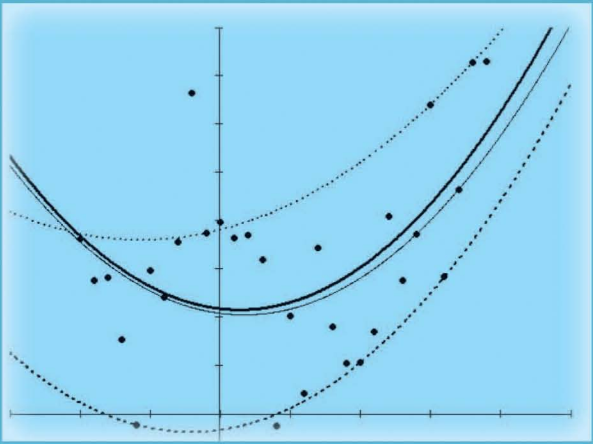
Е.Р. Горяинова
А.Р. Панков
Е.Н. Платонов

ПРИКЛАДНЫЕ МЕТОДЫ анализа статистических данных

Учебное пособие



ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Е.Р. Горяинова
А.Р. Панков
Е.Н. Платонов

ПРИКЛАДНЫЕ МЕТОДЫ анализа статистических данных

Учебное пособие

Рекомендовано УМО в области экономики и менеджмента в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению подготовки «Экономика»

Издательский дом
Высшей школы экономики
Москва, 2012

УДК 519.2(075)
ББК 22.172я7
Г71

Рецензенты:

доктор технических наук, профессор *Ф.Т. Алескеров*;
доктор физико-математических наук *А.В. Борисов*

Г 71 **Горяинова, Е. Р., Панков, А. Р., Платонов, Е. Н.** Прикладные методы анализа статистических данных [Текст] : учеб. пособие / Е. Р. Горяинова, А. Р. Панков, Е. Н. Платонов ; Нац. исслед. ун-т «Высшая школа экономики». — М.: Изд. дом Высшей школы экономики, 2012. — 310, [2] с. — 1000 экз. — 978-5-7598-0866-4 (в обл.).

В учебном пособии излагаются важнейшие понятия математической статистики, описываются статистические модели и методы статистического анализа реальных данных. Все рассмотренные методы проиллюстрированы примерами, которые снабжены подробными решениями и комментариями. В конце каждого раздела приводятся задачи для самостоятельного решения. Наряду с важнейшими базовыми классическими моделями и методами статистической обработки данных в пособии представлены современные непараметрические робастные методы, которые можно эффективно использовать для обработки информации в условиях априорной статистической неопределенности, свойственной реальным статистическим экспериментам.

Для студентов, аспирантов и преподавателей технических и экономических вузов.

УДК 519.2(075)
ББК 22.172я7

ISBN 978-5-7598-0866-4

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Учебное пособие содержит систематическое изложение важнейших понятий математической статистики и методов статистического анализа эмпирических данных. В данном пособии рассмотрены некоторые современные методы анализа данных, например, подробно изучаются непараметрические робастные методы, которые можно использовать для обработки информации в условиях априорной статистической неопределенности, свойственной реальным статистическим экспериментам. Каждый раздел пособия содержит как базовые теоретические положения, так и разнообразные примеры. В конце каждого раздела приведены задачи для самостоятельного решения с ответами и указаниями. Пособие предназначено для студентов факультета «Бизнес-информатика», а также для студентов и аспирантов других факультетов, занимающихся статистической обработкой эмпирических данных.

Структура изложения такова, что это пособие может одновременно играть роль учебника, задачника и справочника.

Данная книга посвящена систематическому изложению основ важного раздела современной прикладной математики — математической статистики. При ее подготовке авторы основывались на следующих базовых принципах:

— математически корректное изложение материала и обоснование всех методов, используемых для решения конкретных задач;

— иллюстрирование основных методов конкретными примерами различного уровня сложности;

— более подробное рассмотрение тех вопросов, которые в настоящее время являются наиболее важными для решения прикладных задач.

Книга состоит из пяти глав. В главе I приведены основные определения и теоретические положения общего характера, необходимые для изучения остального материала, а также кратко описаны важнейшие типы оценок, их свойства и методы их построения.

В главе II описывается постановка и формулируются подходы к решению проблемы первичного анализа статистических данных. Рассматриваются параметрические и непараметрические методы проверки гипотез об однородности выборочных данных. Изучается проблема обнаружения зависимости статистических данных, измеряемых в различных шкалах. Проводится сравнительный анализ асимптотической эффективности классических и ранговых методов.

В главе III книги рассматриваются методы восстановления и прогнозирования зависимостей с использованием как линейных, так и нелинейных регрессионных моделей. Изучаются проблема вырожденности регрессионной модели и методы борьбы с мультиколлинеарностью. Исследуется влияние аномальных ошибок в наблюдениях на точность оценивания параметров регрессии и рассматриваются методы их робастного оценивания. Раскрываются методы проверки адекватности регрессионных моделей по эмпирическим данным и методы построения непараметрических моделей.

Глава IV посвящена анализу временных рядов. Дается подробное описание общей структуры временного ряда. Рассматривается задача выделения детерминированной компоненты временного ряда и идентификация случайной компоненты, которая может быть описана моделью авторегрессии, скользящего среднего или их комбинацией.

Глава V имеет справочный характер и содержит дополнительные сведения по функциональному анализу и теории вероятностей, необходимые для изучения материала в полном объеме. Также в нее включены самые необходимые таблицы, используемые для статистических расчетов.

Авторы выражают благодарность за проявленное внимание профессору Национального исследовательского университета «Высшая школя экономики» Ф.Т. Алескерову — инициатору создания курса «Анализ данных» на факультете «бизнес-информатика», а также коллегам по кафедре «Теория вероятностей» Московского авиационного института К.В. Семенихину и К.В. Степаняну.

Во время работы над этой книгой мы понесли тяжелую утрату. Скоропостижно скончался наш соавтор, старший товарищ и Учитель Алексей Ростиславович Панков. Надеемся, что нам удалось достойно завершить последнюю совместную с А.Р. Панковым работу, и эта книга будет памятью о талантливом и светлом человеке — Алексее Ростиславовиче Панкове.

Е.Р. Горяинова
Е.Н. Платонов

СПИСОК ОСНОВНЫХ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

АОЭ — асимптотическая относительная эффективность;

АР(p) — авторегрессия порядка p;

АРСС(p, q) — модель авторегрессии и скользящего среднего порядков (p, q);

ВР — временной ряд;

МНК — метод наименьших квадратов;

МП-оценка — оценка метода максимального правдоподобия;

НЛН-оценка — наилучшая линейная несмещенная оценка;

ОМНК — обобщенный МНК;

СВ — случайная величина или случайный вектор;

с.к.о. — среднее квадратическое отклонение;

СП — случайная последовательность;

ССП — стационарная СП;

СС(q) — скользящее среднее порядка q;

ЦПТ — центральная предельная теорема;

ЧАКФ — частотная автокорреляционная функция;

$\mathbb{N}$ — множество натуральных чисел;

$\mathbb{R}^n$ — n-мерное (вещественное) евклидово пространство;

$A^\top$ — транспонированная матрица;

A^{-1} — обратная матрица;

I — единичная матрица;

$\mathrm{tr}[A]$ — след матрицы A;

$\det[A]$ — определитель матрицы A;

$A \geqslant 0$ — неотрицательно определенная матрица;

$a \approx b$ — число a приближенно равно числу b;

$\exp\{x\} = e^x$ — экспонента;

$\max(x_1, \ldots, x_n)$ — максимум из $x_1, \ldots, x_n$;

$\arg\min\limits_{x \in X} f(x)$ — точка минимума функции $f(x)$ на множестве X;

$n \gg m$ $(n \ll m)$ — число n намного больше (меньше), чем m;

Ω — пространство элементарных событий (исходов) ω;

$\mathcal{F}$ — σ-алгебра случайных событий (подмножеств Ω);

$\mathbf{P}\{A\}$ — вероятность (вероятностная мера) события A;

$\{\Omega, \mathcal{F}, \mathbf{P}\}$ — основное вероятностное пространство;

$\varnothing$ — невозможное событие;

$F_X(x)$ — функция распределения СВ X;

$X \sim F(x)$ — СВ X имеет распределение $F(x)$;

$p_X(x)$ — плотность вероятности СВ X;

$m_X = \mathbf{M}\{X\}$ — математическое ожидание (среднее) СВ X;

$D_X = \mathbf{D}\{X\}$ — дисперсия СВ X;

$\mathbf{cov}\{X, Y\}$ — ковариация СВ X и Y;

$\overset{\circ}{X}$ — центрированная СВ X;

$\Phi(x)$ — интеграл вероятностей (функция Лапласа);

$X_n \xrightarrow{\mathbf{P}} X$ — сходимость по вероятности;

$X_n \xrightarrow{\text{с.к.}} X$ — сходимость в среднем квадратическом (с.к.-сходимость);

$X_n \xrightarrow{\text{п.н.}} X$ — сходимость почти наверное;

$X_n \xrightarrow{d} X$ — сходимость по распределению (слабая сходимость);

$\Pi(\lambda)$ — распределение Пуассона с параметром λ;

$Bi(N;p)$ — биномиальное распределение с параметрами N, p;

$R[a;b]$ — равномерное распределение на отрезке $[a,b]$;

$E(\lambda)$ — экспоненциальное распределение с параметром λ;

$\mathcal{L}(\lambda)$ — распределение Лапласа с параметром λ;

$Lg(m;\sigma^2)$ — логистическое распределение с параметрами m, σ^2;

$\mathcal{N}(m;D)$ — гауссовское (нормальное) распределение со средним m и дисперсией (ковариационной матрицей) D;

$\Psi_X(\lambda)$ — характеристическая функция n-мерного гауссовского распределения:

$\mathcal{T}_r$ — распределение Стьюдента с r степенями свободы;

$\mathcal{H}_n$ — распределение хи-квадрат с n степенями свободы;

$\chi^2_{n,\delta}$, $\mathcal{H}_{n,\delta}$ — нецентральное распределение хи-квадрат с n степенями свободы и параметром нецентральности δ;

$F(m;n)$ — распределение Фишера с двумя степенями свободы m и n;

$F(m;n;\delta)$ — нецентральное распределение Фишера с двумя степенями свободы m и n и параметром нецентральности δ;

u_α — квантиль уровня α распределения $\mathcal{N}(0;1)$;

$k_\alpha(n)$ — квантиль уровня α распределения $\mathcal{H}_n$;

$t_\alpha(r)$ — квантиль уровня α распределения $\mathcal{T}_r$;

$f_\alpha(m;n)$ — квантиль уровня α распределения Фишера $F(m;n)$.

ГЛАВА I

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ

Для того чтобы познакомить читателей с прикладными методами анализа статистических данных, необходимо определить основные понятия и положения математической статистики, которыми мы будем пользоваться. Предполагается, что читателями уже освоен курс теории вероятностей, тем не менее базовые понятия и сведения по теории вероятностей приведены в математическом приложении в главе 5.

