки $n = 100 - 200$ рекомендуется брать от 10 до 20 интервалов, так чтобы в каждый интервал попадало примерно 10 значений.

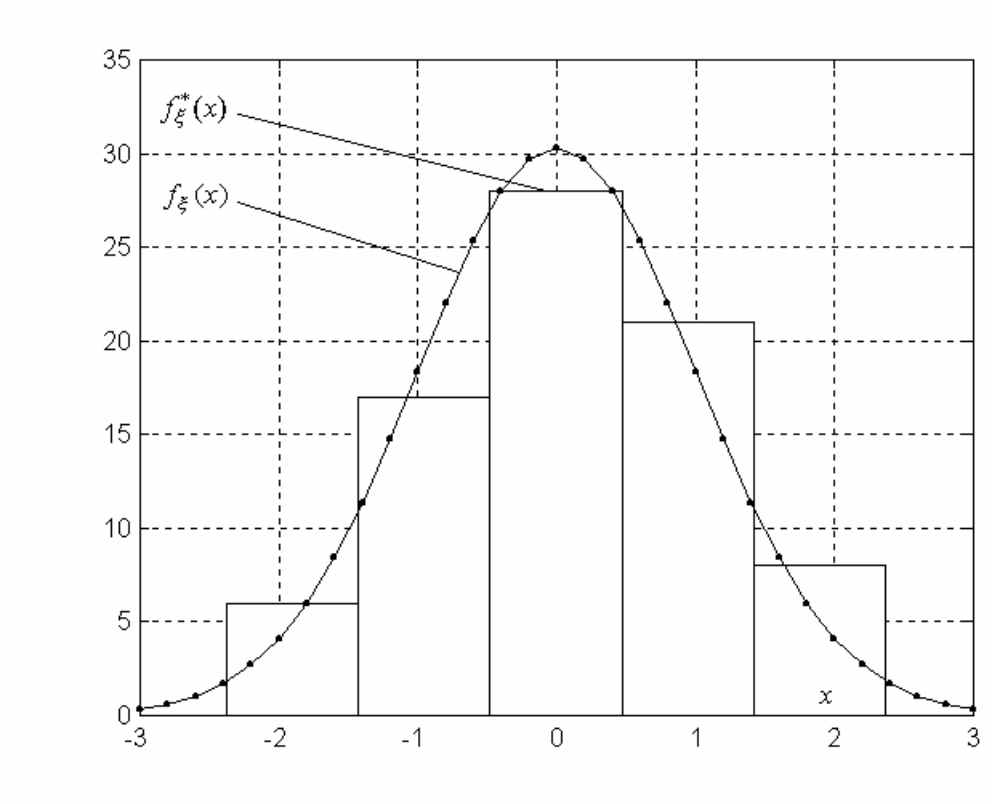

Рис. 1.3. Гистограмма и соответствующая ей теоретическая (генеральная) плотность вероятности

Рассмотрим свойства гистограммы $f^*_\xi(x)$ как оценки генеральной плотности вероятности $f_\xi(x)$.

Теорема 1.4. Пусть $z_1, z_2, ..., z_l$ – точки разбиения области возможных значений случайной величины ξ на l интервалов, m_i – количество выборочных значений, попавших в i-й интервал (z_i, z_{i+1}), n – объем выборки. Если максимальный из интервалов разбиения стремится к нулю при увеличении объема

выборки, то гистограмма $f_\xi^*(x)$ является состоятельной оценкой генеральной плотности вероятности $f_\xi(x)$, то есть для любого $\varepsilon > 0$

$$P(|f_\xi^*(x) - f_\xi(x)| > \varepsilon) \underset{n\to\infty}{\to} 0.$$

Доказательство. Для эмпирической функции распределения можно записать два условия:

$$P(|F_\xi^*(z_{i+1}) - F_\xi(z_{i+1})| > \varepsilon_1) \underset{n\to\infty}{\to} 0,$$

$$P(|F_\xi^*(z_i) - F_\xi(z_i)| > \varepsilon_2) \underset{n\to\infty}{\to} 0.$$

Тогда $P(|F_\xi^*(z_{i+1}) - F_\xi^*(z_i) - F_\xi(z_{i+1}) + F_\xi(z_i)| > \varepsilon_3) \underset{n\to\infty}{\to} 0$, $\varepsilon_3 = \varepsilon_1 + \varepsilon_2$. Выбрав $\varepsilon_3 = \varepsilon(z_{i+1} - z_i)$, $\varepsilon > 0$, получим

$$P\left(\left|\frac{F_\xi^*(z_{i+1}) - F_\xi^*(z_i)}{z_{i+1} - z_i} - \frac{F_\xi(z_{i+1}) - F_\xi(z_i)}{z_{i+1} - z_i}\right| > \varepsilon\right) \underset{n\to\infty}{\to} 0.$$

Если $z_{i+1} \to z_i$, то

$$\frac{F_\xi(z_{i+1}) - F_\xi(z_i)}{z_{i+1} - z_i} \to F'(z_i) = f_\xi(z_i),$$

$$\frac{F_\xi^*(z_{i+1}) - F_\xi^*(z_i)}{z_{i+1} - z_i} = \frac{m}{n(z_{i+1} - z_i)} = f_\xi^*(z_i),$$

$$P(|f_\xi^*(z_i) - f_\xi(z_i)| > \varepsilon) \underset{\substack{n\to\infty \\ z_{i+1}\to z_i}}{\to} 0.$$

Теорема доказана.

Замечание. Если интервалы разбиения не уменьшаются по мере увеличения объема выборки, то гистограмма не будет состоятельной оценкой плотности вероятности.

1.8 Выборочные числовые характеристики

Выборочные числовые характеристики – это числовые характеристики случайной величины ξ^*, рассмотренной в разделе 1.3, то есть выборки, рассматриваемой как дискретная случайная величина с равновозможными значениями. Отсюда выборочные числовые характеристики есть не что иное, как числовые характеристики эмпирического распределения $F^*_{\xi}(x)$. Приведем некоторые из них.

Выборочное среднее $\bar{x}$ – это среднее значение случайной величины ξ^*:

$$\bar{x} = E(\xi^*) = \sum_{t=1}^{n} x_i p_i = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

Мы видим, что выборочное среднее – это среднее арифметическое выборочных значений.

Выборочная дисперсия $\bar{s}^2$ – это дисперсия случайной величины ξ^*:

$$\bar{s}^2 = D(\xi^*) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Для расчета выборочной дисперсии более удобна формула

$$\bar{s}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 ,$$

которая получается простым преобразованием предыдущей формулы.

Выборочные начальные моменты $\bar{\nu}_k$ – это начальные моменты дискретной случайной величины ξ^*:

$$\bar{\nu}_k = E((\xi^*)^k) = \frac{1}{n}\sum_{i=1}^{n} x_i^k \ , \ k = 0,1,2,... \ .$$

Выборочные центральные моменты $\bar{\mu}_k$ – это центральные моменты дискретной случайной величины ξ^*:

$$\bar{\mu}_k = E((\xi^* - E(\xi^*))^k) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k \ , \ k = 0,1,2,... \ .$$

Если выборка извлекается из двухмерного распределения, то есть эксперимент осуществляется над случайным вектором $\overline{\xi} = (\xi_1, \xi_2)$, то эта выборка имеет вид совокупности пар чисел:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n).$$

В этом случае выборочные средние для каждой компоненты определяются формулами

$$\bar{x} = E(\xi_1^*) = \frac{1}{n}\sum_{i=1}^{n} x_i\ , \qquad \bar{y} = E(\xi_2^*) = \frac{1}{n}\sum_{i=1}^{n} y_i\ .$$

Выборочные дисперсии для каждой компоненты будут равны

$$\bar{s}_x^2 = D(\xi_1^*) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\,, \qquad \bar{s}_y^2 = D(\xi_2^*) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\,.$$

Определим также выборочный коэффициент ковариации

$$\bar{s}_{xy} = E\Big(\big(\xi_1^* - E(\xi_1^*)\big)\big(\xi_2^* - E(\xi_2^*)\big)\Big) = \operatorname{cov}(\xi_1^*, \xi_2^*) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

и выборочный коэффициент корреляции

$$\bar{r}_{xy} = \frac{\operatorname{cov}(\xi_1^*, \xi_2^*)}{\sqrt{D(\xi_1^*)D(\xi_2^*)}} = \frac{\bar{s}_{xy}}{\bar{s}_x \bar{s}_y}.$$

Все выборочные характеристики являются случайными величинами (статистиками), имеющими свои законы распределения и числовые характеристики. Приведенные выборочные характеристики являются оценками соответствующих генеральных или теоретических характеристик. Для анализа свойств этих оценок полезна следующая теорема.

Теорема 1.5. Если $x_1, x_2, ..., x_n$ – выборка из распределения $F_\xi(x)$ с конечным математическим ожиданием $E(\xi) = a$ и $q(\xi)$ – некоторая измеримая функция, то для любого $\varepsilon > 0$ выполняются соотношения

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} q(x_i) - E(q(\xi))\Big| > \varepsilon\Big) \underset{n\to\infty}{\to} 0\,,$$

$$P\left(\frac{1}{n}\sum_{i=1}^{n} q(x_i) \underset{n\to\infty}{\to} E(q(\xi))\right) = 1.$$

Для доказательства обозначим $\eta = q(\xi)$, $E(\eta) = E(q(\xi))$. Величины $\eta_i = q(x_i)$ независимые и одинаково распределенные. На основании теорем Хинчина и Колмогорова (раздел 1.1) можно записать соотношения

$$P(|\frac{1}{n}\sum_{i=1}^{n}\eta_i - E(\eta)| > \varepsilon) \underset{n\to\infty}{\to} 0,$$

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\eta_i \underset{n\to\infty}{\to} E(\eta)\right) = 1,$$

которые завершают доказательство.

Следствие. Выборочные начальные и центральные моменты $\overline{\nu}_k$, $\overline{\mu}_k$ является состоятельными и строго состоятельными оценками соответствующих теоретических моментов $\nu_k = E(\xi^k)$, $\mu_k = E((\xi - E(\xi))^k)$. Действительно, для начальных моментов $q(x_i) = x_i^k$, а для центральных моментов $q(x_i) = (x_i - \overline{x})^k$.

Можно также показать, что выборочный коэффициент ковариации $\overline{s}_{xy}$ и выборочный коэффициент корреляции $\overline{r}_{xy}$ являются состоятельными и даже строго состоятельными оценками соответствующих теоретических характеристик $R_{\xi\eta} = cov(\xi,\eta)$, $r_{\xi\eta} = \dfrac{cov(\xi,\eta)}{\sqrt{D(\xi)D(\eta)}}$.

1.9 Свойства выборочного среднего и выборочной дисперсии

Рассмотрим более подробно свойства выборочного среднего $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ и выборочной дисперсии $\bar{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Они, как известно, являются оценками математического ожидания a и дисперсии σ^2 генеральной совокупности и, по изложенному выше, оценками состоятельными и даже строго состоятельными. Исследуем эти оценки на несмещенность и эффективность. Начнем с выборочного среднего.

$$E(\bar{x}) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} a = a.$$

Видим, что $\bar{x}$ – несмещенная оценка. Найдем дисперсию выборочного среднего:

$$D(\bar{x}) = D(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n^2}\sum_{i=1}^{n} D(x_i) = \frac{\sigma^2}{n}.$$

Мы видим, что дисперсия выборочного среднего в n раз меньше генеральной дисперсии. Если выборка извлечена из нормального распределения $N(a, \sigma^2)$, то можно показать, что $\bar{x}$ – эффективная оценка.

Найдем далее среднее значение выборочной дисперсии

$$E(\bar{s}^2) = E(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2).$$

Обозначим $x_i - a = \overset{\circ}{x}_i$. Получим

$$E(\bar{s}^2) = E(\frac{1}{n}\sum_{i=1}^{n}[(x_i - a) - (\frac{1}{n}\sum_{k=1}^{n} x_k - a)]^2) = E(\frac{1}{n}\sum_{i=1}^{n}(\overset{\circ}{x}_i - \frac{1}{n}\sum_{k=1}^{n}\overset{\circ}{x}_k)^2) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}[E(\overset{\circ}{x}_i^2) - \frac{2}{n}E(\sum_{j=1}^{n}\overset{\circ}{x}_i\,\overset{\circ}{x}_j) + \frac{1}{n^2}E(\sum_{k=1}^{n}\sum_{l=1}^{n}\overset{\circ}{x}_k\,\overset{\circ}{x}_l)] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 - \frac{2}{n}\sigma^2 + \frac{1}{n}\sigma^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n\sigma^2 - 2\sigma^2 + \sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Итак, $\bar{s}^2$ – смещенная оценка. Ее смещение

$$b_{\bar{s}^2}(\sigma^2) = E(\bar{s}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \frac{n}{n}\sigma^2 = -\frac{1}{n}\sigma^2.$$

Поскольку $b_{\bar{s}^2}(\sigma^2) \underset{n\to\infty}{\to} 0$, то оценка $\bar{s}^2$ – асимптотически несмещенная. Так как смещение – линейная функция вида $b_{\bar{s}^2}(\sigma^2) = \alpha + \beta\sigma^2$ с $\alpha = 0$, $\beta = -\frac{1}{n}$, то смещение можно устранить. Новая оценка

$$s^2 = \frac{\bar{s}^2}{1-\frac{1}{n}} = \frac{n}{n-1}\bar{s}^2$$

является несмещенной. Несмещенная оценка дисперсии имеет следующее выражение:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

В отличие от выборочной дисперсии здесь деление суммы производится не на весь объем выборки n, а на $(n-1)$. Эту новую оценку s^2 будем называть несмещенной или исправленной выборочной дисперсией. Она обычно и применяется на практике.

Если выборка извлечена из нормального распределения, то можно показать, что $\bar{s}^2$ и s^2 будут асимптотически эффективными оценками генеральной дисперсии σ^2.

Возможен также случай, когда среднее значение генеральной совокупности a известно, и необходимо оценить дисперсию генеральной совокупности σ^2. В этом случае используется оценка

$$\bar{s}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2,$$

которую мы будем называть выборочной дисперсией при известном математическом ожидании. Легко показать, что эта оценка несмещенная и состоятельная. Сравнение выражений для $\bar{s}^2$ и $\bar{s}_0^2$ показывает, что смещение оценки $\bar{s}^2$ возни-

кает из-за того, что при ее расчете вместо неизвестного параметра a используется оценка $\bar{x}$.

1.10 Порядковые статистики

Пусть имеется выборка $x_1,...,x_n$ из некоторого распределения $F_\xi(x)$. Расположив эти значения в порядке возрастания $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, получим вариационный ряд $x_{(1)}, x_{(2)},..., x_{(n)}$. Элемент $x_{(m)}$ вариационного ряда называется m-й порядковой статистикой. Известно, что выборочное значение x_i можно считать случайной величиной, имеющей то же распределение, что и генеральная совокупность $F_\xi(x)$ ($f_\xi(x)$). Порядковую статистику $x_{(m)}$ также можно считать случайной величиной, поскольку от одной выборки к другой она может принимать различные значения. Однако ее распределение отличается от распределения генеральной совокупности. Порядковую статистику как случайную величину будем обозначать $X_{(m)}$, в то время как под $x_{(m)}$ будем понимать возможное значение случайной величины $X_{(m)}$, которое она приняла в процессе извлечения выборки. Получим распределение порядковых статистик. Будем считать, что $F_\xi(x), f_\xi(x)$ непрерывны. В этих условиях вероятность нахождения порядковой статистики $X_{(m)}$ в окрестности точки $x_{(m)}$ определяется выражением

$$P(x_{(m)} \le X_{(m)} < x_{(m)} + dx) = f_{n,m}(x_{(m)})dx, \qquad (1.3)$$

где $f_{n,m}(x)$ – плотность вероятности статистики $X_{(m)}$ при объёме выборки n. Будем читать, что процесс формирования выборки $x_1,...,x_n$ – это независимые испытания Бернулли, в каждом из которых успехом считается появление события $A = (\xi < x_{(m)})$, а неудачей – события $\overline{A} = (\xi \ge x_{(m)})$. Тогда вероятность успеха в одном испытании Бернулли

$$P(A) = p = P(\xi < x_{(m)}) = F_\xi(x_{(m)}),$$

а вероятность неудачи

$$p(\overline{A}) = q = 1 - p = P(\xi \geq x_{(m)}) = 1 - F_\xi(x_{(m)}).$$

Для фиксированного m число успехов равно $(m-1)$, так как $(m-1)$ порядковых статистик расположены левее статистики $x_{(m)}$, а число неудач – $(n-m)$, так как $(n-m)$ статистик расположены правее $x_{(m)}$. Запишем вероятность следующего сложного события: $(m-1)$ элементов выборки находятся слева от $x_{(m)}$, $(n-m)$ – справа и случайная величина ξ примет значение в окрестности $x_{(m)}$. По теореме умножения вероятностей эта вероятность будет равна

$$C_{n-1}^{m-1}[F_\xi(x_{(m)})]^{m-1}[1 - F_\xi(x_{(m)})]^{n-m} f_\xi(x_{(m)})dx. \qquad (1.4)$$

Вероятность (1.4) и есть вероятность, записанная в левой части формулы (1.3), то есть мы можем записать

$$f_{n,m}(x_{(m)}) = C_{n-1}^{m-1}[F_\xi(x_{(m)})]^{m-1}[1 - F_\xi(x_{(m)})]^{n-m} f_\xi(x_{(m)}).$$

Так как $x_{(m)}$ – произвольная переменная, то её можно заменить на x:

$$f_{n,m}(x) = C_{n-1}^{m-1}[F_\xi(x)]^{m-1}[1 - F_\xi(x)]^{n-m} f_\xi(x). \qquad (1.5)$$

Последнее выражение и есть плотность вероятности m-й порядковой статистики $X_{(m)}$ при объеме выборки n. Мы видим, что плотность вероятности порядковой статистики выражается через плотность вероятности $f_\xi(x)$ и функцию распределения $F_\xi(x)$ генеральной совокупности.

Представляет интерес получение плотности вероятности порядковых статистик для наиболее распространенного нормального распределения. Однако удобное аналитическое выражение получить в этом случае не представляется возможным, поскольку функция распределении в этом случае не представляется в элементарных функциях.

Начальный и центральный моменты k-го порядка m-й порядковой статистики определяется формулами

$$E(X_{(m)}^k) = \int_{-\infty}^{\infty} x^k f_{n,m}(x)dx,\ k = 0,1,2,...,$$

$$E((X_{(m)} - E(X_{(m)}))^k) = \int_{-\infty}^{\infty} (x - E(X_{(m)}))^k f_{n,m}(x)dx.$$

При $n = 3$ можно получим следующие формулы для плотностей вероятности порядковых статистик: плотность вероятности минимального выборочного значения

$$f_{3,1}(x) = [1 - F_\xi(x)]^2 f_\xi(x),$$

плотность вероятности выборочной медианы

$$f_{3,2}(x) = 2F_\xi(x)[1 - F_\xi(x)]f_\xi(x),$$

плотность вероятности максимального выборочного значения

$$f_{3,3}(x) = F_\xi{}^2(x) f_\xi(x).$$

2 МЕТОДЫ НАХОЖДЕНИЯ ТОЧЕЧНЫХ ОЦЕНОК ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ

Задача получения точечных оценок параметров распределений является одной из центральных в математической статистике. Она формулируется следующим образом. Известна плотность вероятности генеральной совокупности $f_\xi(x,\overline{\theta})$ с точностью до вектора параметров $\overline{\theta}=(\theta_1,...,\theta_m)$. Требуется по выборке $x_1,...,x_n$ из этой совокупности найти оценку $\overline{\hat{\theta}}=(\hat{\theta}_1,...,\hat{\theta}_m)$ параметра $\overline{\theta}$.

Перейдем к изложению методов решения этой задачи.

2.1 Метод моментов

Метод моментов заключается в следующем. Находим m начальных теоретических моментов

$$\nu_j = E(\xi^j) = \int_{-\infty}^{\infty} x^j f_\xi(x,\overline{\theta})dx,\ j=\overline{1,m}.$$

Из этой формулы видно, что теоретические моменты являются функциями неизвестных параметров, то есть $\nu_j=\nu_j(\overline{\theta})$. Далее находим m выборочных начальных моментов

$$\overline{\nu}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j\ ,\ j=\overline{1,m}.$$

Приравнивая соответствующие теоретические и выборочные моменты, получим систему m уравнений

$$\nu_j(\overline{\theta}) = \frac{1}{n}\sum_{i=1}^{n} x_i^j = \overline{\nu}_j \quad j=\overline{1,m}.$$

Решая эту систему, получаем оценки $\overline{\hat{\theta}}=(\hat{\theta}_1,...,\hat{\theta}_m)$.

Достоинством метода является его простота. Недостатком является то, что в общем случае не доказано, что такие оценки обладают какими-либо хорошими свойствами. Этот вопрос придется решать отдельно.