Методы теории вероятностей позволяют по заданному закону распределения случайной величины (СВ) вычислять ее числовые характеристики, вероятности тех или иных событий, связанных с этой величиной. Однако на практике, за исключением самых простых случаев, точное вероятностное распределение СВ неизвестно. Поэтому естественно возникает вопрос: как найти эти исходные вероятности, функцию распределения, числовые характеристики? Для получения исходных данных, необходимых для построения вероятностной модели, приходится обращаться к эксперименту. Задача восстановления или уточнения закона распределения СВ по результатам проводимых наблюдений является основной задачей математической статистики. Первая глава этой книги будет посвящена описанию статистических моделей, формализации статистических задач и алгоритмам первичной статистической обработки экспериментальных данных.

Например, пусть имеются данные (см. пример 1.1) о росте достаточно большого количества людей. Попытаемся по результатам наблюдений СВ X, где X — рост человека, построить вероятностную модель этой величины, а именно оценить неизвестное математическое ожидание и дисперсию этой величины, восстановить неизвестную функцию распределения и плотность вероятности СВ X, построить интервал, которому с заданной вероятностью принадлежит неизвестное значение среднего роста.

Отметим, что первая глава, в основном, носит теоретический характер, в ней даны определения точечных и интервальных оценок параметров распределений; изучены свойства, характеризующие качество построенных статистических оценок; представлены основные методы нахождения точечных оценок и указаны способы построения доверительных интервалов для параметров основных вероятностных распределений; описан алгоритм проверки статистических параметрических гипотез.

1. Выборка и ее основные характеристики

Как правило, исходным материалом для построения статистической модели являются результаты эксперимента, в котором проводится n независимых наблюдений за некоторой СВ X.

1.1. Теоретические положения

Пусть X — произвольная случайная величина с функцией распределения $F(x) = \mathbf{P}(X \leqslant x)$, $x \in \mathbb{R}^1$.

Определение 1.1. Совокупность $\{X_k,\ k = 1, \ldots, n\}$ независимых случайных величин, имеющих одинаковые функции распределения $F_{X_k}(x) = F(x)$, называется *однородной выборкой объема* n, соответствующей функции распределения $F(x)$.

СВ X_k $(k = 1, \ldots, n)$ называется k-м *элементом выборки.*

Из определения 1.1 следует, что выборку можно рассматривать как случайный вектор $Z_n = \{X_1, \ldots, X_n\}^\top$ с независимыми компонентами. Кроме того, СВ $\{X_k,\ k = 1, \ldots, n\}$ — независимые вероятностные «копии» СВ X, поэтому мы также будем говорить, что *выборка* Z_n *порождена СВ* X с распределением $F(x)$.

Определение 1.2. Выборка $\{X_k,\ k = 1, \ldots, n\}$ называется *гауссовской*, если Z_n — n-мерный гауссовский вектор.

Определение 1.3. Выборка Z_n называется *неоднородной*, если законы распределения $F_{X_k}(x)$ ее элементов неодинаковы.

Далее полагается, что выборка Z_n — однородная, если специально не указано обратное.

Из приведенных определений следует, что выборка является математической моделью последовательности одинаковых опытов со случайными исходами, проводимых в неизменных условиях, причем результаты опытов статистически независимы.

Определение 1.4. *Реализацией выборки* Z_n называется неслучайный вектор $z_n = \{x_1, \ldots, x_n\}^\top$, компонентами которого являются реализации соответствующих элементов выборки.

Определение 1.5. СВ $Y = \varphi\,(X_1, \ldots, X_n)$, где $\varphi(x_1, \ldots, x_n)$ — произвольная (борелевская) функция на $\mathbb{R}^n$, называется *статистикой.*

Пусть $z_n = \{x_1, \ldots, x_n\}^\top$ — некоторая реализация выборки Z_n, а $z_{(n)} = \{x_{(1)}, \ldots, x_{(n)}\}^\top$ — вектор, компонентами которого являются упорядоченные по возрастанию числа $(x_1, \ldots, x_n)$, т.е. $x_{(1)} \leqslant x_{(2)} \leqslant \leqslant \ldots \leqslant x_{(n)}$.

Определение 1.6. СВ $X_{(k)}$, реализацией которой для каждой z_n является число $x_{(k)}$, называется *k-й порядковой статистикой*, $k = 1, \ldots, n$. Случайный вектор $Z_{(n)} = \{X_{(1)}, \ldots, X_{(n)}\}^\top$ называется *вариационным рядом* выборки.

СВ $X_{(1)}$ и $X_{(n)}$ (т.е. крайние элементы вариационного ряда) называются *экстремальными порядковыми статистиками.*

Порядковые статистики используются при анализе свойств распределения СВ X, в частности при оценивании квантилей распределения СВ.

Рассмотрим некоторые важнейшие для приложений виды статистик.

Определение 1.7.

1) $\overline{X}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n} X_k$ называется *выборочным средним.*

2) $\overline{S}^2_n = \dfrac{1}{n}\sum\limits_{k=1}^{n}(X_k - \overline{X}_n)^2$ называется *выборочной дисперсией.*

3) $\overline{\nu}_r(n) = \dfrac{1}{n}\sum\limits_{k=1}^{n}(X_k)^r$, $r = 1, 2, \ldots$, называется *выборочным начальным моментом* r-го порядка.

4) $\overline{\mu}_r(n) = \dfrac{1}{n}\sum\limits_{k=1}^{n}(X_k - \overline{X}_n)^r$, $r = 1, 2, \ldots$, называется *выборочным центральным моментом* r-го порядка.

Заметим, что $\overline{X}_n = \overline{\nu}_1(n)$, а $\overline{S}^2_n = \overline{\mu}_2(n)$.

Для того чтобы описать свойства выборочных моментов, необходимо знать виды сходимости последовательности СВ. Соответствующие определения представлены в математическом приложении (см. разд. 21.6).

Пусть распределение $F(x)$ таково, что следующие теоретические моменты любого элемента X_k выборки: $m_X = \mathbf{M}\{X_k\}$, $D_X = \mathbf{D}\{X_k\}$, $\nu_r = \mathbf{M}\{(X_k)^r\}$, $\mu_r = \mathbf{M}\{(X_k - m_X)^r\}$, $r = 2, 3, \ldots$ существуют и конечны. Тогда справедливо следующее утверждение.

Теорема 1.1. *При неограниченном увеличении объема выборки n выборочные моменты $\overline{\nu}_r(n)$ и $\overline{\mu}_r(n)$, $r = 1, 2, \ldots$ почти наверное сходятся к теоретическим моментам ν_r и μ_r соответственно.*

Следствие 1.1. *Если m_X существует и конечен, то $\overline{X}_n \xrightarrow{\text{п.н.}} m_X$ при $n \to \infty$. Если ν_2 существует и конечен, то $\overline{S}^2_n \xrightarrow{\text{п.н.}} D_X$, $n \to \infty$.*

При определенных дополнительных условиях выборочные моменты обладают свойством асимптотической нормальности.

Теорема 1.2. *Пусть для некоторого* $r \geqslant 1$ *существует и конечен момент* ν_{2r}. *Тогда* $\sqrt{n}\,(\overline{\nu}_r(n) - \nu_r) \xrightarrow{d} \xi \sim \mathcal{N}(0; \nu_{2r} - \nu_r^2)$, $n \to \infty$.

Следствие 1.2. *Если* $D_X < \infty$, *то*

$$\sqrt{n}\,(\overline{X}_n - m_X) \xrightarrow{d} \xi \sim \mathcal{N}(0; D_X), \quad n \to \infty.$$

Если $\nu_4 < \infty$, *то* $\sqrt{n}\,(\overline{\nu}_2(n) - \nu_2) \xrightarrow{d} \xi \sim \mathcal{N}(0; \nu_4 - D_X^2)$, $n \to \infty$.

Из приведенных выше утверждений следует, что при $n \gg 1$ выборочные моменты $\overline{\nu}_r(n)$ и $\overline{\mu}_r(n)$ практически не отличаются от своих теоретических значений ν_r и μ_r. Кроме того, можно считать, что $\overline{\nu}_r(n) \sim \mathcal{N}\left(\nu_r; \dfrac{\nu_{2r} - \nu_r^2}{n}\right)$, если $n \gg 1$.

Пусть выборка $\{X_k,\ k = 1, \ldots, n\}$ порождена СВ X с функцией распределения $F(x)$. Для любого $x \in \mathbb{R}^1$ введем событие $A_X = \{X \leqslant \leqslant x\}$, тогда $\mathbf{P}(A_X) = F(x)$. Обозначим через $M_n(x)$ случайное число элементов выборки, не превосходящих x.

Определение 1.8. Случайная функция $\widehat{F}_n(x) = \dfrac{M_n(x)}{n}$, $x \in \mathbb{R}^1$, называется *выборочной (эмпирической) функцией распределения* СВ X.

При достаточно больших n функция $\widehat{F}_n(x)$ весьма точно аппроксимирует функцию распределения $F(x)$, которой соответствует выборка, о чем свидетельствуют следующие утверждения.

Теорема 1.3 (Гливенко—Кантелли). $\widehat{F}_n(x)$ *сходится к* $F(x)$ *почти наверное равномерно по* x *при* $n \to \infty$, *т.е.*

$$\sup_{x \in \mathbb{R}^1} |\widehat{F}_n(x) - F(x)| \xrightarrow{\text{п.н.}} 0, \quad n \to \infty.$$

Теорема 1.4. *При любом* $x \in \mathbb{R}^1$ *последовательность* $\{\widehat{F}_n(x),\ n = 1, 2, \ldots\}$ *асимптотически нормальна:*

$$\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right) \xrightarrow{d} \xi \sim \mathcal{N}\left(0; F(x)(1 - F(x))\right), \quad n \to \infty.$$

Пусть выборка $\{X_k,\ k = 1, \ldots, n\}$ порождена абсолютно нерерывной СВ X с плотностью вероятности $p(x)$. Если функция $p(x)$ неизвестна, то для ее оценивания можно построить *гистограмму*. Построение гистограммы проводится после предварительной *группировки данных*. Для этого область V_X всех возможных значений СВ X разбивается на $K > 1$ непересекающихся интервалов $\{\delta_m : m = 1, \ldots, K\}$: $\bigcup\limits_{m=1}^{K} \delta_m = V_X$, $\delta_m \cap \delta_i = \varnothing$, $m \neq i$. При выборе числа K интервалов группировки можно воспользоваться *формулой Стерджеса*:

$K = 1 + \{3{,}32 \lg n\}$, где $\{a\}$ — целая часть числа a. Если множество V_X неизвестно, то его можно взять равным $\left[X_{(1)}, X_{(n)}\right]$.

Обозначим через $n_m(x)$ случайное число элементов выборки, попавших в интервал δ_m, которому принадлежит x, а через h_m длину интервала δ_m, $m = 1, \ldots, K$. Очевидно, что $n_m(x) = \sum\limits_{k=1}^{n} I_m(x, X_k)$, где

$$I_m(x, X_k) = \begin{cases} 1, & \text{если } x, X_k \in \delta_m, \\ 0, & \text{в противоположном случае.} \end{cases}$$

Определение 1.9. Случайная функция $\widehat{p}_n(x) = \dfrac{n_m(x)}{nh_m}$, $x \in \mathbb{R}^1$, называется *гистограммой* СВ X.