Метод моментов рекомендуется применять для получения оценок небольшого числа параметров (порядка двух-трех).

Пример 2.1. Найти оценки параметров a и σ^2 нормальной генеральной совокупности $N(a,\sigma^2)$ методом моментов.

Решение. Теоретические моменты равны $\nu_1 = E(\xi) = a$, $\mu_2 = D(\xi) = \sigma^2$. Приравнивая их к соответствующим выборочным моментам, получим в качестве оценок известные нам выборочное среднее и выборочную дисперсию:

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

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \bar{s}^2.$$

Пример 2.2. Найти оценки параметров a, b равномерного в (a,b) распределения методом моментов.

Решение. Теоретические моменты равномерного в (a,b) распределения равны $\nu_1 = E(\xi) = \frac{a+b}{2}$, $\mu_2 = D(\xi) = \frac{(b-a)^2}{12}$. Приравнивая их к соответствующим выборочным моментам, получим систему уравнений

$$\frac{a+b}{2} = \bar{x},$$

$$\frac{(b-a)^2}{12} = \bar{s}^2.$$

Переписав эту систему в виде

$$b + a = 2\bar{x},$$

$$b - a = 2\sqrt{3}\,\bar{s},$$

получаем оценки

$$\hat{a} = \bar{x} - \sqrt{3}\,\bar{s}\,,\ \hat{b} = \bar{x} + \sqrt{3}\,\bar{s}\,.$$

2.2 Метод максимума правдоподобия

Это основной метод нахождения оценок. Он использует понятие функции правдоподобия. Функцией правдоподобия называется совместная плотность вероятности выборочных значений, рассматриваемых как случайные величины. Функция правдоподобия зависит как от переменных $x_1, x_2, ..., x_n$, так и от неизвестных параметров $\theta_1, ..., \theta_m$. Обычно она обозначается зависящей только от неизвестных параметров в виде $L(\theta_1, ..., \theta_n)$. Функция правдоподобия рассчитывается по формуле

$$L(\theta_1, ..., \theta_m) = \prod_{i=1}^{n} f_\xi(x_i, \theta_1, ...\theta_m)\,, \tag{2.1}$$

где $f_\xi(x, \theta_1, ...\theta_m)$ – плотность вероятности генеральной совокупности. Метод максимума правдоподобия заключается в том, что оценки отыскиваются из условия максимума функции правдоподобия:

$$L(\theta_1, ...\theta_m) \to \max_{\theta_1, ..., \theta_m}\,. \tag{2.2}$$

Полученные при этом оценки называются максимально правдоподобными, или МП-оценками. Чтобы найти МП-оценки, необходимо приравнять частные производные функции правдоподобия к нулю и решить полученную систему уравнений:

$$\frac{\partial}{\partial \theta_j} L(\theta_1, ..., \theta_m) = 0\,,\ j = \overline{1, m}\,.$$

Часто эта система упрощается с помощью следующего приема. Поскольку любая функция и ее логарифм достигают экстремума на одних и тех же значениях аргументов, то часто максимизируют не функцию правдоподобия, а ее натуральный логарифм, то есть логарифмическую функцию правдоподобия

$\ln L(\theta_1,...,\theta_m)$. В этом случае для получения оценок необходимо решать следующую систему уравнений:

$$\frac{\partial}{\partial\theta_j}\ln L(\theta_1,...,\theta_m)=0,\ \ j=\overline{1,m}.$$

Если учесть, что для простого случайного выбора функция правдоподобия представляется в виде произведения, то ее логарифм представляется в виде суммы

$$\ln L(\theta_1,...,\theta_m)=\sum_{i=1}^{n} f_\xi(x_i,\theta_1,...,\theta_m),$$

и система уравнений преобразуется к виду

$$\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}\ln f_\xi(x_i,\theta_1,...,\theta_m)=0,\ \ j=\overline{1,m}. \qquad (2.3)$$

Достоинством метода является то, что МП-оценки имеют хорошие свойства. Доказано, что они состоятельны, асимптотически несмещены и асимптотически эффективны. Вместе с тем система уравнений для получения оценок часто оказывается нелинейной, что не позволяет получать оценки в конечном виде.

Пример 2.3. Найти оценки параметров a и σ^2 нормальной генеральной совокупности $N(a,\sigma^2)$ методом максимума правдоподобия.

Решение. Нам известна плотность вероятности генеральной совокупности с точностью до двух параметров a, σ^2:

$$f_\xi(x,a,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Будем максимизировать логарифмическую функцию правдоподобия, для чего найдем

$$\ln f_\xi(x_i,a,\sigma^2)=\ln(\frac{1}{\sqrt{2\pi}})-\frac{1}{2}\ln\sigma^2-\frac{(x_i-a)^2}{2\sigma^2}.$$

Найдем частные производные по оцениваемым параметрам

$$\frac{\partial}{\partial a}(\ln f_\xi(x_i, a, \sigma^2) = \frac{x_i - a}{\sigma^2},$$

$$\frac{\partial}{\partial \sigma^2}\ln f_\xi(x_i, a, \sigma^2) = -\frac{1}{2\sigma^2} + \frac{(x_i - a)^2}{2\sigma^4}.$$

Для получения оценок необходимо решать систему уравнений (2.3), которая имеет вид

$$\begin{cases} \sum_{i=1}^{n} \frac{x_i - a}{\sigma^2} = 0, \\ \sum_{i=1}^{n} \frac{(x_i - a)^2 - \sigma^2}{2\sigma^4} = 0. \end{cases}$$

Из первого уравнения находим

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

Подставляя $\bar{x}$ вместо a во второе уравнение, получим

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \bar{s}^2.$$

Мы видим, что максимально правдоподобными оценками параметров a и σ^2 нормального распределения являются известные нам выборочное среднее $\bar{x}$ и выборочная дисперсия $\bar{s}^2$.

Обратим внимание на структуру оценки параметра a:

$$\hat{a} = \sum_{i=1}^{n} \frac{1}{n} x_i\,.$$

Это линейная функция выборочных значений, взятых с одинаковыми весовыми коэффициентами $\frac{1}{n}$. Такая структура оценки вполне понятна, поскольку все выборочные значения имеют одну и ту же дисперсию, и нет оснований присваивать им различные весовые коэффициенты.

Пример 2.4. Найти оценки параметров a, b равномерного в (a,b) распределения методом максимума правдоподобия.

Решение. Поскольку плотность вероятности равномерного в (a,b) распределения имеет вид

$$f_\xi(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b,\\ 0, & x\le a, x\ge b,\end{cases}$$

то функция правдоподобия равна

$$L(a,b)=\begin{cases}\prod\limits_{i=1}^{n}\dfrac{1}{b-a}, & a<x_1,x_2,...,x_n<b,\\ 0, & \textit{иначе.}\end{cases}$$

Поскольку $b>a$, и функция правдоподобия возрастает с возрастанием b и убыванием a, то с учетом ограничений в выражении функции правдоподобия получаем следующие оценки:

$$\widehat{a}=x_{(1)}=min(x_1,x_2,...,x_n),\ \widehat{b}=x_{(n)}=max(x_1,x_2,...,x_n).$$

Эти оценки отличаются от оценок, полученных в примере 2.2 методом моментов.

2.3 Оценивание параметров по результатам неравноточных измерений

В ряде практических задач приходится решать следующую задачу. Некоторая величина (например, толщина металлической пластины) измеряется приборами, имеющими различную точность измерения. Требуется по результатам этих измерений определить истинное значение величины (толщины пластины).

Поскольку чаще всего ошибки измерений приборов считают распределенными по нормальному закону, то данную задачу можно сформулировать следующим образом. Параметр a является средним значением n нормальных генеральных совокупностей с плотностями вероятности

$$f_{\xi_i}(x,a,\sigma_i^2)=\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(x-a)^2}{2\sigma_i^2}},\ i=\overline{1,n}.$$

Имеется выборка $x_1, x_2, ..., x_n$ из этих совокупностей, содержащая по одному выборочному значению из каждой совокупности. Требуется по этим данным найти оценку $\hat{a}$ параметра a.

Для решения данной задачи можно воспользоваться методом максимума правдоподобия. Логарифмическая функция правдоподобия для одного выборочного значения x_i имеет вид

$$\ln f_{\xi_i}(x_i, a, \sigma_i^2) = \ln(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2}\ln\sigma_i^2 - \frac{(x_i - a)^2}{2\sigma_i^2}.$$

Отсюда получаем производную

$$\frac{d}{da}(\ln f_{\xi_i}(x_i, a, \sigma_i^2) = \frac{x_i - a}{\sigma_i^2}$$

и уравнение для нахождения оценки

$$\sum_{i=1}^{n} \frac{x_i - a}{\sigma_i^2} = 0.$$

После преобразования этого уравнения к виду

$$\sum_{i=1}^{n} \sigma_i^{-2} x_i - a\sum_{i=1}^{n} \sigma_i^{-2} = 0$$

получаем оценку

$$\hat{a} = (\sum_{j=1}^{n} \sigma_j^{-2})^{-1} (\sum_{i=1}^{n} \sigma_i^{-2} x_i).$$

Это также линейная комбинация выборочных значений, как и в примере равноточных измерений раздела 2.2, но с различными весами σ_i^{-2}. В этом случае вес наблюдения x_i обратно пропорционален дисперсии наблюдения σ_i^2. При одинаковых дисперсиях наблюдений такая оценка сводится к оценке, полученной в примере 2.3 раздела 2.2.

2.4 Метод максимума апостериорной плотности вероятности

Данный метод предполагает наличие более полной априорной информации по сравнению с методом максимума правдоподобия. Здесь считается, что векторный параметр $\bar{\theta} = (\theta_1,...,\theta_m)$ является случайным вектором с известной априорной плотностью вероятности $f(\bar{\theta})$, который на время извлечения выборки принял одно из своих значений. Считается известной также плотность вероятности $f(x,\bar{\theta})$, которую удобно теперь считать условной и обозначать $f(x/\bar{\theta})$. Задача состоит в том, чтобы по полученной выборке $X = (x_1,...,x_n)$ найти оценку реализации вектора $\bar{\theta}$.

Метод максимума апостериорной плотности вероятности состоит в том, что оценки определяются из условия максимума апостериорной плотности вероятности параметра $\bar{\theta}$:

$$f(\bar{\theta} / X) \to \max_{\bar{\theta}},$$

где

$$f(\bar{\theta} / X) = \frac{f(\bar{\theta}) f(X / \bar{\theta})}{\int_{-\infty}^{\infty} f(\bar{\theta}) f(X / \bar{\theta}) d\bar{\theta}} - \qquad (2.4)$$

формула Байеса, определяющая апостериорную плотность вероятности вектора $\bar{\theta}$, $f(X / \bar{\theta}) = L(\bar{\theta})$ – известная нам функция правдоподобия (2.1).

Если учесть, что знаменатель формулы Байеса не зависит от параметра $\bar{\theta}$ и, следовательно, не влияет на результат оптимизации, то рассматриваемый метод сводится к решению оптимизационной задачи вида

$$f(\bar{\theta}) L(\bar{\theta}) \to \max_{\bar{\theta}}.$$

Мы видим, что данная задача отличается от задачи (2.2) метода максимума правдоподобия тем, что максимизируемая функция содержит в качестве дополнительного множителя априорную плотность вероятности $f(\bar{\theta})$. Если вместо

исходной функции использовать ее логарифм, то для определения оценок по методу максимума апостериорной плотности вероятности нам необходимо решать систему уравнений

$$\frac{\partial}{\partial\theta_j} ln\, f(\theta_1,...,\theta_m) + \sum_{i=1}^{n} \frac{\partial}{\partial\theta_j} ln\, f_\xi(x_i,\theta_1,...,\theta_m) = 0\,,\ j = \overline{1,m.} \qquad (2.5)$$

Эта система отличается от системы уравнений (2.3) метода максимума правдоподобия наличием дополнительного слагаемого $\frac{\partial}{\partial\theta_j} ln\, f(\theta_1,...,\theta_m)$.

Пример 2.5. Найти оценку параметра a нормального распределения с плотностью вероятности

$$f(x/a) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

методом максимума апостериорной плотности вероятности в предположении, что этот параметр также распределен по нормальному закону с плотностью вероятности

$$f(a) = \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{(a-a_0)^2}{2\sigma_a^2}}.$$

Поскольку

$$ln\, f_\xi(x_i,a,\sigma^2) = ln(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2} ln\,\sigma^2 - \frac{(x_i-a)^2}{2\sigma^2},$$

$$ln\, f(a) = ln(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2} ln\,\sigma_a^2 - \frac{(a-a_0)^2}{2\sigma_a^2},$$

$$\frac{d}{da} ln\, f_\xi(x_i,a,\sigma^2) = \frac{x_i-a}{\sigma^2},$$

$$\frac{d}{da} ln\, f(a) = -\frac{a-a_0}{\sigma_a^2},$$

то система уравнений (2.5) будет состоять из одного уравнения вида

$$-\frac{a-a_0}{\sigma_a^2}+\sum_{i=1}^{n}\frac{x_i-a}{\sigma^2}=0\,.$$

Переписав его следующим образом

$$-\sigma^2(a-a_0)+\sigma_a^2\sum_{i=1}^{n}(x_i-a)=0\,,$$

получим оценку

$$\hat{a}=\frac{\sigma^2 a_0+\sigma_a^2\sum_{i=1}^{n}x_i}{\sigma^2+n\sigma_a^2}\,. \tag{2.6}$$

Выполним анализ структуры полученной оценки $\hat{a}$. Для этого разделим на $\sigma^2\sigma_a^2$ числитель и знаменатель полученного выражения. Будем иметь

$$\hat{a}=\frac{1}{\hat{\sigma}^{-2}}\left(\frac{1}{\sigma_a^2}a_0+\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i\right),$$

где

$$\hat{\sigma}^{-2}=\frac{n}{\sigma^2}+\frac{1}{\sigma_a^2}=\frac{\sigma^2+n\sigma_a^2}{\sigma^2\sigma_a^2}\,. \tag{2.7}$$

Мы видим, что оценка $\hat{a}$ является линейной комбинацией наблюдений x_i и априорных данных в виде априорного среднего. При этом используемые данные имеют различные веса. Наблюдения x_i имеют одинаковые веса $1/\sigma^2$, где σ^2 – дисперсия генеральной совокупности, в то время как априорное среднее a_0 – другой вес $1/\sigma_a^2$, где σ_a^2 – априорная дисперсия параметра a. Полученная сумма делится на сумму всех весов (2.7). Если число наблюдений n мало, то существенным будет вклад в оценку априорных данных a_0. По мере увеличения n вклад наблюдений x_i в оценку $\hat{a}$ будет увеличиваться, а априорных данных a_0 – уменьшаться. Таким образом, оценка по методу максимума апостериорной плотности вероятности представляет собой разумную комбинацию априорных данных и наблюдений.

2.5 Байесовский метод

Исходные данные для этого метода те же, что и для метода максимума апостериорной плотности вероятности, а именно: векторный параметр $\bar{\theta} = (\theta_1,...,\theta_m)$ считается случайным с известной априорной плотностью вероятности $f(\bar{\theta})$, а также известна с точностью до параметра $\bar{\theta}$ плотность вероятности генеральной совокупности $f(x/\bar{\theta})$. Дополнительно для характеристики качества оценивания вводится функция потерь $W(\hat{\bar{\theta}},\bar{\theta})$, зависящая от параметра и его оценки, и средний риск r как математическое ожидание функции потерь:

$$r = E(W(\hat{\bar{\theta}},\bar{\theta})).$$

Требуется по полученной выборке $X = (x_1,...,x_n)$ найти оценку реализации вектора $\bar{\theta}$.

Метод состоит в том, что оценка находится из условия минимума среднего риска:

$$r \to \min_{\hat{\bar{\theta}}}.$$

Для решения данной задачи учтем, что $\hat{\bar{\theta}} = \hat{\bar{\theta}}(X)$ и запишем выражение среднего риска в виде

$$r = \int_{-\infty}^{\infty} W(\hat{\bar{\theta}},\bar{\theta}) f(X,\bar{\theta}) dX d\bar{\theta},\ dX = dx_1 \cdots dx_n,\ d\bar{\theta} = d\theta_1 \cdots d\theta_m.$$

Представим совместную плотность вероятности $f(X,\bar{\theta})$ по теореме умножения в виде произведения

$$f(X,\bar{\theta}) = f(X) f(\bar{\theta}/X),$$

где $f(\bar{\theta}/X)$ – апостериорная плотность вероятности параметра $\bar{\theta}$, определяемая формулой Байеса (2.4). Введем понятие условного риска

$$R = R(X) = \int_{-\infty}^{\infty} W(\hat{\bar{\theta}}, \bar{\theta}) f(\bar{\theta} / X) d\bar{\theta} \qquad (2.8)$$

при условии, что фиксирована выборка X. Тогда средний риск

$$r = \int_{-\infty}^{\infty} R(X) f(X) dX .$$

Данное выражение является функционалом. В вариационном исчислении известна следующая теорема об экстремуме функционала [10].

Теорема 2.1. Для того чтобы функционал

$$J(y) = \int_{a}^{b} F(x, y, y') dx ,$$

определенный на множестве функций $y = y(x)$, имеющих непрерывную первую производную и удовлетворяющих условиям $y(a) = A$, $y(b) = B$, достигал на данной функции $y(x)$ экстремума, необходимо, чтобы эта функция удовлетворяла уравнению Эйлера

$$F_x - \frac{d}{dx} F_{y'} = 0 .$$

Данную теорему можно применить для минимизации среднего риска. Поскольку подынтегральная функция в выражении для среднего риска не зависит от $\hat{\bar{\theta}}'$, то второе слагаемое в уравнении Эйлера будет отсутствовать. Таким образом, мы свели задачу к необходимости решения уравнения

$$\frac{dR}{d\hat{\bar{\theta}}} = 0 ,$$

т.е. к оптимизационной задаче вида

$$R = \int_{-\infty}^{\infty} W(\hat{\bar{\theta}}, \bar{\theta}) f(\bar{\theta} / X) d\bar{\theta} \to \min_{\hat{\bar{\theta}}} .$$

Чаще всего используется квадратичная функция потерь, то есть функция потерь вида

$$W(\hat{\bar{\theta}},\bar{\theta}) = \sum_{i=1}^{m}(\hat{\theta}_i - \theta_i)^2 .$$

Рассмотрим случай скалярного параметра и квадратичной функции потерь $W(\hat{\theta},\theta) = (\hat{\theta} - \theta)^2$ и найдем байесовскую оценку $\hat{\theta}$ параметра θ. Условный риск (2.8) в данном случае имеет вид

$$R = \int_{-\infty}^{\infty}(\hat{\theta} - \theta)^2 f(\theta / X)d\theta .$$

Необходимое условие экстремума условного риска приводит к уравнению

$$\frac{d}{d\hat{\theta}} R(\hat{\theta}) = \int_{-\infty}^{\infty} 2(\hat{\theta} - \theta) f(\theta / X)d\theta = 0 ,$$

откуда получаем

$$\int_{-\infty}^{\infty} \hat{\theta} f(\theta / X)d\theta - \int_{-\infty}^{\infty} \theta f(\theta / X)d\theta = 0 ,$$

$$\hat{\theta} = \int_{-\infty}^{\infty} \theta f(\theta / X)d\theta = E(\theta / X) .$$

Таким образом, байесовская оценка при квадратичной функции потерь равна апостериорному математическому ожиданию. Для получения этой оценки нам необходимо найти апостериорную плотность вероятности параметра и по ней найти апостериорное математическое ожидание.

Пример 2.6. Найти байесовскую оценку параметра a нормального распределения с плотностью вероятности

$$f(x / a) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

в предположении, что этот параметр также распределен по нормальному закону с плотностью вероятности

$$f(a) = \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{(a-a_0)^2}{2\sigma_a^2}}$$

и используется квадратичная функция потерь $W(\hat{a},a)=(\hat{a}-a)^2$.

В соответствии с теорией, изложенной выше, при квадратичной функции потерь байесовская оценка определяется как апостериорное математическое ожидание.