Гистограмма является кусочно-постоянной функцией, причем площадь прямоугольника под функцией для каждого интервала δ_m равна $\dfrac{n_m(x)}{n}$, т. е. совпадает с частотой попадания элементов выборки в интервал. Эта частота будет сходиться к вероятности попадания СВ X с плотностью вероятности $p(x)$ в соответствующий интервал.

Такой способ оценивания неизвестной плотности вероятности можно рекомендовать только на предварительном этапе анализа статистических данных, поскольку он обладает очевидными недостатками: неопределенностью в способе выбора интервалов, потерей информации при группировки данных, разрывностью гистограммы.

Существуют более современные методы оценивания неизвестной плотности вероятности, основанные на использовании *ядерных оценок*. Более подробно с ними можно познакомится в [37].

Пусть двумерная выборка $\{(X_k, Y_k), k = 1, \ldots, n\}$ порождена случайным вектором $\xi = \{X, Y\}^\top$. Обозначим через $k_{XY} = \mathbf{M}\{(X - m_X)(Y - m_Y)\} = \mathbf{M}\{XY\} - m_X m_Y$ ковариацию случайных величин X и Y.

Определение 1.10. Статистика $\widehat{k}_{XY}(n) = \dfrac{1}{n} \sum\limits_{k=1}^{n} X_k Y_k - \overline{X}_n \overline{Y}_n$ называется *выборочной ковариацией* случайных величин X и Y.

Теорема 1.5. *Если СВ X и Y имеют конечные дисперсии, то:*

1) $\mathbf{M}\left\{\widehat{k}_{XY}(n)\right\} = \dfrac{n-1}{n} k_{XY}$;

2) $\widehat{k}_{XY}(n) \xrightarrow{\text{п.н.}} k_{XY}$, $n \to \infty$;

3) *Если* $\mathbf{M}\{|X|^4 + |Y|^4\} < \infty$, *то*

$$\sqrt{n}\left(\widehat{k}_{XY}(n) - k_{XY}\right) \xrightarrow{d} \eta \sim \mathcal{N}\left(0; \mu_{22} - k^2_{XY}\right), \quad n \to \infty,$$

где $\mu_{22} = \mathbf{M}\{(X - m_X)^2 (Y - m_Y)^2\}$.

1.2. Примеры

Пример 1.1. Рассмотрим исторические данные из учебника [20] о росте взрослых мужчин, родившихся в Соединенном Королевстве (данные взяты из: Final Report of the Anthropometric Committee to the British Association, 1883, p. 256). В табл. 1.1 представлена группировка этих данных для 8585 мужчин. В первой и третьей колонке указан рост мужчины с точностью до одного дюйма. Например, значению 57 соответствует рост в пределах от $56\frac{15}{16}$ дюйма до $57\frac{15}{16}$ дюйма.

Таблица 1.1

Рост	Число мужчин	Рост	Число мужчин
57	2	68	1230
58	4	69	1063
59	14	70	646
60	41	71	392
61	83	72	202
62	169	73	79
63	394	74	32
64	669	75	16
65	990	76	5
66	1223	77	2
67	1329	78	0

Вычислите реализации выборочного среднего, выборочной дисперсии, экстремальных порядковых статистик. Постройте графики реализаций выборочной функции распределения и гистограммы.

Решение.

Реализации выборочного среднего, выборочной дисперсии, экстремальных порядковых статистик, вычисленные по выборке объема $n = 8585$ равны:

$$\overline{x}_n = 67{,}46; \quad \overline{s}^2_n = 6{,}6049; \quad x_{(1)} = 57{,}12; \quad x_{(n)} = 77{,}36.$$

Построим график реализации выборочной функции $\widehat{F}_{100}(x)$ для выборки объема 100 и график функции $\widehat{F}_n(x)$, построенной по всей выборке объема $n = 8585$. Функция $\widehat{F}_{100}(x)$ построена по 100 первым реализациям исходной, неупорядоченной по возрастанию выборке. Построенные графики (см. рис. 1.1) соответствуют кусочно-постоянным функциям. Однако при большом n график функции $\widehat{F}_n(x)$ почти не отличим от гладкой кривой.

На рис. 1.2 приведена реализация гистограммы и график плотности вероятности гауссовской СВ $X \sim \mathcal{N}(67{,}46; 6,60)$:

$$\widetilde{p}(x) = \frac{1}{\sqrt{2\pi} \cdot 6,6049} \exp\left\{ -\frac{(x - 67{,}46)^2}{2 \cdot 6{,}6049} \right\}.$$

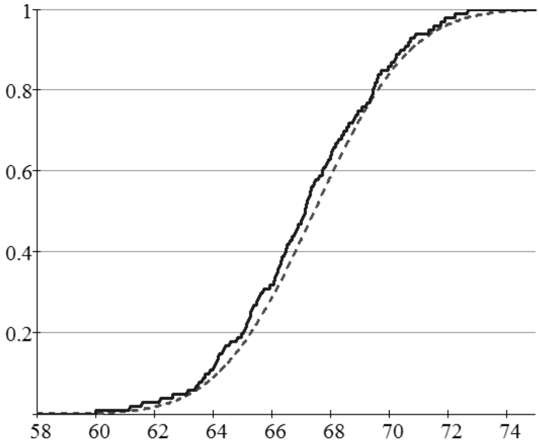


Рис. 1.1. — $\widehat{F}_{100}(x)$ и --- $\widehat{F}_n(x)$

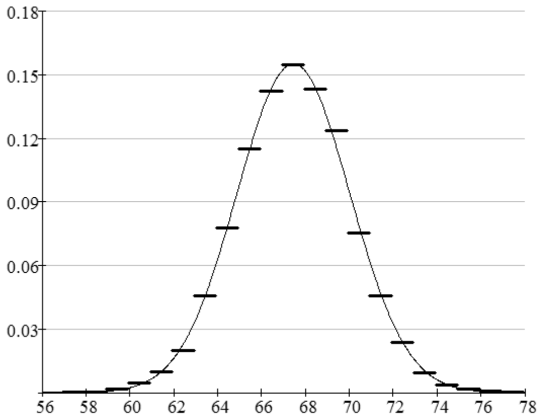


Рис. 1.2. Гистограмма $\widehat{p}_n(x)$ и функция $\widetilde{p}(x)$

Из рис. 1.2. видно, что гистограмма, будучи кусочно-постоянной функцией, удовлетворительно аппроксимируется функцией $\widetilde{p}(x)$, представляющей собой плотность нормального распределения. ■

П р и м е р 1.2. Пусть выборка Z_n соответствует распределению $F(x)$. Докажите, что $X_{(1)}$ и $X_{(n)}$ имеют функции распределения соответственно $F_{(1)}(x) = 1 - (1 - F(x))^n$ и $F_{(n)}(x) = F^n(x)$.

Решение. По определению $X_{(n)} = \max(X_1, \ldots, X_n)$, поэтому $F_{(n)}(x) = \mathbf{P}(X_{(n)} \leqslant x) = \mathbf{P}\left(\{X_1 \leqslant x\} \cdot \{X_2 \leqslant x\} \cdot \ldots \cdot \{X_n \leqslant x\}\right) = \mathbf{P}\left(\prod_{k=1}^{n} \{X_k \leqslant x\}\right)$. Так как элементы выборки статистически независимы и одинаково распределены, получаем

$$F_{(n)}(x) = \prod_{k=1}^{n} \mathbf{P}(X_k \leqslant x) = \prod_{k=1}^{n} F_{X_k}(x) = F^n(x).$$

Аналогично

$$F_{(1)}(x) = 1 - \mathbf{P}(X_{(1)} > x) = 1 - \prod_{k=1}^{n} \mathbf{P}(X_k > x) = $$
$$= 1 - \prod_{k=1}^{n} \left(1 - F_{X_k}(x)\right) = 1 - (1 - F(x))^n. \ \blacksquare$$

Пример 1.3. Пусть выборка Z_n порождена СВ X с конечным моментом ν_r. Докажите, что выборочный начальный момент $\overline{\nu}_r(n)$ обладает по отношению к ν_r *свойством несмещенности*, т.е. $\mathbf{M}\{\overline{\nu}_r(n)\} = \nu_r$, и *свойством сильной состоятельности*, т.е. $\overline{\nu}_r(n) \xrightarrow{\text{п.н.}} \nu_r$ при $n \to \infty$.

Решение. По условию $\mathbf{M}\{(X_k)^r\} = \mathbf{M}\{X^r\} = \nu_r$. Поэтому $\mathbf{M}\{\overline{\nu}_r(n)\} = \mathbf{M}\left\{\frac{1}{n}\sum_{k=1}^{n}(X_k)^r\right\} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{M}\{(X_k)^r\} = \frac{1}{n}\sum_{k=1}^{n}\nu_r = \nu_r$. Свойство несмещенности доказано.

Обозначим $\xi_k = (X_k)^r$, тогда величины $\{\xi_1, \ldots, \xi_n\}$ независимы, одинаково распределены и $\mathbf{M}\{\xi_k\} = \nu_r$. По усиленному закону больших чисел Колмогорова (см. теорему 21.11)

$$\overline{\nu}_r(n) = \frac{1}{n}\sum_{k=1}^{n}\xi_k \xrightarrow{\text{п.н.}} \mathbf{M}\{\xi_1\} = \nu_r \text{ при } n \to \infty.$$

Свойство сильной состоятельности доказано. $\blacksquare$

Пример 1.4. В условиях примера 1.3 для $r = 2$ покажите, что выборочная дисперсия $\overline{S}_n^2$ обладает *свойством асимптотической несмещенности*, т.е. $\mathbf{M}\{\overline{S}_n^2\} \to D_X$, $n \to \infty$, и свойством сильной состоятельности.

Решение. По определению

$$\overline{S}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k^2 - 2X_k\overline{X}_n + \overline{X}_n^{\,2}\right) = \frac{1}{n}\sum_{k=1}^{n}(X_k)^2 -$$
$$-\frac{2}{n}\overline{X}_n\sum_{k=1}^{n}X_k + \frac{n}{n}\overline{X}_n^{\,2} = \frac{1}{n}\sum_{k=1}^{n}(X_k)^2 - (\overline{X}_n)^2 = \overline{\nu}_2(n) - (\overline{\nu}_1(n))^2.$$

Из результата примера 1.3 следует, что $\overline{\nu}_2(n) \xrightarrow{\text{п.н.}} \nu_2$, $\overline{\nu}_1(n) \xrightarrow{\text{п.н.}} \nu_1$, $n \to \infty$. Тогда в силу свойства сходимости почти наверное (см. разд. 21.6) заключаем: $\overline{S}_n^2 = \overline{\nu}_2(n) - (\overline{\nu}_1(n))^2 \xrightarrow{\text{п.н.}} \nu_2 - \nu_1^2 = \mathbf{M}\{X^2\} - (\mathbf{M}\{X\})^2 = D_X$, $n \to \infty$. Свойство сильной состоятельности доказано.

Пусть теперь $\xi_k = (X_k - \overline{X}_n)^2$, а $m_X = \mathbf{M}\{X\}$. Тогда $\mathbf{M}\{\xi_k\} = \mathbf{M}\Big\{\big((X_k - m_X) - (\overline{X}_n - m_X)\big)^2\Big\} = \mathbf{M}\left\{\left(\overset{\circ}{X}_k - \left(\frac{1}{n}\sum_{i=1}^{n}\overset{\circ}{X}_i\right)\right)^2\right\} = \mathbf{M}\Big\{\overset{\circ}{X}{}_k^2\Big\} - \frac{2}{n}\sum_{i=1}^{n}\mathbf{M}\Big\{\overset{\circ}{X}_k\overset{\circ}{X}_i\Big\} + \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbf{M}\Big\{\overset{\circ}{X}_i\overset{\circ}{X}_j\Big\}$. С учетом независимости X_i и X_j при $i \neq j$ получаем $\mathbf{M}\{\xi_k\} = D_X - \frac{2}{n}D_X + \frac{1}{n}D_X = \frac{n-1}{n}D_X$. Поэтому $\mathbf{M}\{\overline{S}_n^2\} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{M}\{\xi_k\} = \frac{n-1}{n}D_X$. Таким образом, $\overline{S}_n^2$ не обладает свойством несмещенности по отношению к дисперсии D_X, так как $\mathbf{M}\{\overline{S}_n^2\} \neq D_X$. Однако $\lim\limits_{n\to\infty}\mathbf{M}\{\overline{S}_n^2\} = \lim\limits_{n\to\infty}\frac{n-1}{n}D_X = D_X$, т.е. свойство асимптотической несмещенности имеет место. ■

Пример 1.5. Выборка $\{X_k,\ k = 1, \ldots, 175\}$ соответствует распределению $R[-1;1]$. Оцените вероятность того, что $|\overline{\nu}_3(175)| \leqslant \frac{1}{70}$.

Решение. По условию $X_k \sim R[-1;1]$, поэтому $\nu_3 = \mathbf{M}\{X_k^3\} = \int\limits_{-1}^{1} x^3 \frac{1}{2}dx = 0$. Так как $n = 175 \gg 1$, то для искомой оценки вероятности можно воспользоваться теоремой 1.2, из которой для $r = 3$ с учетом $\nu_3 = 0$ следует, что $\overline{\nu}_3(175) \sim \mathcal{N}\left(0; \frac{\nu_6}{175}\right)$. Так как $\nu_6 = \frac{1}{2}\int\limits_{-1}^{1} x^6 dx = \frac{1}{7}$, то $\overline{\nu}_3(175) \sim \mathcal{N}\left(0; \frac{1}{1225}\right)$. Отсюда

$$\mathbf{P}\left(|\nu_3(175)| \leqslant \frac{1}{70}\right) \approx \Phi\left(\frac{\sqrt{1225}}{70}\right) - \Phi\left(-\frac{\sqrt{1225}}{70}\right) = 2\Phi_0\left(\frac{1}{2}\right) = 0{,}383.$$

■

Пример 1.6. Выборка Z_n порождена СВ $X \sim R[0;1]$. Для любого $\varepsilon > 0$ оцените $\mathbf{P}\left(|\widehat{F}_n(x) - x| \leqslant \varepsilon\right)$ при каждом $x \in [0;1]$ и $n \gg 1$.

Решение. Так как $X \sim R[0;1]$, функция распределения $F(x) = x$, $x \in [0;1]$. Поэтому $\widehat{F}_n(x) - x = \widehat{F}_n(x) - F(x) \sim \mathcal{N}\left(0; \frac{F(x)(1-F(x))}{n}\right)$ по теореме 1.4. Итак, $\widehat{F}_n(x) - x \sim \mathcal{N}\left(0; \frac{x(1-x)}{n}\right)$ для $x \in [0;1]$ и $n \gg 1$.

Отсюда $\mathbf{P}\left(|\widehat{F}_n(x) - x| \leqslant \varepsilon\right) \approx \Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{x(1-x)}}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{x(1-x)}}\right) = 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{x(1-x)}}\right)$. Так как $x(1-x) \leqslant \frac{1}{4}$, то максимальное значение числа $x(1-x)$ достигается при $x = \frac{1}{2}$, а следовательно, и наихудший результат будет при $x = \frac{1}{2}$. Например, если $\varepsilon = 0{,}1$, $n = 100$, то $\mathbf{P}\left(|\widehat{F}_n(\frac{1}{2}) - \frac{1}{2}| \leqslant 0{,}1\right) \approx 2\Phi_0(2) \approx 0{,}95$. Для сравнения: при $x = 0{,}1$ $\mathbf{P}\left(|\widehat{F}_n(0{,}1) - 0{,}1| \leqslant 0{,}1\right) \approx 2\Phi_0\left(\frac{0{,}1\sqrt{100}}{\sqrt{0{,}09}}\right) \approx 2\Phi_0(3{,}3) \approx 0{,}998$. ■

1.3. Задачи для самостоятельного решения

1. Найдите функцию распределения k-й порядковой статистики $X_{(k)}$, $k = 1, \ldots, n$.

Ответ: $F_{(k)}(x) = \sum\limits_{m=k}^{n} C_n^m F^m(x)(1 - F(x))^{n-m}$.

2. Выборка соответствует распределению $R[0;1]$. Вычислите $\mathbf{M}\left\{X_{(n)}\right\}$ и $\mathbf{D}\left\{X_{(n)}\right\}$.

Ответ: $\mathbf{M}\left\{X_{(n)}\right\} = \frac{n}{n+1}$; $\mathbf{D}\left\{X_{(n)}\right\} = \frac{n}{(n+1)^2(n+2)}$.

3. Выборка соответствует распределению $F(x)$ с конечным моментом ν_r. Докажите, что $\overline{\mu}_r(n) \xrightarrow{\text{п.н.}} \mu_r$, $n \to \infty$.

Указание. Воспользуйтесь формулой $(a-b)^r = \sum\limits_{m=0}^{r} (-1)^m C_r^m a^m b^{r-m}$ и примером 1.4.

4. Выборка объема $n \gg 1$ соответствует распределению $\mathcal{N}(0;\sigma^2)$. Найдите распределение выборочного момента $\overline{\nu}_2(n)$ при $n \to \infty$.

Ответ: $\mathcal{N}\left(\sigma^2; \frac{2\sigma^4}{n}\right)$.

5. Двумерная выборка объема $n \gg 1$ соответствует распределению $\mathcal{N}(\mu; K)$, где ковариационная матрица $K = \begin{bmatrix} D_X & k_{XY} \\ k_{YX} & D_Y \end{bmatrix}$. Докажите, что $\sqrt{n}\left(\widehat{k}_{XY}(n) - k_{XY}\right) \xrightarrow{d} \xi \sim \mathcal{N}\left(0; D_X D_Y + k_{XY}^2\right)$ при $n \to \infty$.

Указание. Вычислите μ_{22}, воспользуйтесь теоремой 1.5.

6. Выборка соответствует распределению $E(\lambda)$, $\lambda > 0$. Найдите предел, к которому почти наверное сходится $\overline{\nu}_2(n)$ при $n \to \infty$.

Ответ: $\frac{2}{\lambda^2}$.

7. Выборка объема $n \gg 1$ порождена СВ $X \sim E(1)$. Оцените $\mathbf{P}\left(|\widehat{F}_n(1) - F_X(1)| \leqslant \frac{1}{\sqrt{n}}\right)$.

Ответ: $2\Phi_0\left(\frac{e}{\sqrt{e-1}}\right)$.

2. Точечные оценки и их свойства

Проблему точечного оценивания можно сформулировать следующим образом. Рассматривается случайная величина, распределение которой принадлежит известному классу распределений, но при этом содержит некоторое число неизвестных параметров. Требуется по выборке, порожденной этой СВ, получить оценки для параметров и определить точность этих оценок. Вообще говоря, существует бесконечное количество различных функций от выборки, которые можно использовать в качестве оценок. Поэтому важно уметь сравнивать свойства различных оценок одного и того же параметра. В частности, для того чтобы оценка была хорошей заменой неизвестному параметру необходимо, чтобы вероятность больших отклонений этой оценки от истинного значения параметра была бы достаточно мала. Желательно также, чтобы при увеличении числа опытов точность результатов оценивания увеличивалась. В связи с этим вводят понятия, определяющие качество построенных оценок.

2.1. Теоретические положения

Пусть $\theta \in \Theta \subseteq \mathbb{R}^1$ — некоторая детерминированная или случайная величина (параметр), а $Z_n = \{X_k,\ k = 1, \ldots, n\}$ — выборка.

Определение 2.1. *Точечной оценкой параметра* θ по выборке Z_n называется любая статистика $\widehat{\theta}_n = \varphi_n(Z_n)$, принимающая значения из множества Θ.

На практике вычисляют реализацию оценки $\widehat{\theta}_n$ (по имеющейся реализации z_n) и принимают ее за приближенное значение параметра θ. Поэтому желательно, чтобы при любом возможном θ величина $\widehat{\theta}_n$ была бы близка к θ.

Определение 2.2. Величина $\Delta\widehat{\theta}_n = \widehat{\theta}_n - \theta$ называется *ошибкой оценки* $\widehat{\theta}_n$.

Определение 2.3. Оценка $\widehat{\theta}_n$ называется *несмещенной*, если $\mathbf{M}\left\{\Delta\widehat{\theta}_n\right\} = 0$. Если же $\mathbf{M}\left\{\Delta\widehat{\theta}_n\right\} \neq 0$, но $\mathbf{M}\left\{\Delta\widehat{\theta}_n\right\} \to 0$, $n \to \infty$, то оценка $\widehat{\theta}_n$ называется *асимптотически несмещенной.*

Часто ограничиваются рассмотрением класса несмещенных оценок. Это требование интуитивно привлекательно: оно означает, что по крайней мере «в среднем» используемая оценка приводит к желаемому результату. К тому же, для класса несмещенных оценок часто удается построить достаточно простую и практически полезную теорию, построение которой невозможно для произвольного класса оценок.

Однако не следует и преувеличивать значение понятия несмещенности: в некоторых случаях это требование оказывается слишком «обременительным» и приводит к нежелательным результатам. Так же может оказаться, что несмещенные оценки значительно уступают по точности (в данной модели) другим оценкам, которые свойством несмещенности не обладают. Следует всегда помнить, что несмещенность не гарантирует того, что ошибка оценки будет маленькой.

Определение 2.4. Оценка $\widehat{\theta}_n$ называется *сильно состоятельной*, если $\Delta\widehat{\theta}_n \xrightarrow{\text{п.н.}} 0$, $n \to \infty$, и *состоятельной в среднем квадратическом* (с.к.-состоятельной), если $\Delta\widehat{\theta}_n \xrightarrow{\text{с.к.}} 0$, $n \to \infty$.