Для решения задачи запишем выражение для функции правдоподобия

$$f(X/a)=\prod_{i=1}^{n}f(x_i/a)=\frac{1}{\sqrt{(2\pi)^n\sigma^{2n}}}exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right).$$

Найдем апостериорную плотность вероятности $f(a/X)$ параметра a по формуле Байеса

$$f(a/X)=\frac{f(a)f(X/a)}{f(X)}, \qquad (2.9)$$

где

$$f(X)=\int_{-\infty}^{\infty}f(a)f(X/a)da.$$

Найдем числитель формулы Байеса:

$$f(a)f(X/a)=$$

$$=\frac{1}{\sqrt{(2\pi)^{n+1}\sigma_a^2\sigma^{2n}}}\exp\left(-\frac{1}{2\sigma^2\sigma_a^2}\left((a-a_0)^2\sigma^2+\sigma_a^2\sum_{i=1}^{n}(x_i-a)^2\right)\right)=$$

$$=\frac{1}{\sqrt{(2\pi)^{n+1}\sigma_a^2\sigma^{2n}}}\exp\left(-\frac{1}{2\sigma^2\sigma_a^2}\left(\sigma^2a^2-2\sigma^2a_0a+\sigma^2a_0^2+\right.\right.$$

$$\left.\left.+\sigma_a^2\sum x_i^2-2\sigma_a^2a\sum x_i+n\sigma_a^2a^2\right)\right)=\frac{1}{\sqrt{(2\pi)^{n+1}\sigma_a^2\sigma^{2n}}}\exp\left(-\frac{1}{2\sigma^2\sigma_a^2}(\sigma^2a_0^2+\right.$$

$$\left.+\sigma_a^2\sum_{i=1}^{n}x_i^2)\right)\exp\left(-\frac{1}{2\sigma^2\sigma_a^2}\left(\sigma^2a^2+n\sigma_a^2a^2-2(\sigma^2a_0+\sigma_a^2\sum_{i=1}^{2}x_i)a\right)\right)=\varphi_1(a_0)\varphi_2(a),$$

где $\varphi_2(a)$ имеет вид

$$\varphi_2(a)=\exp\left(-\frac{1}{2\sigma^2\sigma_a^2}\left(\sigma^2a^2+n\sigma_a^2a^2-2(\sigma^2a_0+\sigma_a^2\sum_{i=1}^{n}x_i)a\right)\right),$$

а выражение $\varphi_1(a_0)$ очевидно. Таким образом, мы представили апостериорную плотность вероятности (2.9) в виде

$$f(a/X) = c\varphi_2(a),$$

где c – нормирующий множитель. Преобразуем это выражение:

$$f(a/X) = c\varphi_2(a) = c\,exp\left(-\frac{\sigma^2 + n\sigma_a^2}{2\sigma^2\sigma_a^2}\left(a^2 - \frac{2}{\sigma^2 + n\sigma_a^2}(\sigma^2 a_0 + \sigma_a^2\sum_{i=1}^{n} x_i)a\right)\right).$$

Дополним сумму, содержащую a^2, до полного квадрата:

$$f(a/X) = c\,exp\left(-\frac{\sigma^2 + n\sigma_a^2}{2\sigma^2\sigma_a^2}\left(a^2 - \frac{2}{\sigma^2 + n\sigma_a^2}(\sigma^2 a_0 + \sigma_a^2\sum_{i=1}^{n} x_i)a + \right.\right.$$

$$\left.\left. + \left(\frac{\sigma^2 a_0 + \sigma_a^2\sum_{i=1}^{n} x_i}{\sigma^2 + n\sigma_a^2}\right)^2 - \left(\frac{\sigma^2 a_0 + \sigma_a^2\sum_{i=1}^{n} x_i}{\sigma^2 + n\sigma_a^2}\right)^2\right)\right) =$$

$$= c\exp\left(-\frac{\sigma^2 + n\sigma_a^2}{2\sigma^2\sigma_a^2}\left(\frac{\sigma^2 a_0 + \sigma_a^2\sum_{i=1}^{n} x_i}{\sigma^2 + n\sigma_a^2}\right)^2\right)\exp\left(-\frac{\sigma^2 + n\sigma_a^2}{2\sigma^2\sigma_a^2}\left(a - \frac{\sigma^2 a_0 + \sigma_a^2\sum_{i=1}^{n} x_i}{\sigma^2 + n\sigma_a^2}\right)^2\right) =$$

$$= c_1\varphi_4(a),$$

где

$$\varphi_4(a) = \exp\left(-\frac{\sigma^2 + n\sigma_a^2}{2\sigma^2\sigma_a^2}\left(a - \frac{\sigma^2 a_0 + \sigma_a^2\sum_{i=1}^{n} x_i}{\sigma^2 + n\sigma_a^2}\right)^2\right),$$

а c_1 – новый нормирующий множитель, который, очевидно, равен

$$c_1 = \frac{1}{\sqrt{2\pi\dfrac{\sigma^2\sigma_a^2}{\sigma^2 + n\sigma_a^2}}}.$$

В итоге получаем, что

$$f(a/X)=\frac{1}{\sqrt{2\pi\dfrac{\sigma^2\sigma_a^2}{\sigma^2+n\sigma_a^2}}}exp\left(-\frac{1}{2\dfrac{\sigma^2\sigma_a^2}{\sigma^2+n\sigma_a^2}}\left(a-\frac{\sigma^2a_0+\sigma_a^2\sum\limits_{i=1}^{n}x_i}{\sigma^2+n\sigma_a^2}\right)^2\right), \quad (2.10)$$

то есть апостериорная плотность вероятности параметра a является нормальной плотностью вероятности вида $f(a/X)=N(\hat{a},\hat{\sigma}^2)$, где

$$\hat{a}=\frac{\sigma^2a_0+\sigma_a^2\sum\limits_{i=1}^{n}x_i}{\sigma^2+n\sigma_a^2}- \quad (2.11)$$

апостериорное среднее, или апостериорное математическое ожидание, которое и является байесовской оценкой. При этом апостериорная дисперсия равна

$$\hat{\sigma}^2=\frac{\sigma^2\sigma_a^2}{\sigma^2+n\sigma_a^2}. \quad (2.12)$$

Мы видим, что получили ту же оценку, что и по методу максимума апостериорной плотности вероятности (2.6). Это понятно, поскольку максимум нормальной плотности вероятности достигается в точке, равной математическому ожиданию. Минимальное значение условного риска определяется выражением

$$R_{min}=E((a-\hat{a})^2)=\int\limits_{-\infty}^{\infty}(a-\hat{a})^2f(a/X)da=\hat{\sigma}^2,$$

где $f(a/X)$, $\hat{a}$ и $\hat{\sigma}^2$ имеют вид (2.10), (2.11), (2.12) соответственно. Мы видим, что это значение равно апостериорной дисперсии (2.12). Таким образом, оценка (2.11) обеспечивает минимальное значение условного и среднего риска $r_{min}=R_{min}=\hat{\sigma}^2$.

2.6 Оценивание параметров по косвенным измерениям (классический метод наименьших квадратов)

До сих пор мы рассматривали случаи, когда оцениваемый параметр измеряется непосредственно измерительным прибором. Возможен также случай, когда оцениваемые параметры непосредственно измерительным прибором не измеряются, а измеряются некоторые другие величины, функционально связанные с оцениваемыми параметрами. Решение такой задачи выполняется методом наименьших квадратов. Классическая задача о методе наименьших квадратов формулируется следующим образом [9]. Предполагается, что результаты измерений (показания приборов) z_i функционально связаны с оцениваемыми параметрами $\theta_1,\ldots,\theta_m$:

$$z_i = \psi_i(\theta_1,\ldots,\theta_m, x_{i,1},\ldots,x_{i,l}) + e_i,\ \ i=\overline{1,n}, \tag{2.13}$$

где $\psi_i(\theta_1,\ldots,\theta_m, x_{i,1},\ldots,x_{i,l})$ – некоторые известные скалярные функции; e_i – ошибки измерений; $x_{i,1},\ldots,x_{i,l}$ – входные переменные, которые измеряются с ошибками, измеряются точно (то есть известны), или отсутствуют. Требуется по измерениям z_i и $x_{i,1},\ldots,x_{i,l}$ найти оценки $\hat{\theta}_1,\ldots,\hat{\theta}_m$ неизвестных параметров $\theta_1,\ldots,\theta_m$. Будем рассматривать случай, когда входные переменные $x_{i,1},\ldots,x_{i,l}$ известны или отсутствуют. В этом случае в выражениях (2.13) их можно не учитывать и вместо (2.13) записать выражения

$$z_i = \psi_i(\bar{\theta}) + e_i,\ \ i=\overline{1,n}, \tag{2.14}$$

где $\bar{\theta} = (\theta_1,\ldots,\theta_m)$ – вектор неизвестных параметров.

Задача в таком виде была сформулирована Гауссом. Для ее решения Гаусс предложил свой знаменитый метод наименьших квадратов (МНК). Метод наименьших квадратов состоит в том, что оценки параметров определяются из условия минимума суммы квадратов ошибок измерений, то есть как решение оптимизационной задачи

$$F(\bar{\theta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (z_i - \psi_i(\bar{\theta}))^2 \to \min_{\bar{\theta}}.$$

В настоящее время эта задача формулируется и решается с использованием векторно-матричного подхода. Зависимости (2.14) записывают в векторной форме:

$$Z = \Psi(\bar{\theta}, X) + E, \tag{2.15}$$

где $Z^T = (z_1, \ldots, z_n)$ – вектор показаний приборов; $\Psi^T(\bar{\theta}, X) = (\psi_1(\bar{\theta}, X_1), \ldots, \psi_n(\bar{\theta}, X_n))$ – векторная функция, $\bar{\theta}^T = (\theta_1, \ldots, \theta_m)$ – вектор параметров, $X = (X_1, \ldots, X_n)$ – массив входных переменных, которые либо известны (измеряются точно), либо отсутствуют, $E^T = (e_1, \ldots, e_n)$ – вектор ошибок измерений с нулевым средним значением и ковариационной матрицей R_E. Требуется по результатам измерений Z, X найти оценку $\hat{\bar{\theta}}$ вектора параметров $\bar{\theta}$.

Оптимизационная задача теперь запишется в виде:

$$F(\bar{\theta}) = E^T E = (Z - \Psi(\bar{\theta}))^T (Z - \Psi(\bar{\theta})) \to \min_{\bar{\theta}}.$$

Чтобы получить аналитическое выражение для МНК-оценки, нужно функцию $\Psi(\bar{\theta})$ в (2.15) линеаризовать в окрестности некоторой опорной точки $\bar{\theta}^{(0)}$, то есть представить ее в окрестности точки $\bar{\theta}^{(0)}$ рядом Тейлора и удержать в разложении только два члена. Получим

$$\Psi(\bar{\theta}) \approx \Psi(\bar{\theta}_0) + Q(\bar{\theta} - \bar{\theta}^{(0)}),$$

где

$$Q = \Psi'(\bar{\theta}^{(0)}) = \frac{d}{d\bar{\theta}^{(0)}} \Psi(\bar{\theta}^{(0)}). \tag{2.16}$$

В результате линеаризации будем иметь следующую оптимизационную задачу:

$$F(\bar{\theta}) = [Z - \Psi(\bar{\theta}^{(0)}) - Q(\bar{\theta} - \bar{\theta}^{(0)})]^T [Z - \Psi(\bar{\theta}^{(0)}) - Q(\bar{\theta} - \bar{\theta}^{(0)})] \to \min_{\bar{\theta}}.$$

Необходимые условия экстремума представляют собой уравнение

$$\frac{d}{d\bar{\theta}} F(\bar{\theta}) = 0.$$

Перепишем функцию $F(\bar{\theta})$ в виде

$$F(\bar{\theta}) = (Z - \Psi(\bar{\theta}^{(0)}))^T (Z - \Psi(\bar{\theta}^{(0)})) - (Z - \Psi(\bar{\theta}^{(0)}))^T Q(\bar{\theta} - \bar{\theta}^{(0)}) -$$

$$- (Q(\bar{\theta} - \bar{\theta}^{(0)}))^T (Z - \Psi(\bar{\theta}^{(0)})) + (Q(\bar{\theta} - \bar{\theta}^{(0)}))^T (Q(\bar{\theta} - \bar{\theta}^{(0)})).$$

Напомним, что функция $F(\bar{\theta})$ и все ее слагаемые являются скалярными функциями вектора $\bar{\theta}$. Воспользуемся правилом транспонирования произведения матриц $(AB)^T = B^T A^T$ и тем фактом, что для скалярной величины $A^T = A$. Тогда

$$F(\bar{\theta}) = (Z - \Psi(\bar{\theta}^{(0)}))^T (Z - \Psi(\bar{\theta}^{(0)})) - 2(Z - \Psi(\bar{\theta}^{(0)}))^T Q(\bar{\theta} - \bar{\theta}^{(0)}) +$$

$$+ (\bar{\theta} - \bar{\theta}^{(0)})^T Q^T Q(\bar{\theta} - \bar{\theta}^{(0)}),$$

и уравнение для нахождения оценки $\hat{\bar{\theta}}$ имеет вид

$$\frac{d}{d\bar{\theta}} F(\bar{\theta}) = -2Q^T (Z - \Psi(\bar{\theta}^{(0)})) + 2Q^T Q(\bar{\theta} - \bar{\theta}^{(0)}) = 0.$$

Отсюда последовательно получаем

$$Q^T Q(\bar{\theta} - \bar{\theta}^{(0)}) = Q^T (Z - \Psi(\bar{\theta}^{(0)})),$$

$$(\bar{\theta} - \bar{\theta}^{(0)}) = (Q^T Q)^{-1} Q^T (Z - \Psi(\bar{\theta}^{(0)})),$$

$$\hat{\bar{\theta}} = \bar{\theta}^{(0)} + (Q^T Q)^{-1} Q^T (Z - \Psi(\bar{\theta}^{(0)})). \qquad (2.17)$$

Последнее выражение и есть оценка неизвестных параметров по методу наименьших квадратов.

Можно показать, что в условиях линейности функции $\Psi(\bar{\theta})$ выражение

$$R_{\hat{\theta}NK} = (Q^T R_E^{-1} Q)^{-1}$$

представляет собой ковариационную матрицу оценки, которая характеризует точность оценки.

Мы видим, что МНК-оценка (2.17) зависит от опорной точки $\bar{\theta}^{(0)}$, то есть от точки, в окрестности которой выполняется линеаризация. Для повышения точности оценивания эта точка должна быть по возможности ближе к истинному значению параметра $\bar{\theta}$. Если опорная точка $\bar{\theta}^{(0)}$ выбрана неудачно, то метод наименьших квадратов может дать плохую оценку.

Пример 2.7. Найти МНК-оценку высоты объекта θ по точным измерениям расстояний x_i от его основания до некоторых точек наблюдения p_i и измерениям с ошибками z_i углов из этих точек на вершину объекта, $i = \overline{1,n}$.

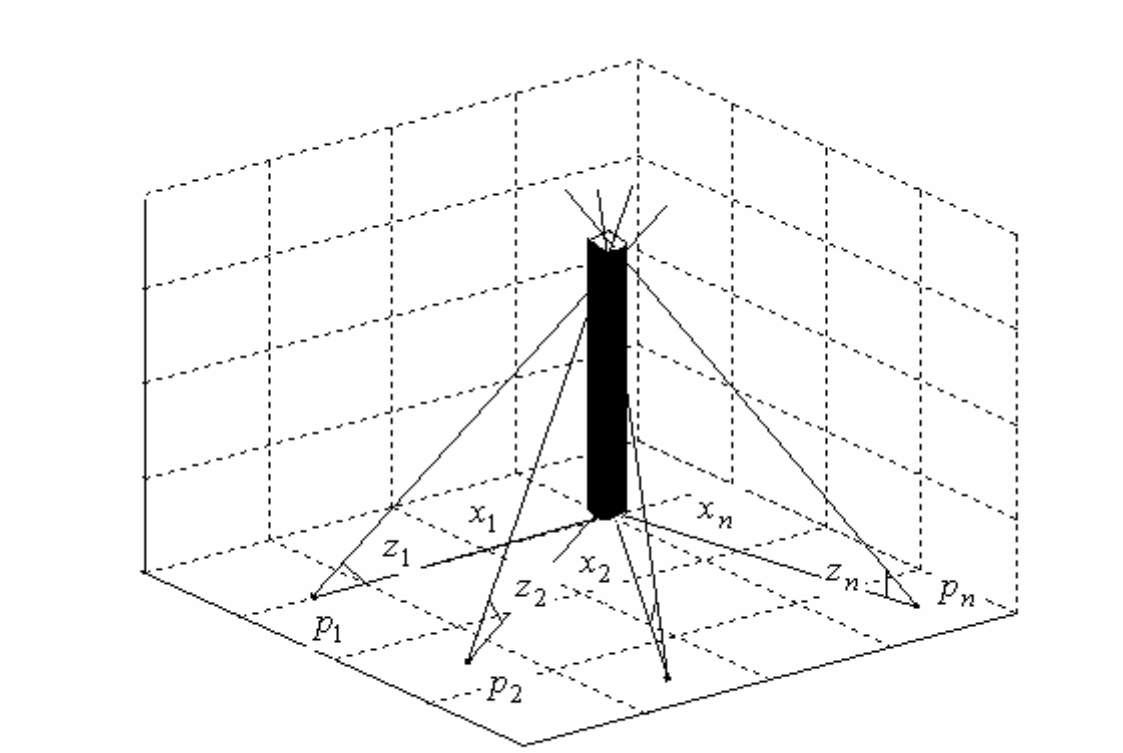

Рис. 2.1. Иллюстрация к примеру 2.7

Графическая иллюстрация задачи приведена на рис. 2.1. Наблюдения углов z_i связаны с высотой объекта θ соотношением

$$z_i = arc\,tg\,\frac{\theta}{x_i} + e_i,\ i = \overline{1,n},$$

где e_i – ошибки наблюдений. Введя вектор-столбец наблюдений $Z = (z_i)$, вектор-столбец функций $\Psi(\theta) = (\phi_i(\theta))$, где

$$\phi_i(\theta) = arc\,tg\,\frac{\theta}{x_i},$$

и вектор-столбец ошибок измерений $E=(e_i)$, мы можем воспользоваться оценкой (2.17). Матрица Q (2.16) в этом выражении представляет собой вектор-столбец $Q=(q_i)$, где

$$q_i=\frac{d\phi_i(\theta^{(0)})}{d\theta^{(0)}}=\frac{1}{\left(1+\frac{(\theta^{(0)})^2}{x_i^2}\right)x_i}=\frac{x_i}{x_i^2+(\theta^{(0)})^2}.$$

Рассчитав указанные матрицы, мы можем затем найти оценку по матричной формуле (2.17). Получим также обычную формулу для оценки. В данном случае Q^TQ и $Q^T(Z-\Psi(\bar{\theta}^{(0)}))$ представляют собой скалярные величины вида

$$Q^TQ=\sum_{i=1}^{n}q_i^2=\sum_{i=1}^{n}\frac{x_i^2}{\left(x_i^2+(\theta^{(0)})^2\right)^2},$$

$$Q^T(Z-\Psi(\bar{\theta}^{(0)}))=\sum_{i=1}^{n}q_i(z_i-\phi_i(\theta))=$$

$$=\sum_{i=1}^{n}\frac{x_i}{x_i^2+(\theta^{(0)})^2}\left(z_i-arc\,tg\,\frac{\theta^{(0)}}{x_i}\right).$$

В итоге для МНК-оценки высоты объекта получаем выражение

$$\hat{\theta}=\theta^{(0)}+\left(\sum_{i=1}^{n}\frac{x_i^2}{\left(x_i^2+(\theta^{(0)})^2\right)^2}\right)^{-1}\left(\sum_{i=1}^{n}\frac{x_i}{x_i^2+(\theta^{(0)})^2}\left(z_i-arc\,tg\,\frac{\theta^{(0)}}{x_i}\right)\right).$$

3 НЕКОТОРЫЕ РАСПРЕДЕЛЕНИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

До сих пор мы ограничивались рассмотрением лишь числовых характеристик статистик. Например, для выяснения несмещенности точечной оценки параметра нам достаточно было найти ее математическое ожидание. Однако некоторые задачи математической статистики, такие как построение доверительных интервалов для параметров или статистическая проверка гипотез, требуют знания законов распределения статистик. Поэтому в данном разделе познакомимся с некоторыми распределениями математической статистики.

3.1 Распределение хи-квадрат

Случайная величина вида

$$\chi_k^2 = \sum_{i=1}^{k} x_i^2 ,$$

где $x_1, x_1, ..., x_k$ – независимые случайные величины, распределенные по нормальному закону $N(0,1)$, имеет распределение, которое называется одномерным распределением хи-квадрат с k степенями свободы и обозначается как $H_1(k)$.

Плотность вероятности распределения хи-квадрат имеет вид

$$f_{\chi_k^2}(x) = \begin{cases} \dfrac{1}{2\Gamma\left(\dfrac{k}{2}\right)} \left(\dfrac{x}{2}\right)^{\frac{k}{2}-1} e^{-\frac{x}{2}}, & x > 0,\ k = 1,2,..., \\ 0, & x \le 0, \end{cases}$$

где $\Gamma\left(\dfrac{k}{2}\right)$ – гамма-функция, определяемая выражением

$$\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy .$$

Гамма-функция обладает следующими свойствами:

$$\Gamma(x+1) = x\Gamma(x),\ \Gamma(k+1) = k!,\ \Gamma(1) = \Gamma(2) = 1,\ \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} .$$

Кривые плотности вероятности этого распределения изображены на рис. 3.1.