Определение 2.5. *Среднеквадратической погрешностью* (с.к.-погрешностью) оценки $\widehat{\theta}_n$ называется величина

$$\Delta_n = \mathbf{M}\Big\{|\Delta\widehat{\theta}_n|^2\Big\}. \tag{2.1}$$

Введенное понятие состоятельности оценок связано только с предельными свойствами последовательности СВ. Поэтому нужна известная осторожность при использовании состоятельности как критерия качества оценивания в практических задачах. Состоятельность, являющаяся, конечно, желательным свойством всякой процедуры оценивания, напрямую не связана со свойством оценки при фиксированном объеме выборки.

Теорема 2.1. *Пусть* $\theta \in \mathbb{R}^1$ *и* $\mathbf{M}\Big\{|\Delta\widehat{\theta}_n|^2\Big\} < \infty$, *тогда*

$$\Delta_n = l_n^2 + d_n, \tag{2.2}$$

где $l_n = \mathbf{M}\Big\{\Delta\widehat{\theta}_n\Big\}$ *— смещение оценки* $\widehat{\theta}_n$, *а* $d_n = \mathbf{D}\Big\{\Delta\widehat{\theta}_n\Big\}$ *— дисперсия ее ошибки.*

Определение 2.6. Оценка $\widehat{\theta}_n$ называется *асимптотически нормальной*, если существует детерминированная последовательность $\{C_n,\ n = 1, 2, \ldots\}$ такая, что $C_n\Delta\widehat{\theta}_n \xrightarrow{d} \xi \sim \mathcal{N}(0; 1)$, $n \to \infty$.

Пусть теперь оценка $\widehat{\theta}_n = \varphi_n(Z_n)$ принадлежит некоторому заданному классу *допустимых оценок*, т.е. $\varphi_n \in \Phi_n$, $n = 1, 2, \ldots$, где Φ_n — фиксированный класс допустимых преобразований выборки Z_n.

Определение 2.7. Оценка $\widehat{\theta}_n = \varphi(Z_n)$ называется *оптимальной в среднем квадратическом* (с.к.-оптимальной) на Φ_n, если

$$\Delta_n = \mathbf{M}\Big\{|\Delta\widehat{\theta}_n|^2\Big\} \leqslant \mathbf{M}\Big\{|\theta_n - \widetilde{\theta}_n|^2\Big\}, \ n = 1, 2, \ldots,$$

где $\widetilde{\theta}_n$ — произвольная допустимая оценка: $\widetilde{\theta}_n = \psi_n(Z_n)$, $\psi_n \in \Phi_n$.

Если $\theta \in \mathbb{R}^m$, где $m \geqslant 2$, то все вышеприведенные определения остаются в силе со следующими уточнениями:

1) в (2.2) величина $l_n^2 = \delta_n^\top \delta_n$, где $\delta_n = \mathbf{M}\Big\{\Delta\widehat{\theta}_n\Big\} \in \mathbb{R}^m$ — вектор смещения оценки $\widehat{\theta}_n$, а $d_n = \mathrm{tr}[K_n]$, где $K_n = \mathbf{cov}(\Delta\widehat{\theta}_n, \Delta\widehat{\theta}_n)$ — ковариационная матрица ошибки $\Delta\widehat{\theta}_n$, $\mathrm{tr}[A]$ — след матрицы A;

2) в определении 2.6 $\{C_n,\, n = 1, 2, \ldots\}$ — последовательность неслучайных матриц размера $m \times m$, а предельное распределение $\mathcal{N}(0; I)$ — m-мерное стандартное гауссовское распределение.

2.2. Примеры

Пример 2.1. Пусть выборка $\{X_k,\, k = 1, \ldots, n\}$ имеет вид

$$X_k = \theta + \varepsilon_k, \quad k = 1, \ldots, n,$$

где θ — неслучайный скалярный параметр, $\{\varepsilon_k,\, k = 1, \ldots, n\}$ — независимые случайные величины, $\mathbf{M}\{\varepsilon_k\} = 0$, $\mathbf{D}\{\varepsilon_k\} = D_k \leqslant \overline{D} < \infty$ для всех $k \geqslant 1$. Докажите, что выборочное среднее $\overline{X}_n$ является несмещенной и состоятельной оценкой θ.

Решение. По определению 1.7 $\widehat{\theta}_n = \overline{X}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n} X_k$, поэтому

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n}(\theta + \varepsilon_k) = \theta + \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k = \theta + \overline{\varepsilon}_n.$$

Отсюда $\Delta\overline{X}_n = \Delta\widehat{\theta}_n = \overline{X}_n - \theta = \overline{\varepsilon}_n$ — ошибка оценки $\widehat{\theta}_n = \overline{X}_n$. Погрешность $\Delta_n = \mathbf{M}\big\{|\Delta\overline{X}_n|^2\big\} = \mathbf{M}\big\{|\overline{\varepsilon}_n|^2\big\} = \dfrac{1}{n^2}\sum\limits_{k=1}^{n}\mathbf{D}\{\varepsilon_k\} \leqslant \dfrac{\overline{D}}{n} \to 0$, $n \to \infty$. Поэтому оценка $\overline{X}_n$ с.к.-состоятельна.

Докажем теперь сильную состоятельность оценки $\overline{X}_n$. Так как $\mathbf{M}\{\varepsilon_k\} = a_k = 0$, а $\sum\limits_{k=1}^{\infty}\dfrac{D_k}{k^2} \leqslant \sum\limits_{k=1}^{\infty}\dfrac{\overline{D}}{k^2} = \overline{D}\sum\limits_{k=1}^{\infty}\dfrac{1}{k^2} < \infty$, то $\overline{\varepsilon}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n}\varepsilon_k \xrightarrow{\text{п.н.}} 0$, $n \to \infty$ по теореме 21.12. Поэтому $\Delta\overline{X}_n \xrightarrow{\text{п.н.}} 0$,

$n \to \infty$, т.е. $\overline{X}_n \xrightarrow{\text{п.н.}} \theta$, $n \to \infty$, что означает сильную состоятельность $\overline{X}_n$.

Наконец, для любого $n \geqslant 1$ $\mathbf{M}\left\{\Delta\widehat{\theta}_n\right\} = \mathbf{M}\{\Delta\overline{X}_n\} = \mathbf{M}\{\overline{\varepsilon}_n\} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{M}\{\varepsilon_k\} = 0$, т.е. оценка $\overline{X}_n$ — несмещенная. ■

Пример 2.2. Пусть в условиях примера 2.1 СВ $\{\varepsilon_k, k = 1, 2, \ldots\}$ одинаково распределены, причем $\mathbf{M}\{\varepsilon_k\} = 0$, $\mathbf{D}\{\varepsilon_k\} = \sigma^2$, где $\sigma < \infty$. Докажите, что оценка $\widehat{\theta}_n = \overline{X}_n$ асимптотически нормальна.

Решение. Из решения примера 2.1 следует, что $\Delta\overline{X}_n = \overline{\varepsilon}_n$, причем $\mathbf{M}\{\varepsilon_k\} = 0$, $\mathbf{D}\{\varepsilon_k\} = \sigma^2$. Тогда из теоремы 21.14 следует, что $\sqrt{n}\overline{\varepsilon}_n \xrightarrow{d} X \sim \mathcal{N}(0; \sigma^2)$, $n \to \infty$. Отсюда $\frac{\sqrt{n}}{\sigma}\Delta\overline{X}_n = \frac{\sqrt{n}}{\sigma}\overline{\varepsilon}_n \xrightarrow{d} \xi \sim \mathcal{N}(0; 1)$, $n \to \infty$. Таким образом, $C_n\Delta\overline{X}_n \xrightarrow{d} \xi \sim \mathcal{N}(0; 1)$, $n \to \infty$, если $\left\{C_n = \frac{\sqrt{n}}{\sigma}, n = 1, 2, \ldots\right\}$. ■

Пример 2.3. Выборка $\{X_k, k = 1, \ldots, n\}$ порождена СВ $X \sim R[0; \theta]$, $\theta > 0$. Докажите, что $\widehat{\theta}_n = X_{(n)}$ — асимптотически несмещенная оценка параметра θ.

Решение. По условию $F(x) = \mathbf{P}(X \leqslant x) = \frac{x}{\theta}$, $x \in [0; \theta]$. Из примера 1.2 следует, что $F_{(n)}(x) = \mathbf{P}(X_{(n)} \leqslant x) = F^n(x) = \frac{x^n}{\theta^n}$, $x \in [0; \theta]$. Тогда $\mathbf{M}\left\{X_{(n)}\right\} = \int_0^\theta x dF_{(n)}(x) = \int_0^\theta x\frac{nx^{n-1}}{\theta^n}dx = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{n+1}\theta$. Отсюда $\mathbf{M}\left\{\Delta\widehat{\theta}_n\right\} = \mathbf{M}\left\{X_{(n)} - \theta\right\} = \mathbf{M}\left\{X_{(n)}\right\} - \theta = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1} \to 0$, $n \to \infty$. Итак, при любом $\theta > 0$ $\mathbf{M}\left\{\Delta\widehat{\theta}_n\right\} < 0$, поэтому смещение $l_n \neq 0$, но $\mathbf{M}\left\{\Delta\widehat{\theta}_n\right\} \to 0$, $n \to \infty$, т.е. $\widehat{\theta}_n = X_{(n)}$ асимптотически несмещенная. ■

Пример 2.4. Выборка $Z_n = \{X_k, k = 1, \ldots, n\}$ соответствует распределению $\mathcal{N}(m; \theta^2)$. Найдите величину C, при которой статистика $\varphi(Z_n) = \frac{C}{n}\sum_{k=1}^{n}|X_k - m|$ будет несмещенной и сильно состоятельной оценкой параметра θ.

Решение. Пусть $\xi = |X - m|$, где $X \sim \mathcal{N}(m; \theta^2)$. Тогда

$$\mathbf{M}\{\xi\} = \mathbf{M}\{|X - m|\} = \frac{1}{\sqrt{2\pi}\theta} \int\limits_{-\infty}^{\infty} |x - m| \exp\left\{-\frac{(x-m)^2}{2\theta^2}\right\} dx =$$
$$= \sqrt{\frac{2}{\pi}}\theta \int\limits_0^{\infty} y \exp\left\{-\frac{y^2}{2}\right\} dy = \sqrt{\frac{2}{\pi}}\theta.$$

С учетом $\mathbf{M}\{|X_k - m|\} = \mathbf{M}\{\xi\}$ получим, что $\mathbf{M}\{\varphi(Z_n) - \theta\} = \frac{C}{n}\sum_{k=1}^{n} \mathbf{M}\{|X_k - m|\} - \theta = \left(C\sqrt{\frac{2}{\pi}} - 1\right)\theta$. Последнее выражение равно нулю при всех θ, только если $C\sqrt{\frac{2}{\pi}} - 1 = 0$. Отсюда $C = \sqrt{\frac{\pi}{2}}$ есть условие несмещенности оценки $\widehat{\theta}_n = \varphi(Z_n)$.

По усиленному закону больших чисел (см. теорему 21.11) имеем: $\nu_n = \frac{1}{n}\sum_{k=1}^{n} |X_k - m| \xrightarrow{\text{п.н.}} \mathbf{M}\{\xi\} = \sqrt{\frac{2}{\pi}}\theta$, $n \to \infty$. Поэтому $\widehat{\theta}_n = C\nu_n \xrightarrow{\text{п.н.}} C\mathbf{M}\{\xi\} = \theta$, если $C = \sqrt{\frac{\pi}{2}}$. Итак, оценка $\widehat{\theta}_n = \frac{\sqrt{\pi}}{n\sqrt{2}} \sum_{k=1}^{n} |X_k - m|$ — несмещенная и сильно состоятельная оценка среднего квадратического отклонения θ. ■

Пример 2.5. Пусть выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$ порождена СВ X, причем $\mathbf{M}\{X\} = \theta$, а $\mathbf{D}\{X\} = \sigma^2$ — известная величина. Докажите, что оценка $\widehat{\theta}_n = \overline{X}_n$ параметра θ с.к.-оптимальна на классе всех линейных несмещенных оценок вида $\widetilde{\theta}_n = \sum_{k=1}^{n} \alpha_k X_k$, где $\{\alpha_k\}$ — некоторые числовые коэффициенты.

Решение. По условию $\widetilde{\theta}_n$ — несмещенная оценка, поэтому $\mathbf{M}\left\{\widetilde{\theta}_n - \theta\right\} = \mathbf{M}\left\{\sum_{k=1}^{n} \alpha_k X_k - \theta\right\} = \sum_{k=1}^{n} \alpha_k \theta - \theta = \left(\sum_{k=1}^{n} \alpha_k - 1\right)\theta$. Таким образом, условие несмещенности $\mathbf{M}\left\{\widetilde{\theta}_n - \theta\right\} = 0$ влечет условие $\sum_{k=1}^{n} \alpha_k = 1$. Обозначим через Φ_n соответствующий класс оценок. Заметим, что если $\alpha_k = \frac{1}{n}$, $k = 1, \ldots, n$, то $\sum_{k=1}^{n} \alpha_k = 1$, поэтому оценка $\widehat{\theta}_n = \sum_{k=1}^{n} \alpha_k X_k = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}_n$ принадлежит классу Φ_n.

Найдем с.к.-погрешность произвольной оценки $\widetilde{\theta}_n$ из Φ_n. Так как $\mathbf{M}\left\{\widetilde{\theta}_n - \theta\right\} = 0$ по доказанному выше, то $\Delta_n = \mathbf{M}\left\{|\widetilde{\theta}_n - \theta|^2\right\} = \mathbf{D}\left\{\widetilde{\theta}_n - \theta\right\} = \mathbf{D}\left\{\widetilde{\theta}_n\right\} = \sigma^2 \sum_{k=1}^{n} \alpha_k^2$. Таким образом, коэффициенты $\{\widehat{\alpha}_k,\ k = 1, \ldots, n\}$, определяющие оптимальную оценку $\widehat{\theta}_n$, удовлетворяют условию

$$\sum_{k=1}^{n} \widehat{\alpha}_k^2 \leqslant \sum_{k=1}^{n} \alpha_k^2 \text{ для любых } \{\alpha_k\} \text{ таких, что } \sum_{k=1}^{n} \alpha_k = 1.$$

Обозначим $e = \{1, \ldots, 1\}^\top$, $\alpha = \{\alpha_1, \ldots, \alpha_n\}^\top$. Из неравенства Коши—Буняковского следует:

$$1 = \left(\sum_{k=1}^{n} \alpha_k\right)^2 = |(e, \alpha)|^2 \leqslant |e|^2 |\alpha|^2,$$

причем равенство достигается только при $\alpha = \lambda e$. Отсюда $|\alpha|^2 = \sum_{k=1}^{n} \alpha_k^2 \geqslant \frac{1}{|e|^2} = \frac{1}{n}$. Если теперь положить $\widehat{\alpha}_k = \frac{1}{n}$, $k = 1, \ldots, n$, то $\sum_{k=1}^{n} \widehat{\alpha}_k^2 = \frac{1}{n} \leqslant \sum_{k=1}^{n} \alpha_k^2$. Итак, $\widehat{\theta}_n = \sum_{k=1}^{n} \widehat{\alpha}_k X_k = \frac{1}{n} \sum_{k=1}^{n} X_k = \overline{X}_n$ — с.к.-оптимальная оценка на классе Φ_n всех линейных несмещенных оценок. Заметим также, что оценка $\widehat{\theta}_n$ — единственная (в силу единственности набора оптимальных коэффициентов $\{\widehat{\alpha}_k = \frac{1}{n}$, $k = 1, \ldots, n\}$). ■

2.3. Задачи для самостоятельного решения

1. Докажите теорему 2.1.
Указание. Учтите, что $\mathbf{M}\{X^2\} = D_X + m_X^2$.

2. Пусть θ — неслучайный параметр, а $\widehat{\theta}_n$ — его несмещенная оценка. Покажите, что $\Delta_n = \mathbf{D}\left\{\widehat{\theta}_n\right\}$.

3. Выборка $\{X_k,\ k = 1, \ldots, n\}$ порождена СВ X с известным средним $m = \mathbf{M}\{X\}$ и неизвестной дисперсией $\theta = \mathbf{D}\{X\}$. Докажите, что статистика $S_n^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - m)^2$ является несмещенной и сильно состоятельной оценкой параметра θ.

4. Выборка $\{X_k,\ k = 1, \ldots, n\}$ соответствует распределению $R[0; \theta]$, $\theta > 0$. Покажите, что $X_{(n)}$ — с.к.-состоятельная оценка параметра θ.

Указание. Вычислите $\mathbf{M}\left\{(X_{(n)})^2\right\}$ и учтите результат примера 2.3.

5. Выборка $\{X_k,\ k=1,\ldots,n\}$ соответствует распределению $E(\theta)$, $\theta>0$. Докажите, что $\widehat{\theta}_n = \sqrt{\dfrac{2n}{\sum\limits_{k=1}^{n} X_k^2}}$ является сильно состоятельной оценкой параметра θ.

Указание. Найдите п.н.-предел $\xi_n = \dfrac{1}{n}\sum\limits_{k=1}^{n} X_k^2$.

6. Пусть выборка $\{X_k,\ k=1,\ldots,n\}$ соответствует нормальному распределению $\mathcal{N}(\theta,\sigma^2)$, где σ — известно. Покажите, что статистика $T_n = (\overline{X}_n)^2 - \dfrac{\sigma^2}{n}$ несмещенно и сильно состоятельно оценивает функцию $g(\theta)=\theta^2$.

Указание. Покажите, что $\mathbf{M}\{(\overline{X}_n)^2\} = \theta^2 + \dfrac{\sigma^2}{n}$.

7. Выборка $\{X_k,\ k=1,\ldots,n\}$ порождена СВ $X \sim R[0;\theta]$, $\theta>0$. Докажите, что $\widehat{\theta}_n = 2\overline{X}_n$ — несмещенная и сильно состоятельная оценка для θ.

8. Пусть выборка $\{X_k,\ k=1,\ldots,n\}$ соответствует распределению $Bi(N;p)$, где N — известно. Покажите, что статистика $T_n = \dfrac{\overline{X}_n(N-\overline{X}_n)}{N}$ является асимптотически несмещенной и сильно состоятельной оценкой параметра $\theta = \mathbf{D}\{X_1\}$.

Указание. Вычислите $\mathbf{M}\{\overline{X}_n^2\} = \mathbf{D}\{\overline{X}_n\} + \left(\mathbf{M}\{\overline{X}_n\}\right)^2$ и учтите, что $\mathbf{M}\{\overline{X}_n\} = pN$.

3. Методы построения точечных оценок параметров

Часто общие соображения позволяют сделать достаточно определенное заключение о типе функции распределения интересующей нас случайной величины. Например, ссылаясь на центральную предельную теорему, можно считать, что цена на некоторый финансовый инструмент является гауссовской СВ, поскольку она формируется под влиянием большого числа слабо зависимых факторов. В этом случае определение неизвестного закона распределения сводится к оцениванию по результатам наблюдений только неизвестных параметров распределения. Для гауссовского распределения этими параметрами являются математическое ожидание и дисперсия.

В качестве другого примера рассмотрим серию из n опытов, удовлетворяющих схеме испытаний Бернулли. Тогда СВ, являющаяся числом «успешных» опытов в каждой серии, имеет биномиальный закон распределения. Здесь неизвестным параметром распределения будет вероятность «успеха» в одном опыте.

В этом разделе будут рассмотрены два важнейших метода нахождения точечных оценок параметров — метод моментов и метод максимального правдоподобия.

3.1. Теоретические положения

Пусть $Z_n = \{X_k,\ k = 1, 2, \ldots, n\}$ — выборка, порожденная СВ X, функция распределения которой $F_X(x; \theta)$ известна с точностью до m-мерного вектора $\theta = \{\theta_1, \ldots, \theta_m\}^\top$ неизвестных неслучайных параметров. Для построения оценок параметров $\theta_1, \ldots, \theta_m$ по выборке Z_n можно использовать *метод моментов*, если СВ X имеет конечные начальные моменты ν_r для всех $r \leqslant m$.

Алгоритм метода моментов:

1) найдите аналитические выражения для моментов ν_r:

$$\nu_r(\theta) = \mathbf{M}\{X^r\} = \int\limits_{-\infty}^{\infty} x^r dF_X(x; \theta), \quad r = 1, \ldots, m; \tag{3.1}$$

2) вычислите соответствующие выборочные начальные моменты:

$$\overline{\nu}_r(n) = \frac{1}{n} \sum_{k=1}^{n} (X_k)^r, \quad r = 1, \ldots, m; \tag{3.2}$$

3) составьте систему из m уравнений для переменных $\{\theta_1, \ldots, \theta_m\}^\top$, приравнивая соответствующие теоретические (3.1) и выборочные (3.2) моменты:

$$\nu_r(\theta) = \overline{\nu}_r(n), \quad r = 1, \ldots, m; \tag{3.3}$$

4) найдите решение $\widehat{\theta}_n$ системы уравнений (3.3).

Определение 3.1. Решение $\widehat{\theta}_n$ системы уравнений (3.3) называется *оценкой метода моментов* вектора параметров θ закона распределения $F_X(x; \theta)$, которому соответствует выборка.

Заметим, что при составлении системы уравнений (3.3) можно использовать не только начальные моменты, но также и центральные моменты $\mu_r(\theta)$ и $\overline{\mu}_r(n)$, если это удобно.

Основным достоинством метода моментов является простота его практической реализации.

Важнейшим методом построения точечных оценок вектора θ является *метод максимального правдоподобия* (ММП). Предположим, что $\theta \in \Theta$, где Θ — множество допустимых значений вектора θ. Если СВ X, порождающая выборку, является дискретной, то пусть

$$p(x; \theta) = \mathbf{P}\{(X = x; \theta)\}, \quad x \in \mathcal{X}, \tag{3.4}$$

где $\mathcal{X}$ — множество всех возможных значений СВ X, а $\mathbf{P}(X = x; \theta)$ — закон распределения дискретной СВ X.