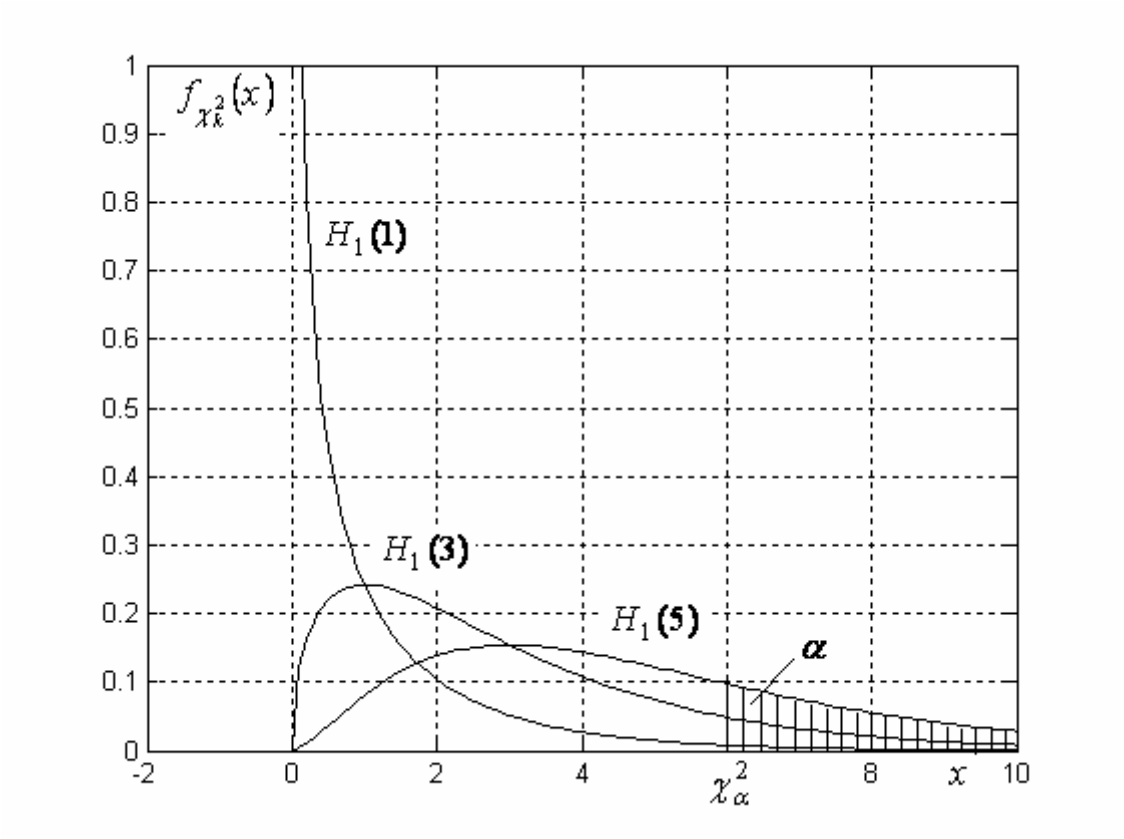

Рис. 3.1. Плотность вероятности распределения хи-квадрат

Это несимметричные кривые, расположенные на положительной полуоси абсцисс. Кривые имеют по одному максимуму в точке $x = k - 2$. Математическое ожидание и дисперсия распределения хи-квадрат равны

$$E(\chi_k^2) = k ,$$

$$D(\chi_k^2) = 2k .$$

Для практических приложений составлены таблицы процентных отклонений распределения хи-квадрат. Эти таблицы позволяют решать уравнения вида

$$P(\chi_k^2 > \chi_\alpha^2) = \alpha,\ 0 \le \alpha \le 1.$$

Это значит, что с помощью таблиц для некоторого k и некоторой вероятности α можно найти величину χ_α^2, удовлетворяющую указанному равенству, и наоборот, по χ_α^2 найти α. Величины α и χ_α^2 приведены на рис. 3.1: α равна площади, заштрихованной на рис. 3.1, а χ_α^2 есть левая граница этой области. Величина χ_α^2 называется 100α-процентным отклонением случайной величины χ_k^2.

Из определения распределения хи-квадрат очевидно свойство: если $\chi_p^2 \in H_1(p)$, $\chi_q^2 \in H_1(q)$ и χ_p^2, χ_q^2 независимы, то

$$\chi_p^2 + \chi_q^2 \in H_1(p+q). \tag{3.1}$$

3.2 Распределение Стьюдента (t-распределение)

Случайная величина

$$t = \frac{u}{\sqrt{v}}\sqrt{n},$$

где u и v – независимые случайные величины, причем $u \in N(0,1)$, $v \in H_1(n)$, имеет распределение, которое называется распределением Стьюдента с n степенями свободы и обозначается $T_1(n)$. Плотность вероятности распределения Стьюдента имеет вид

$$f_t(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}.$$

Кривые плотности вероятности этого распределения изображены на рис 3.2.

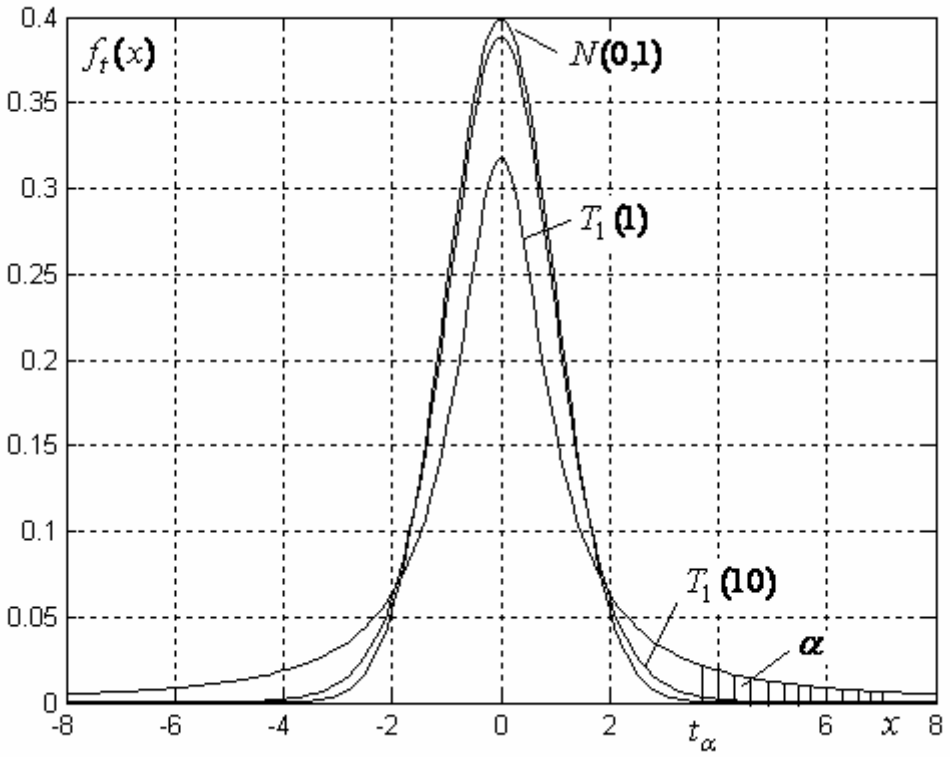

Рис. 3.2. Плотность вероятности распределения Стьюдента

Это симметричные относительно оси ординат кривые. При $n=1$ это распределение Коши, которое известно тем, что не имеет ни одного момента. При $n>2$ математическое ожидание $E(t)=0$, а дисперсия $D(t)=n/(n-2)$.

При $n\to\infty$ распределение Стьюдента приближается к нормальному распределению $N(0,1)$. Однако при малых n ($n<30$) оно заметно отличается от нормального. Существуют таблицы процентных отклонений распределения Стьюдента. В этих таблицах для фиксированных n и α можно найти число t_α, удовлетворяющее равенству

$$P(t>t_\alpha)=\alpha,$$

и наоборот, по n и t_α найти α. Величины t_α и α отмечены на рис. 3.2 штриховкой.

3.3 Распределение Фишера (f -распределение)

Случайная величина

$$f = \frac{v}{m} : \frac{w}{n},$$

где v и w – независимые случайные величины, $v \in H_1(m)$, $w \in H_1(n)$, имеет распределение, которое называется одномерным распределением Фишера с m, n степенями свободы и обозначается $F_1(m,n)$.

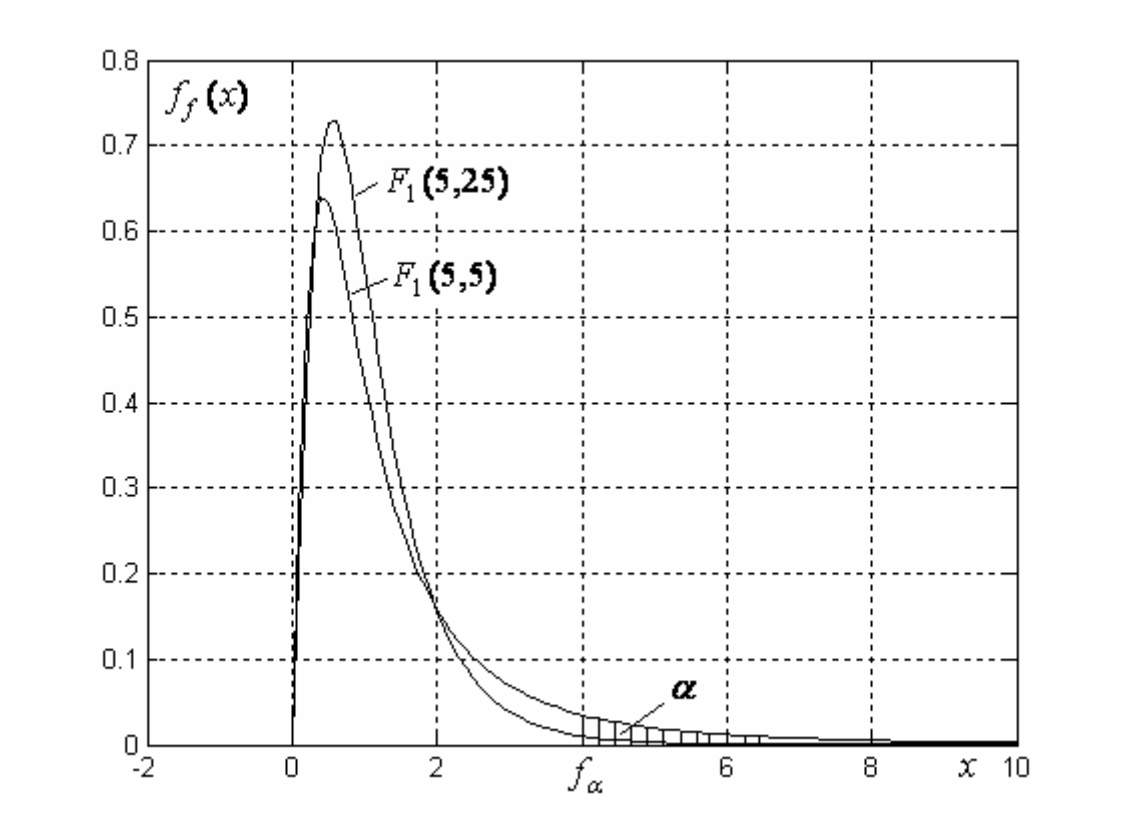

Рис. 3.3. Плотность вероятности распределения Фишера

Кривые плотности вероятности этого распределения приведены на рис 3.3. Они определяются выражением

$$f_f(x)=\begin{cases}\dfrac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\,n^{\frac{n}{2}}m^{\frac{m}{2}}\,\dfrac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{n+m}{2}}}, & x\geq 0,\\ 0, & x<0.\end{cases}$$

Это несимметричные кривые, расположенные на положительной полуоси абсцисс, которые достигают максимума вблизи точки $x=1$.

Для практического использования этого распределения существуют таблицы процентных отклонений, которые позволяют для фиксированных m, n и α найти число f_α, удовлетворяющее равенству

$$P(f>f_\alpha)=\alpha,$$

и наоборот, по m, n и f_α найти α. Величины f_α и α отмечены на рис. 3.3 штриховкой.

Из определения распределения Фишера вытекает, что если $f\in F_1(m,n)$, то $f^{-1}\in F_1(n,m)$.

3.4 Распределения некоторых статистик для нормальной генеральной совокупности

Теорема 3.1. Если $x_1,x_2,...,x_n$ – выборка из распределения $N(a,\sigma^2)$, $\bar{x}$ – выборочное среднее, $\bar{s}^2$ – выборочная дисперсия, s^2 – несмещенная выборочная дисперсия, $\bar{s}_0^2$ – выборочная дисперсия при известном математическом ожидании (см. разделы 1.8, 1.9), то

$$u=\frac{\bar{x}-a}{\sigma}\sqrt{n}\in N(0,1), \tag{3.2}$$

$$v=\frac{n\bar{s}_0^2}{\sigma^2}\in H_1(n), \tag{3.3}$$

$$w = \frac{n\bar{s}^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \in H_1(n-1), \tag{3.4}$$

$$t = \frac{\bar{x}-a}{\bar{s}}\sqrt{n-1} = \frac{\bar{x}-a}{s}\sqrt{n} \in T_1(n-1), \tag{3.5}$$

причем величины u и v независимы.

Доказательство утверждения (3.2). Рассмотрим выборочное среднее

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

где $x_1, x_2, ..., x_n$ – независимые случайные величины. Обозначим $\eta_i = \frac{x_i}{n}$. Эти величины имеют распределение $N(\frac{a}{n}, \frac{\sigma^2}{n^2})$ и характеристические функции

$$\theta_{\eta_i}(v) = \exp\left(j\frac{a}{n}v - v^2\,\frac{\sigma^2}{2n^2} \right).$$

Поскольку $\bar{x} = \sum_{i=1}^{n} \eta_i$, то по теореме о произведении характеристических функций [14] получим

$$\theta_{\bar{x}}(v) = \prod_{i=1}^{n} \theta_{\eta_i}(v) = \exp\left(jv\sum_{i=1}^{n}\frac{a}{n} - \frac{v^2}{2}\sum_{i=1}^{n}\frac{\sigma^2}{n^2} \right).$$

Отсюда видно, что

$$\bar{x} \in N(a, \frac{\sigma^2}{n}),$$

$$\frac{\bar{x}-a}{\sigma}\sqrt{n} \in N(0,1).$$

Доказательство утверждения (3.3). Поскольку

$$(x_i - a) \in N(0, \sigma^2),$$

то

$$v = \sum_{i=1}^{n} \frac{(x_i - a)^2}{\sigma^2} = \sum_{i=1}^{n} u_i^2,$$

где $u_i = \dfrac{(x_i - a)}{\sigma} \in N(0,1)$. По определению распределения хи-квадрат (раздел 3.1) получаем утверждение (3.3).

Доказательство утверждения (3.4). Для доказательства нам понадобятся понятие ортогональной матрицы и лемма Фишера.

Квадратная матрица $C = (c_{i,j})$, $i, j = \overline{1,n}$, называется ортогональной, если $CC^T = I$. Отсюда следует, что $C^T = C^{-1}$, $C^T C = I$, матрица C^T также ортогональная, и справедливы равенства

$$\sum_{k=1}^{n} c_{i,k} c_{j,k} = \begin{cases} 1 & при\ \ i = j, \\ 0 & при\ \ i \neq j, \end{cases} \qquad (3.6)$$

$$\sum_{k=1}^{n} c_{k,i} c_{k,j} = \begin{cases} 1 & при\ \ i = j, \\ 0 & при\ \ i \neq j. \end{cases}$$

Если дано произвольное число $p < n$ строк матрицы C, для которых выполняются соотношения (3.6), то можно всегда найти еще $n - p$ строк, таких, что, добавляя их, мы получим ортогональную матрицу размером $n \times n$. Линейное преобразование $Y = CX$, где $X^T = (x_1, x_2, ..., x_n)$, $Y^T = (y_1, y_2, ..., y_n)$, называется ортогональным преобразованием. Квадратичная форма $X^T X = x_1^2 + x_2^2 + ... + x_n^2$ инвариантна относительно такого преобразования, т.е. она преобразуется в форму $Y^T C^T CY = Y^T Y = y_1^2 + y_2^2 + ... + y_n^2$ с той же матрицей I.

Лемма Фишера [12]. Пусть $X^T = (x_1, x_2, ..., x_n)$ – вектор из независимых случайных величин x_i, каждая из которых имеет нормальное распределение $N(0, \sigma^2)$, и $Y = CX$ – ортогональное преобразование, вводящее вместо $x_1, x_2, ..., x_n$ новые величины $y_1, y_2, ..., y_n$. Если задано лишь некоторое число $p < n$ случайных величин $y_1, y_2, ..., y_p$ этого преобразования, то величина

$$Q = \sum_{i=1}^{n} x_i^2 - y_1^2 - y_2^2 - ... - y_p^2 \qquad (3.7)$$

независима от $y_1, y_2, ..., y_p$ и Q/σ^2 имеет распределение $H_1(n-p)$, т.е. распределение хи-квадрат с $n-p$ степенями свободы.

Доказательство леммы Фишера. В силу изложенного выше относительно ортогонального преобразования величины y_i некоррелированы и нормальны с распределением $N(0,\sigma^2)$, следовательно, и независимы. Если задано лишь некоторое число $p < n$ случайных величин $y_1, y_2, ..., y_p$, удовлетворяющих ортогональному преобразованию, то, как указано выше, можно найти еще $n-p$ строк матрицы C с номерами $p+1$, $p+2$, …, n, таких, что полная матрица C будет ортогональной. Если к переменным $x_1, x_2, ..., x_n$, $y_1, y_2, ..., y_p$ из квадратичной формы (3.7) применить указанное ортогональное преобразование C, то $\sum_{i=1}^{n} x_i^2$ преобразуется в $\sum_{i=1}^{n} y_i^2$, и вместо (3.7) мы получим

$$Q = y_{p+1}^2 + y_{p+2}^2 + ... + y_n^2.$$

Таким образом, Q равна сумме квадратов $n-p$ независимых нормальных $N(0,\sigma^2)$ величин, которые, кроме того, не зависят от $y_1, y_2, ..., y_p$. По определению распределения хи-квадрат (раздел 3.1) делаем вывод о том, что величина Q/σ^2 имеет распределение $H_1(n-p)$. Лемма Фишера доказана.

Перейдем к доказательству утверждения (3.4). Предположим, что среднее значение генеральной совокупности равно нулю, т.е. каждое x_i нормально $N(0,\sigma^2)$. Это предположение не ограничивает общности, так как не меняет $\bar{s}^2$. Рассмотрим тождество

$$Q = n\bar{s}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - y_1^2. \qquad (3.8)$$

где $y_1^2 = n\bar{x}^2$. Тогда

$$y_1 = \sqrt{n}\,\bar{x} = (\sqrt{n}\,x_1 + \sqrt{n}\,x_2 + ... + \sqrt{n}\,x_n) = (c_{1,1}\,x_1 + c_{1,2}\,x_2 + ... + c_{1,n}\,x_n),$$

где $c_{1,k} = \sqrt{n}$. Мы видим, что y_1 есть линейное преобразование переменных $x_1, x_2, ..., x_n$, а $(c_{1,1}, c_{1,2}, ...c_{1,n})$ есть первая строка матрицы C этого преобразования. Поскольку $\sum_{k=1}^{n} c_{i,k} c_{i,k} = 1$, т.е. удовлетворяется равенство (3.6), то это преобразование C можно дополнить до ортогонального преобразования. Поэтому к выражению (3.8) можно применить лемму Фишера, положив в (3.7) $p = 1$ и $y_1 = \sqrt{n}\,\bar{x}$. В итоге получаем, что величина $n\bar{s}^2 / \sigma^2$ имеет распределение хи-квадрат с $(n-1)$ степенями свободы.

Доказательство утверждения (3.5). Исходя из статистик $u \in N(0,1)$ (3.1) и $v \in H_1(n-1)$ (3.4) мы можем записать, что

$$t = \frac{u}{\sqrt{v}} \sqrt{n-1} \in T_1(n-1).$$

Подставляя сюда выражения для u и v, получим утверждение (3.5).

4 ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ

4.1 Постановка задачи

Доверительным интервалом для некоторого параметра θ называется интервал $(\theta_н,\theta_в)$, накрывающий параметр θ с доверительной вероятностью γ:

$$P(\theta_н < \theta < \theta_в) = \gamma . \qquad (4.1)$$

Величина γ называется доверительной вероятностью или коэффициентом доверия, $\theta_н$, $\theta_в$ – нижним и верхним доверительными пределами соответственно, а разность $(\theta_в - \theta_н)$ – длиной доверительного интервала. Доверительная вероятность γ выбирается близкой к 1 из набора чисел $\{0{,}9;\ 0{,}95;\ 0{,}975\}$. Доверительный интервал для параметра θ называют также интервальной оценкой этого параметра.

Задача построения доверительного интервала для параметра распределения формулируется следующим образом. Известна плотность вероятности генеральной совокупности с точностью до параметра θ, то есть известна $f_\xi(x,\theta)$. Требуется по выборке $x_1, x_2, ..., x_n$ из этой совокупности найти интервал $(\theta_н,\theta_в)$ вида (4.1). Поскольку доверительный интервал строится по выборке, то доверительные пределы будут различными для различных выборок, то есть величины $\theta_н$, $\theta_в$ являются случайными.