Если же СВ X абсолютно непрерывна, то

$$p(x; \theta) = \frac{dF(x; \theta)}{dx}, \tag{3.5}$$

т.е. является плотностью вероятности СВ X.

Определение 3.2. *Функцией правдоподобия* выборки Z_n называется функция $L_n(\theta; Z_n)$, $\theta \in \Theta$ вида

$$L_n(\theta; Z_n) = \prod_{k=1}^{n} p\,(X_k; \theta). \tag{3.6}$$

Заметим, что для случая (3.5) функция $L_n(\theta; x)$ является *плотностью вероятности* случайного вектора Z_n в точке $x \in \mathbb{R}^n$.

Определение 3.3. Пусть $\widehat{\theta}_n$ — точка глобального максимума функции $L_n(\theta; Z_n)$ на Θ. Статистика $\widehat{\theta}_n$ называется *оценкой максимального правдоподобия* вектора θ (МП-оценкой).

Итак, $\widehat{\theta}_n = \arg\max\limits_{\theta \in \Theta} L_n(\theta; Z_n)$ — МП-оценка.

Обычно в расчетах используют *логарифмическую функцию правдоподобия*

$$\widetilde{L}_n(\theta) = \ln L_n(\theta; Z_n) = \sum_{k=1}^{n} \ln p\,(X_k; \theta). \tag{3.7}$$

Очевидно, что $\arg\max\limits_{\theta \in \Theta} L_n(\theta; Z_n) = \arg\max\limits_{\theta \in \Theta} \widetilde{L}_n(\theta)$.

Для построения МП-оценки $\widehat{\theta}_n$ можно использовать *необходимые условия экстремума* функции $\widetilde{L}_n(\theta)$:

$$\frac{\partial \widetilde{L}_n(\theta)}{\partial \theta_k} = 0, \quad k = 1, \ldots, m. \tag{3.8}$$

Система уравнений (3.8), решением которой при определенных условиях является оценка $\widehat{\theta}_n$, называется *системой уравнений правдоподобия.*

Следующее утверждение называется *принципом инвариантности* для оценивания по методу максимального правдоподобия.

Теорема 3.1. *Пусть выборка Z_n соответствует распределению $F(x; \theta)$, $\theta \in \Theta$, а функция $g(\theta)$ отображает Θ в некоторый промежуток Δ действительной оси. Тогда, если $\widehat{\theta}_n$ — МП-оценка вектора θ, то $g(\widehat{\theta}_n)$ — МП-оценка функции $g(\theta)$.*

При определенных условиях МП-оценка параметра θ обладает замечательными асимптотическими свойствами. Предположим, что θ_0 — истинное значение скалярного параметра θ, Θ — замкнутое ограниченное подмножество $\mathbb{R}^1$, а θ_0 лежит внутри Θ. Пусть также выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$ соответствует распределению с плотностью вероятности $p(x;\theta)$.

Теорема 3.2. *Пусть выполнены следующие условия:*

1) *при каждом* $\theta \in \Theta$ $\left|\dfrac{\partial^{(k)} p(x;\theta)}{\partial\theta^{(k)}}\right| \leqslant g_k(x)$, $k = 1, 2, 3$, *причем* $g_1(x)$ *и* $g_2(x)$ *интегрируемы на* $\mathbb{R}^1$, *а* $\sup\limits_{\theta\in\Theta} \int\limits_{-\infty}^{\infty} g_3(x) p(x;\theta)dx < \infty$;

2) *при каждом* $\theta \in \Theta$ *функция*

$$i(\theta) = \int\limits_{-\infty}^{\infty} \left[\frac{\partial \ln p(x;\theta)}{\partial\theta}\right]^2 p(x;\theta)dx$$

конечна и положительна.

Тогда уравнение правдоподобия (3.8) *имеет решение* $\widehat{\theta}_n$, *обладающее следующими свойствами*:

а) $\mathbf{M}\left\{\widehat{\theta}_n - \theta_0\right\} \to 0$, $n \to \infty$ (*асимптотическая несмещенность*);

б) $\widehat{\theta} \xrightarrow{\text{п.н.}} \theta_0$, $n \to \infty$ (*сильная состоятельность*);

в) $\sqrt{n\, i(\theta_0)}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \xi \sim \mathcal{N}(0;1)$, $n \to \infty$ (*асимптотическая нормальность*).

Утверждения теоремы 3.2 могут быть обобщены на случай многомерного параметра θ.

3.2. Примеры

Пример 3.1. Выборка Z_n порождена СВ $X \sim R[\theta_1;\theta_2]$, $\theta_1 < \theta_2$. Найдите оценку вектора $\theta = \{\theta_1, \theta_2\}^\top$ методом моментов.

Решение. Известно, что $\nu_1(\theta) = \mathbf{M}\{X\} = \dfrac{\theta_1 + \theta_2}{2}$, а $\mu_2(\theta) = \mathbf{M}\left\{(X - \nu_1(\theta))^2\right\} = \mathbf{D}\{X\} = \dfrac{(\theta_2 - \theta_1)^2}{12}$. Выборочными оценками моментов $\nu_1(\theta)$ и $\mu_2(\theta)$ являются соответственно выборочное среднее и выборочная дисперсия (см. раздел 2):

$$\overline{\nu}_1(n) = \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

$$\overline{\mu}_2(n) = \overline{S}^2_n = \frac{1}{n}\sum_{k=1}^{n} (X_k - \overline{X}_n)^2.$$

Подставляя найденные теоретические и выборочные моменты в систему уравнений метода моментов (3.3), получаем

$$\begin{cases} \theta_1 + \theta_2 = 2\overline{X}_n, \\ \theta_2 - \theta_1 = 2\sqrt{3} \cdot \overline{S}_n. \end{cases}$$

Решая полученную систему уравнений относительно θ_1, θ_2, находим окончательный вид оценок:

$$\widehat{\theta}_1 = \overline{X}_n - \sqrt{3} \cdot \overline{S}_n, \quad \widehat{\theta}_2 = \overline{X}_n + \sqrt{3} \cdot \overline{S}_n. \ \blacksquare$$

Пример 3.2. В условиях примера 3.1 найдите оценки максимального правдоподобия параметров θ_1 и θ_2.

Решение. По условию $p(x;\theta) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1}, & \text{если } x \in [\theta_1, \theta_2]; \\ 0, & \text{если } x \notin [\theta_1, \theta_2]. \end{cases}$

Отсюда

$$L_n(\theta; Z_n) = \begin{cases} \dfrac{1}{(\theta_2 - \theta_1)^n}, & \text{если } X_i \in [\theta_1, \theta_2],\ i = 1, \ldots, n; \\ 0, & \text{если } \exists j : X_j \notin [\theta_1, \theta_2]. \end{cases}$$

Из полученного выражения следует, что при любых $\theta_1 < \theta_2$ $L_n(\theta; Z_n) \leqslant \dfrac{1}{(X_{(n)} - X_{(1)})^n} = L_{\max}$, где $X_{(1)} = \min(X_1, \ldots, X_n)$, $X_{(n)} = \max(X_1, \ldots, X_n)$. Отсюда $\widehat{\theta}_1 = X_{(1)}$, $\widehat{\theta}_2 = X_{(n)}$, так как $L_n(\widehat{\theta}_1, \widehat{\theta}_2; Z_n) = L_{\max}$. Заметим, что МП-оценки $\widehat{\theta}_1$, $\widehat{\theta}_2$ не совпадают с оценками метода моментов, построенными в примере 3.1. $\blacksquare$

Пример 3.3. Пусть выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$ соответствует распределению $Bi(N;\theta)$, где N — известно. Найдите МП-оценку параметра θ (с учетом $\theta \in (0;1)$).

Решение. Из условия следует, что $p(x;\theta) = C_N^x \theta^x (1-\theta)^{N-x}$, где $x = 0, 1, \ldots, N$, а $C_N^x = \dfrac{N!}{x!(N-x)!}$. Поэтому функция правдоподобия имеет вид

$$L_n(\theta; Z_n) = \prod_{k=1}^{n} p(X_k; \theta) = \prod_{k=1}^{n} C_N^{X_k} \theta^{X_k} (1-\theta)^{N-X_k}. \tag{3.9}$$

Логарифмируя (3.9), найдем логарифмическую функцию правдоподобия:

$$\widetilde{L}_n(\theta) = \ln L_n(\theta; Z_n) = \sum_{k=1}^{n} \left(\ln C_N^{X_k} + X_k \ln\theta + (N - X_k)\ln(1-\theta) \right) =$$

$$= \sum_{k=1}^{n} \ln C_N^{X_k} + \ln\theta \sum_{k=1}^{n} X_k + \ln(1-\theta)\left(Nn - \sum_{k=1}^{n} X_k \right).$$

Уравнение правдоподобия (3.8) имеет вид

$$\frac{d\widetilde{L}_n(\theta)}{d\theta} = \frac{1}{\theta}\sum_{k=1}^{n} X_k - \frac{1}{1-\theta}\left(Nn - \sum_{k=1}^{n} X_k\right) = 0.$$

Решая полученное уравнение относительно θ, находим $\widehat{\theta}_n = \dfrac{\sum_{k=1}^{n} X_k}{Nn} = \dfrac{\overline{X}_n}{N}$. Оценка $\widehat{\theta}_n$ будет несмещенной, сильно состоятельной и асимптотически нормальной. ■

Пример 3.4. Дана гауссовская выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$, где $X_k \sim \mathcal{N}(\theta_1; \theta_2)$. Найдите МП-оценку среднего θ_1 и дисперсии $\theta_2 > 0$.

Решение. По условию для $x \in \mathbb{R}^1$ и $\theta = \{\theta_1, \theta_2\}^\top$ имеем $p(x;\theta) = \dfrac{1}{\sqrt{2\pi\theta_2}}\exp\left\{-\dfrac{(x-\theta_1)^2}{2\theta_2}\right\}$. Поэтому

$$L_n(\theta; Z_n) = \prod_{k=1}^{n} p\,(X_k;\theta) = (2\pi\theta_2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\theta_2}\sum_{k=1}^{n}(X_k-\theta_1)^2\right\}.$$

Отсюда

$$\widetilde{L}_n(\theta) = \ln L_n(\theta; Z_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{k=1}^{n}(X_k-\theta_1)^2.$$

Для нахождения максимума функции $\widetilde{L}_n(\theta)$ по θ воспользуемся уравнениями правдоподобия (3.8):

$$\begin{cases} \dfrac{\partial\widetilde{L}_n(\theta)}{\partial\theta_1} = \dfrac{1}{\theta_2}\displaystyle\sum_{k=1}^{n}(X_k-\theta_1) = 0, \\ \dfrac{\partial\widetilde{L}_n(\theta)}{\partial\theta_2} = -\dfrac{n}{2\theta_2} + \dfrac{1}{2\theta_2^2}\displaystyle\sum_{k=1}^{n}(X_k-\theta_1)^2 = 0. \end{cases}$$

Решая полученную систему уравнений относительно θ_1 и θ_2, находим требуемые оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$:

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}_n; \qquad \widehat{\theta}_2 = \frac{1}{n}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2 = \overline{S}_n^2.$$

Итак, выборочное среднее $\overline{X}_n$ и выборочная дисперсия $\overline{S}_n^2$ являются МП-оценками соответственно математического ожидания θ_1 и дисперсии θ_2 по гауссовской выборке. ■

Из результатов примеров 1.3 и 1.4 следует, что $\widehat{\theta}_1$ — несмещенная и сильно состоятельная оценка θ_1, $\widehat{\theta}_2$ — асимптотически несмещенная и сильно состоятельная оценка для θ_2. Можно показать, что обе оценки асимптотически нормальны.