Поскольку две границы доверительного интервала определяются из одного уравнения (1), то существует бесконечное множество интервалов, удовлетворяющих этому уравнению.

Для придания задаче однозначности от уравнения (4.1) переходят к двум уравнениям:

$$\begin{cases} P(\theta > \theta_в) = \alpha_1, \\ P(\theta < \theta_н) = \alpha_2, \end{cases} \tag{4.2}$$

где $\alpha_1 + \alpha_2 = 1 - \gamma$.

Доверительный интервал называется симметричным, если в (4.2) $\alpha_1 = \alpha_2 = \alpha = (1 - \gamma)/2$. Симметричный интервал строится на основании следующей системы уравнений:

$$\begin{cases} P(\theta > \theta_в) = \dfrac{1-\gamma}{2}, \\ P(\theta < \theta_н) = \dfrac{1-\gamma}{2}. \end{cases} \tag{4.3}$$

Чаще всего строят симметричные доверительные интервалы.

4.2 Методика построения симметричного доверительного интервала

Для построения симметричного доверительного интервала для неизвестного параметра θ обычно используется статистика, представляющая собой некоторую функцию $g = g(\hat{\theta})$ точечной оценки $\hat{\theta}$. Должен быть известен закон распределения этой статистики, например, плотность вероятности $f_g(x)$. В силу того, что параметр θ нам неизвестен, статистика $g = g(\hat{\theta})$ оказывается зависящей также от параметра θ, то есть $g = g(\hat{\theta}, \theta)$. Для построения доверительного интервала для параметра θ строят "доверительный интервал" для статистики g, то есть находят $g_н$ и $g_в$ из системы уравнений

$$\begin{cases} P(g > g_в) = \dfrac{1-\gamma}{2}, \\ P(g < g_н) = \dfrac{1-\gamma}{2}, \end{cases} \tag{4.4}$$

где γ – доверительная вероятность. Уравнения (4.4) иллюстрируются с помощью рис. 4.1, на котором приведен график плотности вероятности $f_g(x)$ и обозначены величины $g_н$, $g_в$, $\frac{1-\gamma}{2}$.

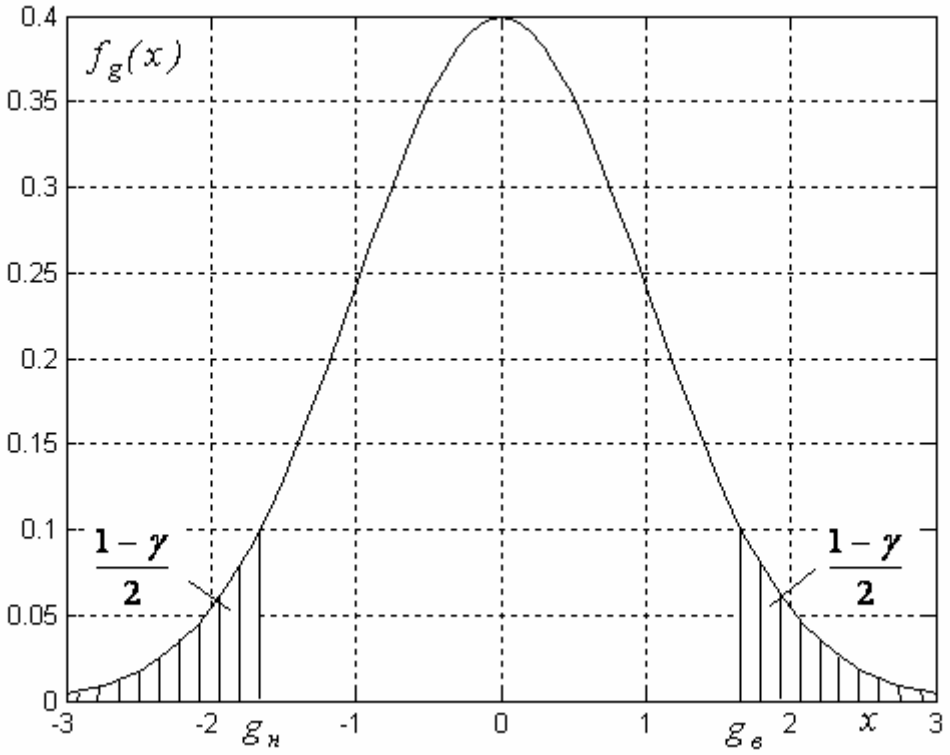

Рис. 4.1. Иллюстрация методики построения доверительного интервала

Решение задачи (4.4) эквивалентно решению задачи (4.3) и в итоге – задачи (4.1). Действительно, в силу зависимости $g = g(\hat{\theta}, \theta)$ интервал вида

$$P(g_н < g(\hat{\theta}, \theta) < g_в) = \gamma$$

можно преобразовать к виду (4.1).

В разделах 4.3 – 4.7 по изложенной методике получены доверительные интервалы для параметров нормального распределения и вероятности появления случайного события A.

4.3 Доверительный интервал для математического ожидания нормальной генеральной совокупности при известной дисперсии

Пусть известна выборка $x_1, x_2, ..., x_n$ из генеральной совокупности $N(a, \sigma^2)$, и требуется построить доверительный интервал для математического ожидания a при условии, что дисперсия σ^2 по каким-то соображениям нам известна.

Воспользуемся изложенной в п. 4.2 методикой. Возьмем точечную оценку $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ параметра a. Для построения доверительного интервала выберем статистику $u = \frac{\bar{x} - a}{\sigma}\sqrt{n}$, которая, как известно, имеет распределение $N(0,1)$. Найдем $u_н$, $u_в$ из уравнений

$$\begin{cases} P(u > u_в) = \dfrac{1-\gamma}{2}, \\ P(u < u_н) = \dfrac{1-\gamma}{2}. \end{cases}$$

Эти уравнения можно проиллюстрировать с помощью рис. 4.1 при замене g на u. В силу симметричности распределения $N(0,1)$ имеем $u_н = -u_в$. Обозначим $u_в = u_{\frac{1-\gamma}{2}}$ и напомним, что $u_в$ – это $100\frac{1-\gamma}{2}$-процентное отклонение распределения $N(0,1)$ (таблица 4.1). Более подробную таблицу можно найти в [13].

Таблица 4.1

Процентные отклонения нормального распределения $N(0,1)$, $P(u > u_\alpha) = \alpha$

α	0,0010	0,005	0,010	0,015	0,020	0,030	0,040	0,050
u_α	3,0902	2,5758	2,3264	2,1701	2,0938	1,8808	1,7507	1,6449

В итоге получим

$$P\left(-u_{\frac{1-\gamma}{2}} < u < u_{\frac{1-\gamma}{2}}\right) = \gamma,$$

или

$$P\left(-u_{\frac{1-\gamma}{2}} < \frac{\bar{x}-a}{\sigma}\sqrt{n} < u_{\frac{1-\gamma}{2}}\right) = \gamma,$$

или

$$P\left(\bar{x} - u_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{n}} < a < \bar{x} + u_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{n}}\right) = \gamma.$$

Из последнего выражения делаем вывод, что доверительный интервал для a имеет вид

$$\bar{x} - u_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{n}} < a < \bar{x} + u_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{n}}.$$

Длина доверительного интервала

$$l = 2u_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{n}}.$$

4.4 Доверительный интервал для математического ожидания нормальной генеральной совокупности при неизвестной дисперсии

Пусть имеется выборка $x_1, x_2, ..., x_n$ из генеральной совокупности $N(a, \sigma^2)$, и требуется построить доверительный интервал для математического ожидания a при условии, что дисперсия σ^2 неизвестна.

Для построения доверительного интервала в данном случае используется статистика

$$t = \frac{\bar{x}-a}{\bar{s}}\sqrt{n-1} \in T_1(n-1)$$

(см. п. 3.4). Аналогично предыдущему разделу получаем следующий доверительный интервал

$$\bar{x} - t_{\frac{1-\gamma}{2}} \frac{\bar{s}}{\sqrt{n-1}} < a < \bar{x} + t_{\frac{1-\gamma}{2}} \frac{\bar{s}}{\sqrt{n-1}}$$

и его длину

$$l = 2t_{\frac{1-\gamma}{2}} \frac{\bar{s}}{\sqrt{n-1}},$$

где $t_{\frac{1-\gamma}{2}}$ – $100\dfrac{1-\gamma}{2}$-процентное отклонение распределения $T_1(n-1)$ (Стьюдента с $n-1$ степенями свободы). Графическая иллюстрация в данном случае аналогична приведенной на рис. 4.1 с заменой g на t.

Если мы воспользуемся соотношением $\bar{s}^2 = (n-1)s^2 / n$, то получим доверительный интервал в виде

$$\bar{x} - t_{\frac{1-\gamma}{2}} \frac{s}{\sqrt{n}} < a < \bar{x} + t_{\frac{1-\gamma}{2}} \frac{s}{\sqrt{n}}.$$

4.5 Доверительный интервал для дисперсии нормальной генеральной совокупности при известном математическом ожидании

Пусть имеется выборка $x_1, x_2, ..., x_n$ из генеральной совокупности $N(a, \sigma^2)$, и требуется построить доверительный интервал для дисперсии σ^2 при условии, что математическое ожидание a известно.

Для построения данного доверительного интервала используется статистика (см. (3.3))

$$v = \frac{n\bar{s}_0^2}{\sigma^2} \in H_1(n).$$

Найдем $v_н$, $v_в$ из уравнений

$$\begin{cases} P(v > v_в) = \dfrac{1-\gamma}{2}, \\ P(v < v_н) = \dfrac{1-\gamma}{2}. \end{cases}$$

Чтобы решить с помощью существующих таблиц процентных отклонений распределения хи-квадрат второе из этих уравнений, приведем его к виду первого уравнения:

$$P(v > v_н) = \frac{1+\gamma}{2}.$$

Обозначим $v_в = v_{\frac{1-\gamma}{2}}$, $v_н = v_{\frac{1+\gamma}{2}}$, где $v_{\frac{1-\gamma}{2}}$, $v_{\frac{1+\gamma}{2}}$ – $100\dfrac{1-\gamma}{2}$- и $100\dfrac{1+\gamma}{2}$-процентные отклонения распределения $H_1(n)$ соответственно (рис. 4.2).

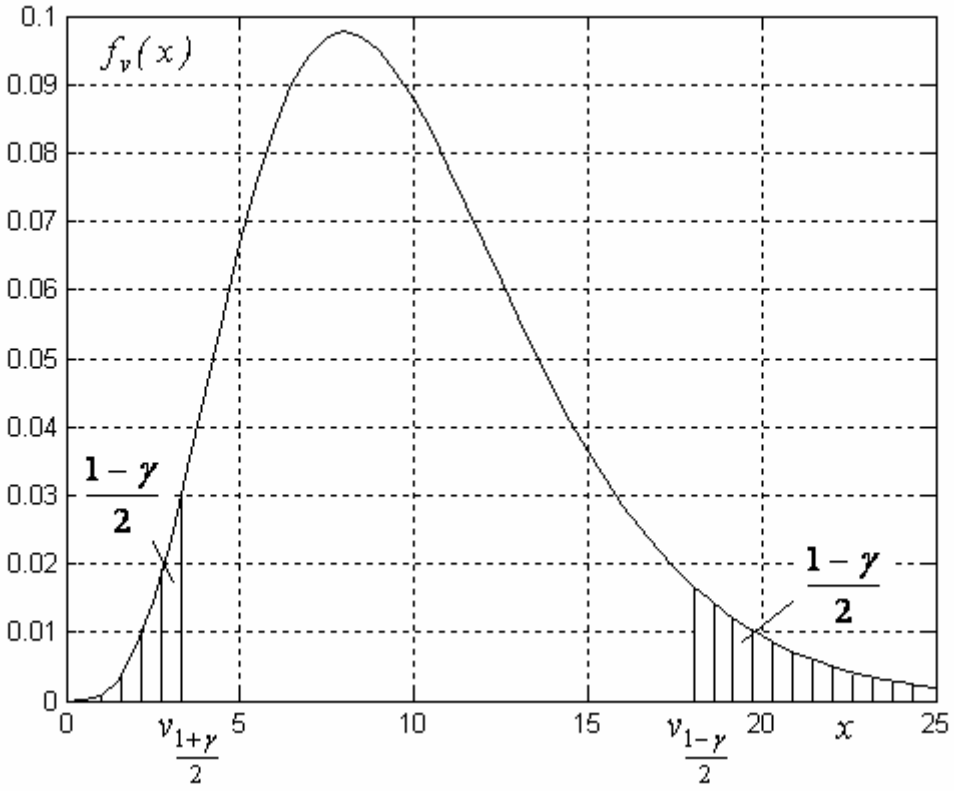

Рис. 4.2. Иллюстрация методики построения доверительного интервала для дисперсии нормальной генеральной совокупности

В итоге получим

$$P\left(v_{\frac{1+\gamma}{2}} < v < v_{\frac{1-\gamma}{2}} \right) = \gamma ,$$

или

$$P\left(v_{\frac{1+\gamma}{2}} < \frac{n\bar{s}_0^2}{\sigma^2} < v_{\frac{1-\gamma}{2}} \right) = \gamma ,$$

или

$$P\left(\frac{n\bar{s}_0^2}{v_{\frac{1-\gamma}{2}}} < \sigma^2 < \frac{n\bar{s}_0^2}{v_{\frac{1+\gamma}{2}}} \right) = \gamma .$$

Из последнего выражения получаем доверительный интервал для σ^2:

$$\frac{n\bar{s}_0^2}{v_{\frac{1-\gamma}{2}}} < \sigma^2 < \frac{\overline{n}\bar{s}_0^2}{v_{\frac{1+\gamma}{2}}} .$$

4.6 Доверительный интервал для дисперсии нормальной генеральной совокупности при неизвестном математическом ожидании

Пусть имеется выборка $x_1, x_2, ..., x_n$ из генеральной совокупности $N(a, \sigma^2)$, и требуется построить доверительный интервал для дисперсии σ^2 при условии, что математическое ожидание a неизвестно.

Для построения доверительного интервала воспользуемся статистикой (см. (4.4))

$$w = \frac{n\bar{s}^2}{\sigma^2} \in H_1(n-1) .$$

Повторяя выкладки предыдущего раздела с заменой v на w, получим следующий доверительный интервал:

$$\frac{n\bar{s}^2}{w_{\frac{1-\gamma}{2}}} < \sigma^2 < \frac{n\bar{s}^2}{w_{\frac{1+\gamma}{2}}},$$

где $w_{\frac{1-\gamma}{2}}$, $w_{\frac{1+\gamma}{2}}$ – $100\frac{1-\gamma}{2}$- и $100\frac{1+\gamma}{2}$-процентные отклонения распределения $H_1(n-1)$ соответственно. Графическая иллюстрация методики построения данного интервала та же, что и на рис. 4.2, с заменой v на w.

Если воспользоваться соотношением $\bar{s}^2 = (n-1)s^2/n$, то получим доверительный интервал в виде

$$\frac{(n-1)s^2}{w_{\frac{1-\gamma}{2}}} < \sigma^2 < \frac{(n-1)s^2}{w_{\frac{1+\gamma}{2}}}.$$

4.7 Доверительный интервал для вероятности случайного события

Пусть A – случайное событие, имеющее вероятность $P(A) = p$. Выполним над этим событием n независимых испытаний Бернулли и зафиксируем число m появлений события A в n испытаниях. Построим по этим данным доверительный интервал для вероятности p события A.

Для построения доверительного интервала воспользуемся относительной частотой события $\hat{p} = \frac{m}{n}$, которая, в соответствии с теоремой Бернулли, является несмещенной и состоятельной оценкой вероятности p события A. На основании локальной предельной теоремы Муавра-Лапласа [14] при большом n имеем

$$\hat{p} \in N\left(p, \sqrt{\frac{pq}{n}}\right),$$

$$u = \frac{\hat{p} - p}{\sqrt{pq}}\sqrt{n} \in N(0,1).$$

Статистика u пригодна для построения доверительного интервала. По изложенной в п. 4.2 методике можно получить следующие границы доверительного интервала:

$$\frac{n}{n + u^2_{\frac{1-\gamma}{2}}}\left(\hat{p} + \frac{u^2_{\frac{1-\gamma}{2}}}{2n} \pm u_{\frac{1-\gamma}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{u^2_{\frac{1-\gamma}{2}}}{4n^2}} \right),$$

где $\hat{q} = 1 - \hat{p}$, $u_{\frac{1-\gamma}{2}}$ – $100\dfrac{1-\gamma}{2}$-процентное отклонение распределения $N(0,1)$, определяемое по таблицам распределения $N(0,1)$ (см. таблицу 4.1 раздела 4.3). Знак минус в этой формуле соответствует нижнему доверительному пределу, а знак плюс – верхнему.

5 СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

5.1 Понятие статистической гипотезы. Классификация гипотез

Статистической гипотезой называется любое непротиворечивое множество утверждений

$$H = \{H_0, H_1, ..., H_{k-1}\}$$

относительно распределения генеральной совокупности. Такая гипотеза называется k-альтернативной. Каждое утверждение гипотезы H_i, $i = \overline{0, k-1}$, называется альтернативой k-альтернативной гипотезы или также гипотезой.

Проверить гипотезу – значит по выборке $x_1, x_2, ..., x_n$ из генеральной совокупности принять обоснованное решение о том, какая из альтернатив является истинной.

Если в результате проверки гипотезы какая-то из альтернатив принята, то другие альтернативы отклоняются, то есть считаются ложными.

Гипотеза проверяется на основе так называемого критерия проверки гипотезы. Критерий – это правило, позволяющее принять или отклонить ту или иную альтернативу по имеющейся выборке. Обычно принимают или отклоняют нулевую гипотезу H_0.

Альтернатива H_i называется параметрической, если она задает значение некоторого параметра θ распределения. В противном случае она называется непараметрической.

Многоальтернативная гипотеза H называется параметрической, если все ее альтернативы параметрические, и непараметрической, если хотя бы одна альтернатива непараметрическая.

Альтернатива H_i называется простой, если она однозначно определяет распределение генеральной совокупности, и сложной в противном случае.

Многоальтернативная гипотеза H называется простой, если все ее альтернативы простые, и сложной, если хотя бы одна из альтернатив сложная.

Приведем примеры гипотез с их классификацией. Пусть выборка извлекается из нормальной совокупности $N(a,\sigma^2)$ и a_0, a_1, σ_0^2 – некоторые фиксированные числа. Сформулируем следующие гипотезы:

1. $\{H_0 : a = a_0;\ H_1 : a = a_1\}$.

Это двухальтернативная параметрическая простая гипотеза о параметре a нормальной генеральной совокупности.

2. $\{H_0 : a = a_0;\ H_1 : a \neq a_1\}$.

Это двухальтернативная параметрическая сложная гипотеза, так как H_1 – сложная.

3. $\{H_0 : \sigma^2 = \sigma_0^2;\ H_1 : \sigma^2 > \sigma_0^2\}$.

Это двухальтернативная параметрическая сложная гипотеза о параметре σ^2 нормальной генеральной совокупности.

Пусть $f_0(x)$ – некоторая полностью известная плотность вероятности и $f_\xi(x)$ – плотность вероятности генеральной совокупности. Гипотеза вида

4. $\{H_0 : f_\xi(x) = f_0(x);\ H_1 : f_\xi(x) \neq f_0(x)\}$

является двухальтернативной непараметрической сложной гипотезой – так называемой гипотезой о законе распределения. Здесь проверяется утверждение о том, что наша выборка извлечена из распределения $f_0(x)$.

5.2 Критерий значимости

Пусть проверяется двухальтернативная сложная гипотеза $\{H_0, H_1\}$, где H_0 – простая гипотеза, а H_1 – сложная. Такая гипотеза проверяется с помощью так называемого критерия значимости.

В основе критерия значимости лежит некоторая скалярная статистика $g = g(x_1,...,x_n)$, которая представляет собой отклонение эмпирических (выборочных) данных от гипотетических.

Пусть $f_g(x)$ – плотность вероятности статистики g. Эта плотность вероятности предполагается известной (при условии, что H_0 верна). Критерий значимости имеет вид

$$P(|g| > g_{\alpha/2}) = \alpha, \tag{5.1}$$

или

$$P(g > g_\alpha) = \alpha, \tag{5.2}$$

или

$$P(g < g_\alpha) = \alpha, \tag{5.3}$$

где α – вероятность, которая выбирается из следующего набора малых чисел: $\{0{,}1;\ 0{,}05;\ 0{,}025;\ 0{,}01\}$. Событие, имеющее такую вероятность, можно считать практически невозможным, то есть не появляющимся в результате одного эксперимента. Величины $g_{\alpha/2}$, g_α называются пределами значимости или критическими значениями, α – уровнем значимости. Области, определяемые условиями $|g| > g_{\alpha/2}$, или $g > g_\alpha$, или $g < g_\alpha$, называются критическими областями. Эти области отмечены на рис. 5.1 – 5.3 штриховкой.

Критерий (5.1) называется двухсторонним или критерием с двухсторонней критической областью. Критерий (5.2) – правосторонний. Критерий (5.3) – левосторонний. Гипотеза проверяется следующим образом. Выбирается уровень значимости α. По таблицам распределения статистики g определяется предел значимости $g_{\alpha/2}$ или g_α, в зависимости от вида критерия. Затем по имеющейся выборке и формуле для статистики g подсчитывается эмпирическое значение статистики $g_э$. Если окажется, что $|g_э| > g_{\alpha/2}$ для двухстороннего критерия (5.1), или $g_э > g_\alpha$ для правостороннего критерия (5.2), или $g_э < g_\alpha$ для левостороннего критерия (5.3), то проверяемая гипотеза H_0 отклоняется.

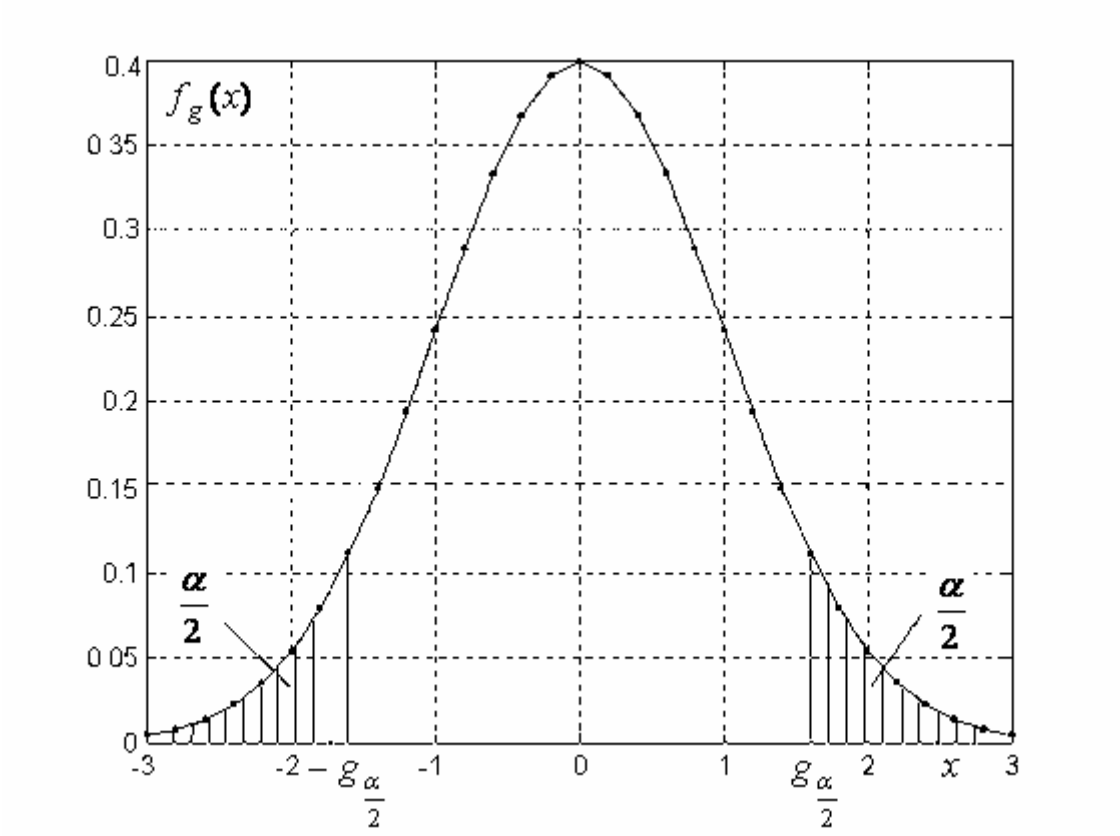

Рис. 5.1. Критические области для двухстороннего критерия значимости

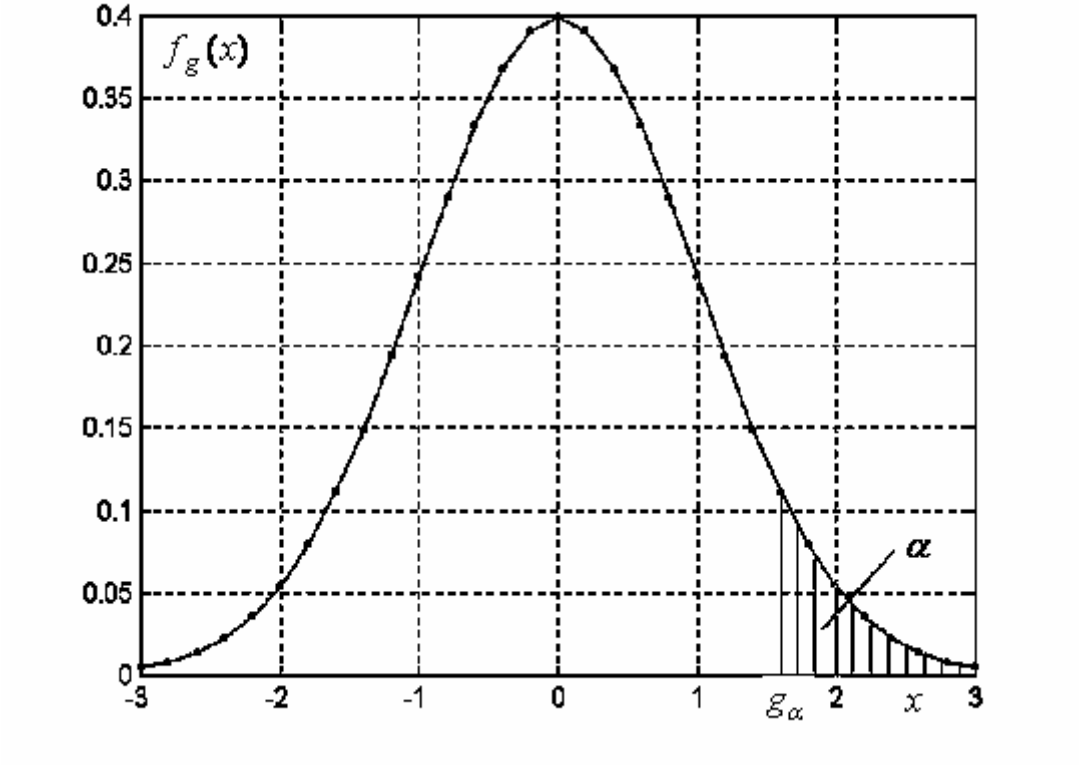

Рис. 5.2. Критическая область для правостороннего критерия значимости

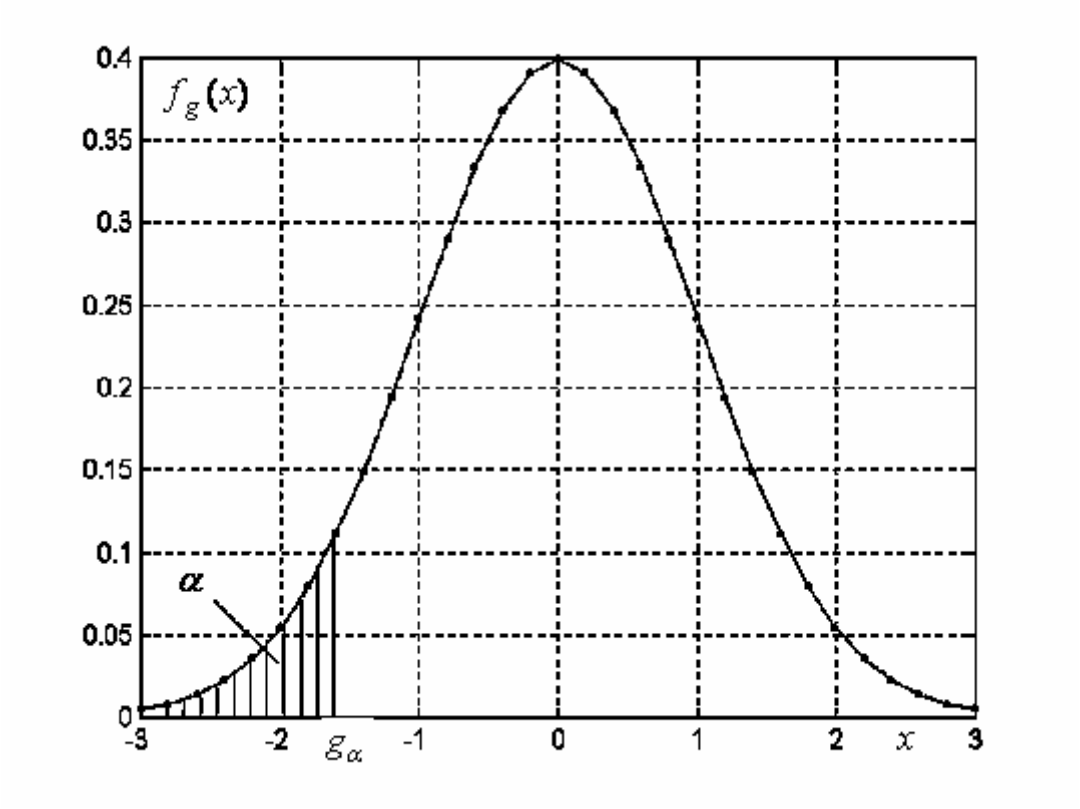

Рис. 5.3. Критическая область для левостороннего критерия значимости

Иначе говоря, если эмпирическое значение статистики $g_э$ попадает в критическую область, то проверяемая гипотеза H_0 отклоняется.

Гипотеза H_0 отклоняется в силу того, что имеется противоречие между гипотетическими и эмпирическими данными, которое проявилось в том, что произошло событие, которое не должно было произойти в результате единичного эксперимента.

Критерий значимости позволяет установить, значимо ли отклонение эмпирических данных от гипотетических, то есть значимо ли значение статистики g. Отсюда происходит название критерия.

Отметим, что поскольку мы используем распределение статистики g при условии, что H_0 верна, то α представляет собой вероятность отклонения верной гипотезы H_0 и называется также вероятностью ошибки 1-го рода. Вероятность принятия верной гипотезы H_0 равна $(1-\alpha)$.

5.3 Проверка гипотезы о законе распределения

Пусть по выборке $x_1,...,x_n$ из некоторой генеральной совокупности нужно проверить гипотезу о том, что генеральная совокупность имеет заданное распределение. Критерии для проверки такой гипотезы получили название критериев согласия.

5.3.1 Критерий согласия хи-квадрат (Пирсона)

Пусть $f_\xi(x)$ – плотность вероятности генеральной совокупности, $f_0(x,\theta_1,...,\theta_m)$ – гипотетическая плотность вероятности, известная с точностью до m параметров $\theta_1,...,\theta_m$, причем m может быть равным нулю. Требуется проверить двухальтернативную непараметрическую сложную гипотезу

$$\{H_0 : f_\xi(x) = f_0(x,\theta_1,...,\theta_m),\ H_1 : f_\xi(x) \neq f_0(x,\theta_1,...,\theta_m)\}.$$

Это гипотеза о том, что выборка извлечена из распределения $f_0(x,\theta_1,...,\theta_m)$. Для проверки этой гипотезы критерием χ^2 множество возможных значений случайной величины ξ разбивается на l интервалов Δ_i и подсчитывается количество выборочных значений m_i, попавших в каждый интервал (как при построении гистрограммы). Для проверки гипотезы используется статистика

$$v = \sum_{i=1}^{l} \frac{(m_i - n\hat{p}_i)^2}{n\hat{p}_i}, \qquad (5.4)$$

где $\hat{p}_i$ – гипотетическая вероятность попадания случайной величины ξ в i-й интервал. Она определяется по формуле

$$\hat{p}_i = \int_{\Delta_i} f_0(x,\hat{\theta}_1,...,\hat{\theta}_m)dx.$$

Интегрирование в этой формуле осуществляется по *i*-му интервалу Δ_i. Здесь $f_0(x,\widehat{\theta}_1,...,\widehat{\theta}_m)$ – гипотетическая плотность вероятности, в которую вместо неизвестных параметров подставлены их МП-оценки $\widehat{\theta}_1,...,\widehat{\theta}_m$.

В случае истинности гипотезы H_0 статистика (5.4) имеет распределение, которое при $n \to \infty$ приближается к распределению $H_1(l-m-1)$ (хи-квадрат с $(l-m-1)$ степенями свободы).

Критерий значимости для проверки этой гипотезы – это правосторонний критерий вида

$$P(v > v_\alpha) = \alpha,$$

где v_α – 100α-процентное отклонение распределения $H_1(l-m-1)$.

Если гипотетическая плотность вероятности известна полностью, то необходимо считать $m = 0$, то есть пользоваться таблицами распределения $H_1(l-1)$.

5.3.2 Критерий согласия λ (Колмогорова)

Проверяется гипотеза

$$H_0 : F_\xi(x) = F_0(x)$$

против альтернативы

$$H_1 : F_\xi(x) \neq F_0(x),$$

где $F_\xi(x)$ – функция распределения генеральной совокупности, $F_0(x)$ – непрерывная гипотетическая функция распределения (полностью известная функция).

Для проверки гипотезы используется статистика

$$\lambda = \Delta\sqrt{n}, \tag{5.5}$$

где

$$\Delta = \max_x | F_0(x) - F_\xi^*(x) | -$$

максимальный модуль отклонения гипотетической функции распределения $F_0(x)$ от эмпирической функции распределения $F_\xi^*(x)$.

Если гипотеза H_0 верна, то статистика λ (5.5) имеет распределение, приближающееся при $n \to \infty$ к распределению Колмогорова. Критерий для проверки гипотезы имеет следующий вид:

$$P(\lambda > \lambda_\alpha) = \alpha,$$

где $\lambda_\alpha - 100\alpha$ -процентное отклонение распределения Колмогорова (табл. 5.1).

Таблица 5.1

Процентные отклонения распределения Колмогорова, $P(\lambda > \lambda_\alpha) = \alpha$

α	0.01	0.02	0.03	0.04	0.05
λ_α	1.627	1.520	1.45	1.40	1.358

5.3.3 Критерий согласия ω^2 (Мизеса-Смирнова)

Проверяется двухальтернативная гипотеза

$$\{H_0 : F_\xi(x) = F_0(x);\ \ H_1 : F_\xi(x) \neq F_0(x)\}.$$

Количественной мерой отклонения эмпирических данных от гипотетических служит величина [2, 15]

$$\omega^2 = \int_{-\infty}^{\infty} \left(F_\xi^*(x) - F_0(x)\right)^2 dF_0(x), \qquad (5.6)$$

где $F_\xi^*(x)$ – эмпирическая функция распределения. Получим выражение для численного расчета статистики ω^2 в предположении, что гипотетическая функция распределения $F_0(x)$ непрерывна и существует производная $F_0'(x)$

(плотность вероятности). Учитывая выражение (1.2) для эмпирической функции распределения, разобьем действительную прямую на интервалы

$$(-\infty, x_{(1)}),\ [x_{(1)}, x_{(2)}),\ \ldots,\ [x_{(n-1)}, x_{(n)}),\ [x_{(n)}, +\infty),$$

где $x_{(k)}$ – порядковая статистика, и вычислим интеграл (5.6) по этим интервалам. Получим

$$\omega^2 = \int_{-\infty}^{x_{(1)}} (F_0(x))^2\, dF_0(x) + \sum_{k=1}^{n-1} \int_{x_{(k)}}^{x_{(k+1)}} \left(\frac{k}{n} - F_0(x)\right)^2 dF_0(x) + \int_{x_{(n)}}^{\infty} (1 - F_0(x))^2\, dF_0(x) =$$

$$= \frac{(F_0(x))^3}{3}\Bigg|_{-\infty}^{x_{(1)}} + \sum_{k=1}^{n-1} \frac{(F_0(x) - k/n)^3}{3}\Bigg|_{x_{(k)}}^{x_{(k+1)}} - \frac{(1 - F_0(x))^3}{3}\Bigg|_{x_{(n)}}^{\infty} =$$

$$= \frac{F_0^3(x_{(1)})}{3} + \sum_{q=1}^{n-1} \frac{(F_0(x_{(q+1)}) - q/n)^3}{3} - \sum_{k=1}^{n-1} \frac{(F_0(x_{(k)}) - k/n)^3}{3} + \frac{(1 - F_0(x_{(n)}))^3}{3}.$$

Легко заметить, что первое слагаемое в последнем выражении можно включить в первую сумму, а последнее слагаемое – во вторую сумму, то есть записать

$$\omega^2 = \sum_{q=0}^{n-1} \frac{(F_0(x_{(q+1)}) - q/n)^3}{3} - \sum_{k=1}^{n} \frac{(F_0(x_{(k)}) - k/n)^3}{3}.$$

Если ввести в первой сумме новую переменную суммирования $k = q + 1$, то получим

$$\omega^2 = \sum_{k=1}^{n} \frac{(F_0(x_{(k)}) - (k-1)/n)^3}{3} - \sum_{k=1}^{n} \frac{(F_0(x_{(k)}) - k/n)^3}{3}.$$

Обозначив $V = F_0(x_{(k)}) - k/n$, будем иметь

$$\omega^2 = \sum_{k=1}^{n} \frac{(V + 1/n)^3}{3} - \sum_{k=1}^{n} \frac{V^3}{3} = \sum_{k=1}^{n} \frac{(V + 1/n)^3 - V^3}{3} =$$

$$= \sum_{k=1}^{n} \frac{3V^2/n + 3V/n^2 + 1/n^3}{3} = \frac{1}{n} \sum_{k=1}^{n} \left(V^2 + \frac{V}{n} + \frac{1}{3n^2}\right).$$

Дополняя выражение под знаком суммы до полного квадрата по V, получим

$$\omega^2 = \frac{1}{n}\sum_{k=1}^{n}\left(V^2 + 2\frac{V}{2n} + \frac{1}{4n^2} - \frac{1}{4n^2} + \frac{1}{3n^2}\right) = \frac{1}{n}\sum_{k=1}^{n}\left[\left(V + \frac{1}{2n}\right)^2 + \frac{1}{12n^2}\right].$$

Поскольку

$$V + \frac{1}{2n} = F_0(x_{(k)}) - \frac{k}{n} + \frac{1}{2n} = F_0(x_{(k)}) - \frac{2k-1}{2n},$$

то окончательно будем иметь

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n}\sum_{k=1}^{n}\left(F_0(x_{(k)}) - \frac{2k-1}{2n}\right)^2 .$$

Статистика критерия ω^2 имеет вид

$$z = n\omega^2 = \frac{1}{12n} + \sum_{k=1}^{n}\left(F_0(x_{(k)}) - \frac{2k-1}{2n}\right)^2 . \qquad (5.7)$$

Для статистики z (5.7) при $n \to \infty$ существует предельное распределение, для которого составлены таблицы процентных отклонений (табл. 5.2). Критерий ω^2 является правосторонним:

$$P(z > z_\alpha) = \alpha .$$

Таблица 5.2.

Процентные отклонения предельного распределения статистики z,

$$P(z > z_\alpha) = \alpha$$

α	0.01	0.02	0.03	0.04	0.05
z_α	0.74	0.62	0.55	0.50	0.46

5.4 Проверка гипотез о параметрах распределений

5.4.1 Проверка гипотезы о математическом ожидании нормальной генеральной совокупности при известной дисперсии

Имеется выборка $x_1, x_2, ..., x_n$ из нормального распределения $N(a, \sigma^2)$, причем дисперсия σ^2 известна, и нужно проверить гипотезу о математическом ожидании $H_0 : a = a_0$, где a_0 – известное число.

Для проверки гипотезы используется статистика

$$u = \frac{\bar{x} - a}{\sigma}\sqrt{n}.$$

Если H_0 верна, то есть $a = a_0$, то $u \in N(0,1)$. Для проверки гипотезы вида $\{H_0 : a = a_0;\ H_1 : a \neq a_0\}$ используется двухсторонний критерий значимости $P(|u| > u_{\alpha/2}) = \alpha$ (рис. 5.1), гипотезы $\{H_0 : a = a_0;\ H_1 : a > a_0\}$ – правосторонний критерий $P(u > u_\alpha) = \alpha$ (рис. 5.2), гипотезы $\{H_0 : a = a_0;\ H_1 : a < a_0\}$ – левосторонний критерий $P(u < u_\alpha) = \alpha$ (рис. 5.3). Здесь u_α – 100α-процентное отклонение распределения $N(0,1)$, $u_{\alpha/2}$ – $100\dfrac{\alpha}{2}$-процентное отклонение распределения $N(0,1)$ (таблица 4.1 раздела 4.3).

5.4.2 Проверка гипотезы о математическом ожидании нормальной генеральной совокупности при неизвестной дисперсии

Имеется выборка $x_1, x_2, ..., x_n$ из нормального распределения $N(a, \sigma^2)$, причем дисперсия σ^2 неизвестна, и нужно проверить гипотезу о математическом ожидании $H_0 : a = a_0$, где a_0 – известное число.

Для проверки гипотезы используется статистика

$$t = \frac{\bar{x} - a}{\bar{s}}\sqrt{n-1} = \frac{\bar{x} - a}{s}\sqrt{n}\,.$$

Если H_0 верна, то $t \in T_1(n-1)$. Для проверки гипотезы $\{H_0 : a = a_0;\ H_1 : a \neq a_0\}$ применяется двухсторонний критерий значимости $P(|t| > t_{\alpha/2}) = \alpha$ (рис. 5.1), гипотезы $\{H_0 : a = a_0;\ H_1 : a > a_0\}$ – правосторонний критерий $P(t > t_\alpha) = \alpha$ (рис. 5.2), гипотезы $\{H_0 : a = a_0;\ H_1 : a < a_0\}$ – левосторонний критерий $P(t < t_\alpha) = \alpha$ (рис. 5.3). Здесь t_α – 100α -процентное отклонение распределения $T_1(n-1)$, $t_{\alpha/2}$ – $100\dfrac{\alpha}{2}$-процентное отклонение распределения $T_1(n-1)$.

5.4.3 Проверка гипотезы о дисперсии нормальной генеральной совокупности при известном математическом ожидании

Выборка $x_1, x_2, ..., x_n$ извлечена из распределения $N(a, \sigma^2)$, математическое ожидание a известно, и нужно проверить гипотезу о дисперсии $H_0 : \sigma^2 = \sigma_0^2$, где σ_0^2 – известное число.

Для проверки гипотезы используется статистика

$$v = \frac{n\bar{s}_0^2}{\sigma^2}\,.$$

Если H_0 верна, то $v \in H_1(n)$. Для проверки гипотезы $\{H_0 : \sigma^2 = \sigma_0^2;\ H_1 : \sigma^2 \neq \sigma_0^2\}$ применяется двухсторонний критерий значимости, который в силу асимметрии распределения $H_1(n)$ имеет вид

$$\begin{cases} P(v > v_{\alpha/2}) = \dfrac{\alpha}{2}, \\ P(v < v_{(2-\alpha)/2}) = \dfrac{\alpha}{2}. \end{cases}$$

Критическая область критерия отмечена штриховкой на рис. 5.4. Величины $v_{\frac{\alpha}{2}}$, $v_{\frac{(2-\alpha)}{2}}$ – $100\dfrac{\alpha}{2}$- и $100\dfrac{2-\alpha}{2}$-процентные отклонения распределения $H_1(n)$.

Для проверки гипотезы $\{H_0 : \sigma^2 = \sigma_0^2;\ H_1 : \sigma^2 > \sigma_0^2\}$ используется правосторонний критерий $P(v > v_\alpha) = \alpha$, гипотезы $\{H_0 : \sigma^2 = \sigma_0^2;\ H_1 : \sigma^2 < \sigma_0^2\}$ – левосторонний критерий $P(v < v_\alpha) = \alpha$,

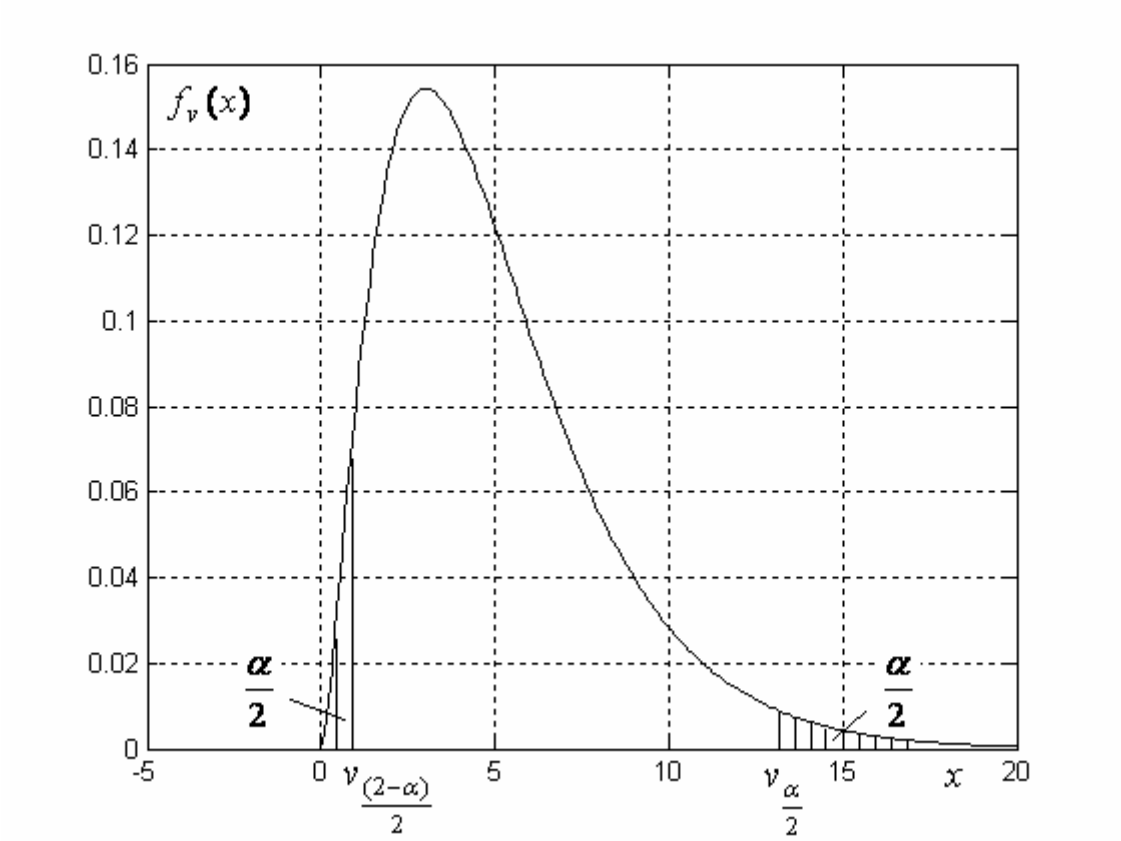

Рис. 5.4. Критическая область для гипотезы о дисперсии нормальной генеральной совокупности

5.4.4 Проверка гипотезы о дисперсии нормальной генеральной совокупности при неизвестном математическом ожидании

Выборка $x_1, x_2, ..., x_n$ извлечена из распределения $N(a, \sigma^2)$, математическое ожидание a неизвестно, и нужно проверить гипотезу о дисперсии $H_0 : \sigma^2 = \sigma_0^2$, где σ_0^2 – известное число.

Для проверки гипотезы используется статистика

$$w = \frac{n\bar{s}^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}.$$

Если H_0 верна, то $v \in H_1(n-1)$. Для проверки гипотезы $\{H_0 : \sigma^2 = \sigma_0^2;$ $H_1 : \sigma^2 \neq \sigma_0^2\}$ применяется двухсторонний критерий значимости, который в силу асимметрии распределения $H_1(n-1)$ имеет вид

$$\begin{cases} P(w > w_{\alpha/2}) = \alpha/2, \\ P(w < w_{(2-\alpha)/2}) = \alpha/2. \end{cases}$$

Иллюстрация данного критерия та же, что и на рис. 5.4, с заменой v на w. Величины $w_{\alpha/2}$, $w_{(2-\alpha)/2}$ – $100\frac{\alpha}{2}$- и $100\frac{2-\alpha}{2}$-процентные отклонения распределения $H_1(n-1)$.

Для проверки гипотезы $\{H_0 : \sigma^2 = \sigma_0^2;\ H_1 : \sigma^2 > \sigma_0^2\}$ используется правосторонний критерий $P(w > w_\alpha) = \alpha$, гипотезы $\{H_0 : \sigma^2 = \sigma_0^2;\ H_1 : \sigma^2 < \sigma_0^2\}$ – левосторонний критерий $P(w < w_\alpha) = \alpha$.

5.4.5 Проверка гипотезы о равенстве математических ожиданий двух нормальных генеральных совокупностей при неизвестных, но равных дисперсиях

Имеются две выборки разных объемов $x_1,...,x_m$ и $y_1,...,y_n$ из двух нормальных генеральных совокупностей $N(a_1,\sigma^2)$ и $N(a_2,\sigma^2)$ соответственно, в предположении, что дисперсии генеральных совокупностей равны между собой, но неизвестны. Требуется проверить гипотезу о том, что математические ожидания этих распределений равны, то есть проверить двухальтернативную параметрическую сложную гипотезу

$$\{H_0 : a_1 = a_2;\ H_1 : a_1 \neq a_2\}.$$

Для решения этой задачи рассмотрим t-критерий Стьюдента, который основан на сравнении выборочных средних. Пусть $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ – выборочное среднее первой генеральной совокупности, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – выборочное среднее второй генеральной совокупности. Как известно,

$$\bar{x} \in N(a_1, \frac{\sigma^2}{m}),\ \bar{y} \in N(a_2, \frac{\sigma^2}{n}).$$

Тогда

$$z = \bar{x} - \bar{y} \in N(a_1 - a_2, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}).$$

Если проверяемая гипотеза H_0 верна, то есть $a_1 = a_2$, то

$$z \in N(0, \frac{\sigma^2 (m+n)}{mn})$$

и

$$u = \frac{z}{\sqrt{D(z)}} = \frac{(\bar{x} - \bar{y})\sqrt{mn}}{\sigma\sqrt{m+n}} \in N(0,1).$$

Поскольку σ в данном выражении неизвестна, статистика u неприменима для проверки гипотезы, и мы продолжим поиск подходящей для этого статистики. Мы знаем, что

$$v_x = \frac{m\bar{s}_x^2}{\sigma^2} \in H_1(m-1), \qquad v_y = \frac{n\bar{s}_y^2}{\sigma^2} \in H_1(n-1).$$

Тогда статистика

$$v = v_x + v_y = \frac{m\bar{s}_x^2 + n\bar{s}_y^2}{\sigma^2} \in H_1(m+n-2).$$

Отметим, что v не зависит от u, поскольку v_x и v_y не зависят от u. Тогда статистика

$$t = \frac{u}{\sqrt{v}}\sqrt{m+n-2} \in T_1(m+n-2).$$

Подставляя сюда выражения для u и v, получим

$$t = \frac{(\bar{x}-\bar{y})\sqrt{mn(m+n-2)}}{\sqrt{m+n}\,\sqrt{m\bar{s}_x^2 + n\bar{s}_y^2}} \in T_1(m+n-2). \qquad (5.8)$$

Статистика t не содержит неизвестных параметров и известен ее закон распределения. Следовательно, она пригодна для проверки сформулированной гипотезы. Гипотеза проверяется следующим образом. Задавшись уравнением значимости α, по таблице процентных отклонений распределения $T_1(m+n-2)$ находим величину $t_{\alpha/2}$, удовлетворяющую равенству $p(|t| > t_{\alpha/2}) = \alpha$. Затем находим эмпирическое значение статистики $t_э$ по формуле (5.8). Если окажется, что $|t_э| > t_{\alpha/2}$, то гипотеза H_0 отклоняется в пользу гипотезы H_1.

5.4.6 Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей

Имеются две выборки разных объемов $x_1,...,x_m$ и $y_1,...,y_n$ из двух нормальных генеральных совокупностей $N(a_1,\sigma_1^2)$ и $N(a_2,\sigma_2^2)$ соответственно, в условиях, когда все параметры неизвестны. Требуется проверить гипотезу о том, что дисперсии этих распределений равны, то есть проверить двухальтернативную параметрическую гипотезу $\{H_0, H_1\}$, где

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Для проверки этой гипотезы используется статистика

$$f = \frac{s_x^2}{s_y^2},$$

где s_x^2, s_y^2 – несмещенные выборочные дисперсии

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 ,$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 .$$

Легко показать, что если H_0 верна, то $f \in F_1(m-1, n-1)$. Действительно, поскольку

$$v = \frac{(m-1)s_x^2}{\sigma_1^2} \in H_1(m-1),\ w = \frac{(n-1)s_y^2}{\sigma_2^2} \in H_1(n-1),$$

и $\sigma_1^2 = \sigma_2^2$, то

$$f = \frac{v}{m-1} : \frac{w}{n-1} = \frac{s_x^2}{s_y^2} \in F_1(m-1, n-1).$$

Для проверки гипотезы $\{H_0 : \sigma_1^2 = \sigma_2^2;\ H_1 : \sigma_1^2 \neq \sigma_2^2\}$ используется двухсторонний критерий

$$P(f > f_{\frac{\alpha}{2}}) = \frac{\alpha}{2},$$

$$P(f < f_{\frac{2-\alpha}{2}}) = \frac{\alpha}{2}.$$

Для проверки гипотезы $\{H_0 : \sigma_1^2 = \sigma_2^2;\ H_1 : \sigma_1^2 > \sigma_2^2\}$ используется правосторонний критерий $P(f > f_\alpha) = \alpha$, а гипотезы $\{H_0 : \sigma_1^2 = \sigma_2^2;\ H_1 : \sigma_1^2 < \sigma_2^2\}$ – левосторонний критерий $P(f < f_\alpha) = \alpha$.

5.5 Критерий Неймана-Пирсона

Будем рассматривать двухальтернативную гипотезу $\{H_0, H_1\}$. Пусть S – пространство выборок $X = (x_1, x_2, ..., x_n)$. Проверка сформулированной гипотезы сводится к разбиению пространства выборок S на две области G_0 и G_1. Если конкретная выборка $X = (x_1, x_2, ..., x_n)$ попадает в область G_0, то принимается гипотеза H_0, и принимается гипотеза H_1 в противном случае. При вынесении решения возможны следующие ошибки: *ошибка первого рода*, когда гипотеза H_0 верна, но отклоняется, и *ошибка второго рода*, когда гипотеза H_0 не верна, но принимается.

Задача проверки двухальтернативной гипотезы актуальна в радиолокации при обнаружении воздушных целей, когда гипотеза H_0 – цель есть, а гипотеза H_1 – цели нет. В этом случае ошибка первого рода называется ошибкой ложного отбоя, а ошибка второго рода – ошибкой ложной тревоги.

Пусть α и β – вероятности ошибок первого и второго рода соответственно, а $f(X / H_0)$, $f(X / H_1)$ – плотности вероятности выборки при условии истинности гипотез H_0 и H_1 соответственно. Тогда эти вероятности определяются формулами

$$\alpha = P(X \in G_1 / H_0) = \int_{G_1} f(X / H_0) dX,$$

$$\beta = P(X \in G_0 / H_1) = \int_{G_0} f(X / H_1)dX .$$

Величина $\gamma = 1 - \beta$ называется мощностью критерия, соответствующего разбиению G_0, G_1. Она представляет собой вероятность отклонить неверную гипотезу и определяется формулой

$$\gamma = 1 - \beta = P(X \in G_1 / H_1) = \int_{G_1} f(X / H_1)dX .$$

Иллюстрация вероятностей ошибок первого и второго рода и мощности критерия для одномерного пространства выборок $S = R^1$ приведена на рис. 5.5.

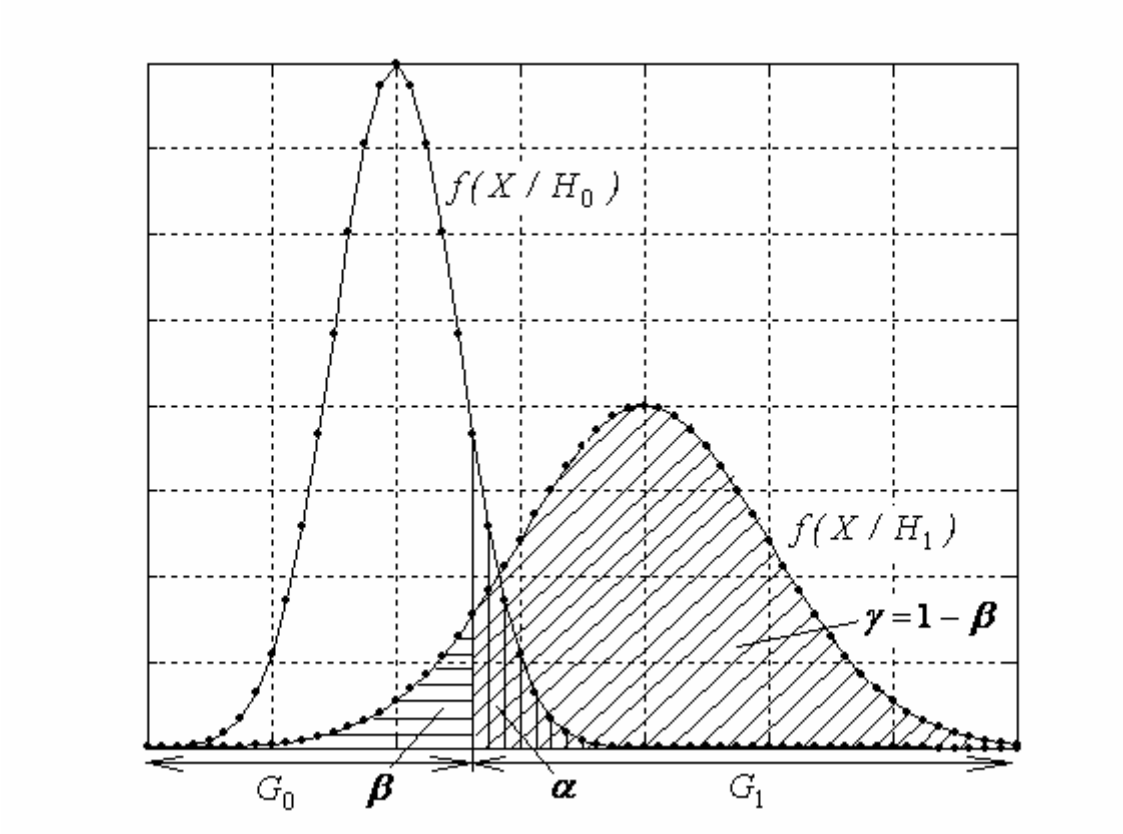

Рис. 5.5 Иллюстрация вероятностей ошибок первого и второго рода

При проверке двухальтернативной гипотезы критерием значимости мы задаем малую вероятность ошибки первого рода (уровень значимости α, см. раздел 5.2), и не контролируем вероятность ошибки второго рода. Вместе с тем нежелательными являются ошибки как первого, так и второго рода. Нейман и Пир-

сон предложили подход к проверке гипотезы, согласно которому задается некоторая малая вероятность ошибки первого рода и минимизируется вероятность ошибки второго рода (максимизируется мощность критерия). При таком подходе решается следующая оптимизационная задача:

$$\int_{G_1} f(X / H_0)dX = \alpha = const,$$

$$\gamma = \int_{G_1} f(X / H_1)dX \to max.$$

Критерий проверки гипотезы, получаемый в результате решения этой задачи, называется (и является) наиболее мощным по сравнению с другими критериями. Решение этой задачи для двухальтернативной простой гипотезы дается следующей леммой.

Лемма (Неймана-Пирсона) [15]. Пусть G – область пространства выборок S такая, что

$$\int_{G} f(X / H_0)dX = \alpha, \tag{5.9}$$

и существует постоянная k такая, что для области G_1, внутри которой

$$f(X / H_1) \geq kf(X / H_0) \tag{5.10}$$

и вне которой $f(X / H_1) \leq kf(X / H_0)$, выполняется условие (5.9), т.е.

$$\int_{G} f(X / H_0)dX = \int_{G_1} f(X / H_0)dX = \alpha. \tag{5.11}$$

Тогда

$$\int_{G_1} f(X / H_1)dX \geq \int_{G} f(X / H_1)dX. \tag{5.12}$$

Для доказательства обозначим общую часть областей G и G_1 как GG_1. Вычитая из обеих частей равенства (5.11) интеграл по общей области, получим

$$\int_{G-GG_1} f(X / H_0)dX = \int_{G_1-GG_1} f(X / H_0)dX. \tag{5.13}$$

Рассмотрим теперь разность

$$\int_{G_1} f(X/H_1)dX - \int_{G} f(X/H_1)dX .$$

Прибавляя и вычитая здесь интеграл по общей области, получим

$$\int_{G_1} f(X/H_1)dX - \int_{G} f(X/H_1)dX = \int_{G_1-GG_1} f(X/H_1)dX - \int_{G-GG_1} f(X/H_1)dX \geq$$

$$\geq \int_{G_1-GG_1} kf(X/H_0)dX - \int_{G-GG_1} kf(X/H_0)dX .$$

Неравенство здесь записано на основании (5.10). На основании равенства (5.11) последняя разность здесь равна нулю, и мы получаем (5.12). Лемма доказана.

Таким образом, в лемме Фишера доказано, что оптимальной областью G_1 является область

$$\{X : f(X/H_1) \geq kf(X/H_0)\}, \qquad (5.14)$$

поскольку, в соответствии с (5.12), мощность критерия для этой области больше мощности для любой другой области G, удовлетворяющей условию (5.11).

Критерий (5.14) мы можем записать в виде критерия значимости

$$P(l(X) \geq k/H_0) = \alpha, \qquad (5.15)$$

где статистика $l(X)$ определяется выражением

$$l(X) = \frac{f(X/H_1)}{f(X/H_0)}$$

и называется отношением правдоподобия. Критерий (5.15) называется критерием Неймана-Пирсона для проверки двухальтернативной простой гипотезы. Для проверки гипотезы необходимо получить распределение статистики $l(X)$ при условии истинности гипотезы H_0 и найти предел значимости k для этой статистики по таблице процентных отклонений распределения статистики $l(X)$ на уровне значимости α. Если эмпирическое значение $l_э(X)$ статистики удовлетворяет неравенству $l_э(X) \geq k$, то гипотеза H_0 отклоняется.

Вместо отношения правдоподобия можно использовать логарифмическое отношение правдоподобия, поскольку если $P(l(X) \geq k / H_0) = \alpha$, то и $P(\ln l(X) \geq k' / H_0) = \alpha$, где k' – некоторый новый порог.

Пример 5.1. Проверим гипотезу $\{H_0 : a = a_0;\ H_1 : a = a_1\}$ о математическом ожидании a нормальной генеральной совокупности $N(a, \sigma^2)$ при известной дисперсии σ^2 критерием Неймана-Пирсона, где a_0, a_1 – некоторые числа. Воспользуемся логарифмическим отношением правдоподобия. Поскольку

$$f_\xi(x, a, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

(см. пример 2.3 раздела 2.2), то

$$f(X / H_0) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x - a_0)^2\right),$$

$$\ln l(X) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a_0)^2 = \frac{a_1 - a_0}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n(a_1^2 - a_0^2)}{2\sigma^2} =$$

$$= \frac{n(a_1 - a_0)}{\sigma^2}\bar{x} - \frac{n(a_1^2 - a_0^2)}{2\sigma^2},$$

где $\bar{x} = (\sum_{i=1}^{n} x_i^2) / n$. Вместо критерия $P(\ln l(X) \geq k' / H_0) = \alpha$ мы можем использовать критерий

$$P(\bar{x} \geq k'' / H_0) = \alpha,$$

где

$$k'' = \frac{(a_0 + a_1)}{2} + \frac{\sigma^2 k'}{n(a_1 - a_0)}.$$

Поскольку статистика $\bar{x}$ при условии истинности гипотезы H_0 имеет нормальное распределение $N(a_0, \frac{\sigma^2}{n})$ (см. раздел (3.4)), то для определения предела значимости k'' необходимо пользоваться таблицами этого распределения. Так

как обычно составляются таблицы для распределения $N(0,1)$, то целесообразно перейти к статистике $u = \frac{\bar{x} - a_0}{\sigma}\sqrt{n}$, имеющей распределение $N(0,1)$, и использовать критерий

$$P(u \geq h) = \alpha,$$

где $h = a_0 + \frac{\sigma}{\sqrt{n}} u_\alpha$ и u_α – 100α-процентное отклонение распределения $N(0,1)$ (таблица 4.1 раздела 4.3).

5.6 Непараметрические критерии проверки гипотез

Гипотезы, при определении которых не указываются значения параметров, называются непараметрическими. Критерии для проверки таких гипотез также называются непараметрическими. В данном разделе рассмотрим некоторые из таких критериев.

5.6.1 Критерий квантилей

Этот критерий применяется для проверки гипотезы о том, что выборка $x_1,...,x_n$ извлечена из совокупности с функцией распределения $F_\xi(x)$, квантиль x_p порядка p которой имеет значение x_{p^*} [18]. Напомним, что квантилью порядка p распределения $F_\xi(x)$ называется величина x_p, определяемая формулой

$$P(\xi < x_p) = F_\xi(x_p) = p.$$

Опишем гипотезу более аккуратно. Пусть G_{p^*} – класс (множество) функций распределения, для которых x_p есть квантиль порядка p^*, $x_p = x_{p^*}$, а G –

класс всех возможных функций распределения. Будем проверять двухальтернативную гипотезу

$$\{H_0 : F(x) \in G_{p^*} \;\; H_1 : F(x) \in G \setminus G_{p^*}\}.$$

Критерий для проверки этой гипотезы строится следующим образом. Обозначим событие $A = \{\xi < x_p\}$, имеющее вероятность p. Пусть η – число появлений события A в n независимых испытаниях. Если выборка объема n извлечена из распределения, принадлежащего классу G_{p^*} (то есть если гипотеза H_0 верна), то случайная величина η имеет биномиальное распределение с числом испытаний n и вероятностью успеха в одном испытании p. Вероятность, что в промежутке $(-\infty, x_p)$ будет ровно r значений этой величины, определяется формулой Бернулли:

$$P\{\eta = r\} = C_n^r p^r (1-p)^{n-r}.$$

Для заданной малой вероятности α определяется двухсторонняя критическая область следующими условиями:

$$P(\eta \le r_{\alpha/2}) = \sum_{i=0}^{r_{\alpha/2}} C_n^i p^i (1-p)^{n-i} = \frac{\alpha}{2}, \tag{5.16}$$

$$P(\eta \ge r'_{\alpha/2}) = \sum_{i=r'_{\alpha/2}}^{n} C_n^i p^i (1-p)^{n-i} = \frac{\alpha}{2}. \tag{5.17}$$

Ввиду целочисленности случайной величины η равенства (5.16), (5.17) могут не выполняться. Поэтому нижнее критическое значение $r_{\alpha/2}$ определяется как наибольшее целое x, для которого

$$P(\eta \le x) = \sum_{i=0}^{x} C_n^i p^i (1-p)^{n-i} \le \frac{\alpha}{2}, \tag{5.18}$$

а верхнее критическое значение $r'_{\alpha/2}$ – как наименьшее целое y, для которого

$$P(\eta \ge y) = \sum_{i=y}^{n} C_n^i p^i (1-p)^{n-i} \le \frac{\alpha}{2}. \tag{5.19}$$

При выбранном α по таблицам биномиального распределения находятся числа $r_{\alpha/2}$ и $r'_{\alpha/2}$, удовлетворяющие условиям (5.18), (5.19). Подсчитывается число выборочных значений $r_э$, лежащих в промежутке $(-\infty, x_{p*})$. Гипотеза H_0 отклоняется, если $r_э \le r_{\alpha/2}$, или $r_э \ge r'_{\alpha/2}$.

Если критическую область определить условием

$$P(\eta \le r_\alpha) = \alpha,$$

то будем иметь левосторонний критерий. При определении критической области условием

$$P(\eta \ge r'_\alpha) = \alpha$$

будем иметь правосторонний критерий. Для левостороннего (правостороннего) критерия альтернатива проверяемой гипотезе будет иметь вид $x_p > x_{p*}$ ($x_p < x_{p*}$).

Таблицы сумм (5.18), (5.19) для определения критической области приведены в [22]. Расчет критических значений можно организовать на компьютере с использованием алгоритмов (5.18), (5.19).

5.6.2 Критерий знаков

Критерий знаков является частным случаем рассмотренного в предыдущем разделе критерия квантилей. Этот критерий применяется для проверки гипотезы о том, что выборка $x_1,...,x_n$ извлечена из совокупности с функцией распределения $F_\xi(x)$, квантиль порядка $0{,}5$ которой равна нулю, то есть $x_{0,5} = 0$.

Если обозначить G_{p*} – класс (множество) функций распределения, для которых x_p есть квантиль порядка $0{,}5$, $x_p = 0{,}5$, и $x_{0,5} = 0$, а G – класс всех возможных функций распределения, то проверяется двухальтернативная гипотеза

$$\{H_0 : F(x) \in G_{p*};\ H_1 : F(x) \in G \setminus G_{p*}\}.$$

Обычно важен случай, когда выборочные значения $x_1,...,x_n$ являются разностями между независимыми парами наблюдений некоторой случайной величины, $x_i = u_i - v_i$, $i = \overline{1,n}$, причем одно наблюдение u_i осуществляется при одном условии A, а другое v_i – при другом условии B. При этом проверяется гипотеза, что медиана равна нулю, $x_{0,5} = m = 0$, то есть что условия A и B дают один и тот же эффект.

Критерий знаков состоит в следующем. Подсчитывают число значений x_i, меньших нуля, и больших нуля (разности $u_i - v_i$ имеют знак "+" или "-"). Учитывают число случаев реже встречающегося знака. Пусть это число $\eta = r_э$, а $n - r_э$ – число случаев чаще встречающегося знака. Если проверяемая гипотеза H_0 верна, случайная величина η распределена по биномиальному закону $Bi(n,\ 0.5)$, так что по формуле Бернулли

$$P\{\eta = r\} = C_n^r\, 0{,}5^n.$$

Двухсторонняя критическая область задается условиями

$$P(\eta \le r_{\alpha/2}) = 0{,}5^n \sum_{i=0}^{r_{\alpha/2}} C_n^i = \frac{\alpha}{2}, \tag{5.20}$$

$$P(\eta \ge r'_{\alpha/2}) = 0{,}5^n \sum_{i=r'_{\alpha/2}}^{n} C_n^i = \frac{\alpha}{2}, \tag{5.21}$$

причем

$$r_{\alpha/2} + r'_{1-\alpha/2} = n.$$

Поскольку случайная величина η принимает целочисленные значения и равенства в (5.20), (5.21) могут не выполняться, то нижнее критическое значение $r_{\alpha/2}$ определяется как целочисленное решение относительно x неравенств

$$0{,}5^n \sum_{i=0}^{x} C_n^i \le \frac{\alpha}{2},\ 0{,}5^n \sum_{i=0}^{x+1} C_n^i > \frac{\alpha}{2}. \tag{5.22}$$

Соответственно, верхнее критическое значение $r'_{\alpha/2}$ определяется как целочисленное решение относительно y неравенств

$$0{,}5^n \sum_{i=y+1}^{n} C_n^i \geq \frac{\alpha}{2},\ 0{,}5^n \sum_{i=y}^{n} C_n^i < \frac{\alpha}{2}. \qquad (5.23)$$

Расчет критических значений $r_{\alpha/2}$, $r'_{\alpha/2}$ можно организовать на компьютере по алгоритмам (5.22), (5.23). В работах [2, 16] имеются таблицы для определения критических значений. Однако описание этих таблиц требует дополнительных выкладок и здесь не приводится. Гипотеза H_0 отклоняется, если $r_э < r_{\alpha/2}$, или $r_э < r'_{\alpha/2}$. Если используется левосторонний критерий

$$P(\eta \leq r_\alpha) = 0{,}5^n \sum_{i=0}^{r_\alpha} C_n^i = \alpha,$$

то альтернатива состоит в том, что медиана меньше нуля, и гипотеза H_0 отклоняется при $r_э < r_\alpha$. Если используется правосторонний критерий

$$P(\eta \geq r'_\alpha) = 0{,}5^n \sum_{r'_\alpha}^{n} C_n^i = \alpha,$$

то альтернатива состоит в том, что медиана больше нуля, и гипотеза H_0 отклоняется при $r_э > r'_\alpha$.

При больших n статистика η распределена асимптотически нормально $N(n/2, n/4)$, и для расчета критических значений можно пользоваться таблицами нормального распределения.

5.6.3 Критерий Уилкоксона

Критерий Уилкоксона – это критерий однородности двух выборок $x_1, ..., x_n$ и $y_1, ..., y_m$. Элементы выборок предполагаются взаимно независимыми с не-

прерывными функциями распределения $F_1(x)$ и $F_2(x)$ соответственно. Проверяется гипотеза

$$H_0 : F_1(x) = F_2(x).$$

При проверке гипотезы критерием Уилкоксона полученные $n+m$ наблюдений записываются в порядке возрастания значений, независимо от их принадлежности к той или иной выборке. В результате получается некоторый ряд, содержащий n величин x и m величин y, перемешанных между собой. Критерий Уилкоксона основан на ранговой статистике

$$W = \sum_{j=1}^{m} s(r_j), \tag{5.24}$$

где r_j – ранги (номера) чисел y_j в общем вариационном ряду x_i и y_j, а функция $s(r)$, $r = 1,2,...,n+m$, определяется заранее фиксированной подстановкой

$$\begin{pmatrix} 1, & 2, & ..., n+m \\ s(1), s(2), ..., s(n+m) \end{pmatrix}, \tag{5.25}$$

где $s(1)$, $s(2)$, …, $s(n+m)$ – одна из возможных перестановок чисел $1,2,...,n+m$. Выбор подстановки (5.25) осуществляется так, чтобы мощность критерия для заданной альтернативы H_1 была наибольшей. Распределение статистики W (5.24) зависит лишь от объема выборок и не зависит от выбора подстановки (если справедлива гипотеза H_0). Математическое ожидание и дисперсия статистики W определяются выражениями

$$E(W) = \frac{m(m+n+1)}{2}, \; D(W) = \frac{mn(m+n+1)}{12}. \tag{5.26}$$

Если проверяемая гипотеза H_0 верна, то можно получить закон распределения статистики W. Нижнее критическое значение $w_{\alpha/2}$ критерия Уилкоксона определяется как целочисленное решение относительно x неравенств

$$P\{W < x\} \le \alpha/2, \; p\{W < x+1\} > \alpha/2.$$

Так как распределение статистики W симметрично относительно математического ожидания, то верхнее критическое значение $w'_{\alpha/2}$ связано с нижним $w_{\alpha/2}$ соотношением

$$w'_{\alpha/2} = 2E(W) - w_{\alpha/2} = m(m+n+1) - w_{\alpha/2}. \qquad (5.27)$$

Пара чисел $w_{\alpha/2}$ и $w'_{\alpha/2}$ определяет критические значения двухстороннего критерия Уилкоксона с уровнем значения α.

Таблица для определения нижних критических значений $w_{\alpha/2}$ приведена в [2] для $m,n=\overline{1,25}$. Верхние критические значения находятся по формуле (5.27).

Если $m \to \infty$ и $n \to \infty$, то случайная величина W распределена асимптотически нормально с параметрами (5.26).

Вариант критерия рассмотренного типа предложен впервые Ф. Уилкоксоном для выборок равного объема и был основан на статистике (5.24) с функцией $s(r)=r$. Критерий с такой функцией $s(r)=r$ целесообразно использовать при альтернативах H_1 вида $F_1(x) < F_2(x)$ или $F_1(x) > F_2(x)$ при всех действительных x.

5.6.4 Критерий Манна-Уитни

Критерий Манна-Уитни – статистический критерий для проверки гипотезы H_0 об однородности двух выборок $x_1,...,x_n$ и $y_1,...,y_m$, все $n+m$ элементов которых взаимно независимы и подчиняются непрерывным распределениям $F_1(x)$ и $F_2(x)$ соответственно. Проверяемая гипотеза записывается в виде

$$H_0 : F_1(x) = F_2(x).$$

Критерий основан на статистике

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{i,j},$$

где

$$\delta_{i,j} = \begin{cases} 1, \ x_i < y_j, \\ 0, \ \textit{иначе.} \end{cases}$$

Эта статистика представляет собой общее число случаев, когда элементы выборки $x_1,...,x_n$ предшествуют элементам выборки $y_1,...,y_m$ в общем вариационном ряду. Доказано, что если H_0 верна, то

$$E(U) = \frac{mn}{2}, \ D(U) = \frac{mn(m+n+1)}{12}. \qquad (5.28)$$

Критерий Манна-Уитни является односторонним. Нижнее критическое значение u_α определяется как наибольшее целое x, для которого

$$P\{U \le x\} \le \alpha .$$

Манном и Уитни были рассчитаны вероятности значений статистики U для $1 \le m \le n \le 8$ (при условии истинности гипотезы H_0), позволяющие определить нижнее критическое значение u_α. Эти таблицы приведены в книге Химмельблау [22].

При $m \to \infty$, $n \to \infty$ статистика U асимптотически нормальна с параметрами (5.28). Таблицами нормального распределения можно пользоваться при $min\{m,n\} > 25$.

5.6.5 Критерий Ван-дер-Вардена

Критерий Ван-дер-Вардена [5] применяется для проверки однородности двух выборок против их различия в положениях. Точнее, альтернативой гипотезе об однородности

$$H_0 : F_1(x) = F_2(x)$$

двух непрерывных генеральных совокупностей $F_1(x)$, $F_2(x)$, представленных независимыми выборками $x_1, x_2, ..., x_n$ и $y_1, y_2, ..., y_m$ служит различие их положений при постоянстве формы (альтернатива сдвига).

Статистика X критерия Вар-дер-Вардена представляет собой сумму

$$X = \sum_{i=1}^{m} \Psi\left(\frac{s(r_i)}{N+1}\right), \tag{5.29}$$

где r_i – ранг (номер) значения y_i в общем упорядоченном ряду (предполагается, что $m \le n$), $N = m + n$ – общий объем выборки, функция $s(r)$ определяется заранее фиксированной подстановкой (5.25), и $\Psi(p)$ – функция квантилей распределения $N(0,1)$. Функция $\Psi(p)$ – это функция, обратная функции Лапласа $\Phi(x)$,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz .$$

Если $p = \Phi(x)$, p – вероятность, x – значение случайной величины, то $x = \Psi(p)$, или, иначе, $\Phi(\Psi(p)) = p$. Функция $\Psi(p)$ удовлетворяет условию $\Psi(p) = -\Psi(1-p)$.

Если проверяемая гипотеза H_0 верна, то распределение статистики X (5.29) зависит только от m и n.

Нижнее и верхнее критические значения для односторонних критериев определяются уравнениями

$$P(X < x_\alpha) = \alpha ,$$

$$P(X > x'_\alpha) = \alpha ,$$

причем $x_\alpha = -x'_\alpha$. Критическая область для двухстороннего критерия Ван-дер-Вардена задается условием

$$P(|X| > x'_{\alpha/2}) = \alpha .$$

Таблица верхних критических значений $x'_{\alpha/2}$ имеется в [2] для $m+n=\overline{6,50}$ и $n-m=\overline{0,5}$.

Числовые характеристики статистики X определяются выражениями

$$E(X)=0,\ D(X)=\frac{mnS}{m+n+1}, \tag{5.30}$$

где

$$S=\frac{1}{N}\sum_{i=1}^{N}\Psi^2\left(\frac{i}{N+1}\right).$$

Если $N=n+m\to\infty$, то случайная величина X распределена асимптотически нормально с параметрами (5.30) вне зависимости от того, стремится ли в отдельности m и n к бесконечности. Потому в данном случае верхнее критическое значение $x'_{\alpha/2}$ определяется формулой

$$x'_{\alpha/2}=\Psi\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{mnS}{m+n-1}}.$$

Критерий Ван-дер Вардена рекомендуется применять, если предполагается, что наблюдения близко следуют нормальному закону.

5.6.6 Критерий Смирнова

Критерий Смирнова – статистический критерий для проверки гипотезы о том, что две независимые выборки подчиняются общему (непрерывному) распределению. Иначе говоря, проверке подлежит гипотеза однородности

$$H_0: F_1(x)=F_2(x)$$

по двум независимым выборкам $x_1,...,x_m$ и $y_1,...,y_n$ из двух непрерывных распределений $F_1(x)$ и $F_2(x)$ соответственно.

Критерий Смирнова основан на статистике

$$D_{m,n}=\sup_x\left|F_1^*(x)-F_2^*(x)\right|,$$

где $F_1^*(x)$ и $F_2^*(x)$ – выборочные (эмпирические) функции распределения для первой и второй выборок соответственно.

Если гипотеза H_0 верна, то распределение статистики $D_{m,n}$ не зависит от теоретического распределения.

Смирнов Н.В. доказал, что если гипотеза H_0 верна и объемы выборок неограниченно увеличиваются $(m \to \infty,\ n \to \infty,\ m/n \to \rho > 0)$ то статистика

$$z = \sqrt{\frac{mn}{m+n}} D_{m,n}$$

имеет приближенно функцию распределения вида

$$K(y) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 y^2},$$

то есть

$$K(y) = \lim_{\substack{m \to \infty \\ m \le n}} P\left\{ \sqrt{\frac{mn}{m+n}} D_{m,n} < y \right\}. \qquad (5.31)$$

Функция $K(y)$ называется функцией распределения Колмогорова. Зная функцию распределения (5.31), можно определить критическую область условием

$$P\{z > z'_\alpha\} = \alpha.$$

Если эмпирическое значение статистики $z_э$ попадает в критическую область, то гипотеза H_0 отклоняется.

Для критерия Смирнова построены таблицы критических значений для $n, m = \overline{1,20}$ [2]. По этим таблицам при заданном уровне значимости α и объемах выборок n, m находится не критическое значение z'_α, а целое число r. Критическое значение z'_α затем можно получить по формуле

$$z'_\alpha = \frac{r}{k},$$

где k – наименьшее общее кратное чисел n, m. Число k также указывается в таблице.

ОГЛАВЛЕНИЕ