3.3. Задачи для самостоятельного решения

1. Докажите, что оценки параметров, построенные в примерах 3.1 и 3.3, являются асимптотически несмещенными и сильно состоятельными.

Указание. Используйте асимптотические свойства выборочных моментов.

2. Выборка объема n порождена СВ $X \sim E(\theta)$, $\theta > 0$. Найдите МП-оценку параметра θ и докажите ее сильную состоятельность.

Ответ: $\widehat{\theta}_n = \dfrac{1}{\overline{X}_n}$.

3. Выборка объема n соответствует распределению Пуассона $\Pi(\theta)$, $\theta > 0$. Найдите МП-оценку для θ, докажите ее несмещенность, сильную состоятельность и асимптотическую нормальность.

Ответ: $\widehat{\theta}_n = \overline{X}_n$.

4. Выборка $\{X_k,\ k = 1, \ldots, n\}$ порождена СВ $X \sim E(\theta_1, \theta_2)$, $\theta_2 > 0$, т.е.

$$p_X(x) = \begin{cases} \theta_2 \exp\{-\theta_2(x - \theta_1)\}, & \text{если } x \geqslant \theta_1, \\ 0, & \text{если } x < \theta_1. \end{cases}$$

Найдите оценки параметров θ_1 и θ_2 методом моментов. Докажите сильную состоятельность полученных оценок.

Ответ: $\widehat{\theta}_1 = \overline{X}_n - \overline{S}_n$; $\widehat{\theta}_2 = \dfrac{1}{\overline{S}_n}$.

5. Выборка Z_n соответствует распределению Рэлея с функцией распределения $F(x;\theta) = 1 - \exp\left\{-\dfrac{x^2}{\theta}\right\}$, $x \geqslant 0$, $\theta > 0$. Найдите МП-оценку параметра θ.

Ответ: $\widehat{\theta}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n}(X_k)^2$.

6. Выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$ объема $n = 2m + 1$ (m — натуральное) соответствует распределению Лапласа с плотностью $p(x;\theta) = \dfrac{1}{2}\exp\{-|x - \theta|\}$. Найдите МП-оценку параметра θ.

Ответ: $\widehat{\theta}_n = X_{(m+1)}$.

7. В условиях задачи 6 для случая $n = 2m$ покажите, что МП-оценкой для θ является любая статистика вида $\widehat{\theta}_n = (1-\lambda)X_{(m)} + \lambda X_{(m+1)}$, $\lambda \in [0;1]$.

8. Пусть $\widehat{\theta}_n$ — МП-оценка параметра θ распределения Бернулли $Bi(1;\theta)$. Покажите, что последовательность $\sqrt{n}(\widehat{\theta}_n - \theta)$ асимптотически нормальна с параметрами $(0; \theta(1-\theta))$.

Указание. См. пример 3.3.

9. Выборка Z_n соответствует нормальному распределению с параметрами $(\sqrt{\theta}; 2)$, $\theta \geqslant 0$. Найдите МП-оценку для θ.

Ответ: $\widehat{\theta}_n = (\overline{X}_n)^2$, если $\overline{X}_n \geqslant 0$, и $\widehat{\theta}_n = 0$ — в противном случае.

10. Выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$ соответствует логнормальному распределению с параметрами $(\theta_1; \theta_2)$, т.е. $\ln X_k \sim \mathcal{N}(\theta_1; \theta_2)$. Найдите МП-оценку параметра $\theta = \mathbf{M}\{X_k\}$, докажите ее сильную состоятельность.

Указание. Покажите, что $\theta = \exp\left\{\theta_1 + \dfrac{\theta_2}{2}\right\}$; воспользуйтесь принципом инвариантности.

Ответ: $\widehat{\theta}_n = \exp\left\{\overline{Y}_n + \dfrac{\overline{S}_n^2}{2}\right\}$, где $\overline{Y}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n} Y_k$, $\overline{S}_n^2 = \dfrac{1}{n}\sum\limits_{k=1}^{n}(Y_k - \overline{Y}_n)^2$, $Y_k = \ln X_k$, $k = 1, \ldots, n$.

11. Найдите МП-оценку параметра $\theta = \mathbf{M}\{X_k\}$ по выборке $\{X_k,\ k = 1, \ldots, n\}$, соответствующей распределению $R[\theta_1, \theta_2]$.

Ответ: $\widehat{\theta}_n = \dfrac{X_{(1)} + X_{(n)}}{2}$.

4. Эффективность точечных оценок

Если потребовать от оценки некоторого параметра, чтобы она была несмещенной, то может оказаться, что таких оценок бесконечно много. Предположим для простоты, что у нас есть выборка, порожденная СВ X всего из двух наблюдений, и мы хотим оценить математическое ожидание наблюдаемой СВ. Любое взвешенное среднее этих наблюдений будет несмещенной оценкой математического ожидания. Таким образом, количество несмещенных оценок может быть бесконечным. Как выбрать лучшую из них? Что является мерой сравнения качества двух оценок? Возникает задача построения оценки, которая является наилучшей в некотором смысле. Одним из ответов на этот вопрос и является эффективная оценка, определению и построению которой посвящен этот раздел.

4.1. Теоретические положения

Для определенности предположим, что выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$ соответствует абсолютно непрерывному распределению $F(x; \theta)$ с плотностью $p(x; \theta)$, где $\theta \in \Theta \subseteq \mathbb{R}^1$, Θ — произвольный промежуток.

Определение 4.1. Распределение $F(x; \theta)$ называется *регулярным*, если выполнены следующие два условия:

R.1) функция $\sqrt{p(x; \theta)}$ непрерывно дифференцируема по θ на Θ для почти всех x (по мере Лебега);

R.2) функция

$$i(\theta) = \mathbf{M}_\theta \left\{ \left(\frac{\partial \ln p(X;\theta)}{\partial \theta} \right)^2 \right\} = \int\limits_{-\infty}^{\infty} \left(\frac{\partial \ln p(x;\theta)}{\partial \theta} \right)^2 p(x;\theta) dx \qquad (4.1)$$

конечна, положительна и непрерывна по θ на Θ.

В формуле (4.1) СВ X имеет плотность распределения $p(x;\theta)$, $\theta \in \Theta$, а $\mathbf{M}_\theta \{\xi\}$ означает усреднение СВ ξ по этому распределению.

Определение 4.2. Функция $i(\theta)$ называется *информационным количеством Фишера одного наблюдения* с распределением $p(x;\theta)$.

Если СВ X, порождающая выборку Z_n, является дискретной, а $\mathcal{X}$ — множество ее допустимых значений, $p(x;\theta) = \mathbf{P}_\theta(X = x)$, $x \in \mathcal{X}$, $\theta \in \Theta$, а в формуле (4.1) интеграл заменяется суммой:

$$i(\theta) = \mathbf{M}_\theta \left\{ \left(\frac{\partial \ln p(X;\theta)}{\partial \theta} \right)^2 \right\} = \sum_{x \in \mathcal{X}} \left(\frac{\partial \ln p(x;\theta)}{\partial \theta} \right)^2 p(x;\theta). \qquad (4.2)$$

Пусть $L_n(\theta; Z_n)$ — функция правдоподобия выборки Z_n (см. определение 3.2), а $\widehat{L}_n(\theta) = \ln L_n(\theta; Z_n)$ — логарифмическая функция правдоподобия.

Определение 4.3. Функция

$$U_n(\theta; Z_n) = \frac{d\widehat{L}_n(\theta)}{d\theta} \qquad (4.3)$$

называется *вкладом выборки* Z_n.

Определение 4.4. Функция $I_n(\theta)$, определенная на Θ формулой

$$I_n(\theta) = \mathbf{M}_\theta \left\{ U_n^2(\theta; Z_n) \right\} = \int\limits_{\mathbb{R}^n} U_n^2(\theta; x) L_n(\theta; x) dx, \qquad (4.4)$$

называется *количеством информации Фишера* о параметре θ, содержащемся в выборке Z_n, соответствующей распределению $p(x;\theta)$, $\theta \in \Theta$.

Теорема 4.1. *Пусть выполнены условия регулярности* $R.1$ *и* $R.2$, *тогда* $I_n(\theta) = n\, i(\theta)$, *где* $i(\theta)$ *имеет вид* (4.1) *или* (4.2).

Пусть $\widehat{\theta}_n$ — произвольная несмещенная оценка для θ, построенная по выборке Z_n: $\mathbf{M}\left\{\widehat{\theta}_n - \theta\right\} = 0$. Пусть также $\Delta_n = \mathbf{M}\left\{|\widehat{\theta}_n - \theta|^2\right\}$ — с.к.-погрешность оценки $\widehat{\theta}_n$.

Теорема 4.2 (неравенство Рао—Крамера). *Пусть выполнены условия регулярности* $R.1$ *и* $R.2$, *тогда справедливы следующие утверждения*:

1)
$$\Delta_n \geqslant \frac{1}{I_n(\theta)} = \Delta_n^{\min}, \tag{4.5}$$

где $\Delta_n^{\min}$ *— нижняя граница Рао—Крамера с.к.-погрешности несмещенной оценки* $\widehat{\theta}_n$*;*

2) *если в* (4.5) *для некоторой оценки* $\widehat{\theta}_n$ *достигается равенство, то ее можно представить в виде*

$$\widehat{\theta}_n = \theta + a(\theta) U_n(\theta; Z_n), \tag{4.6}$$

где $a(\theta)$ *— детерминированная функция, а* $U_n(\theta; Z_n)$ *— вклад выборки* (4.3)*.*

Определение 4.5. Несмещенная оценка $\widehat{\theta}_n$, с.к.-погрешность которой совпадает при всех $n \geqslant 1$ с нижней границей $\Delta_n^{\min}$, называется *эффективной по Рао—Крамеру.*

Из приведенных определений и утверждений следует:

1) эффективная оценка является с.к.-оптимальной на классе всех несмещенных оценок параметра θ;

2) если эффективная оценка существует, то она имеет вид (4.6).

Следующее утверждение поясняет связь между эффективной оценкой и МП-оценкой.

Теорема 4.3. *Пусть в условиях теоремы 4.2 существует эффективная оценка* $\widehat{\theta}_n$*, тогда она единственна и является МП-оценкой.*

Заметим, что с учетом несмещенности эффективной оценки $\widehat{\theta}_n$ и утверждения теоремы 4.1

$$\Delta_n(\theta) = \mathbf{M}\Big\{(\widehat{\theta}_n - \theta)^2\Big\} = \mathbf{D}\Big\{\widehat{\theta}_n\Big\} = \frac{1}{n\, i(\theta_0)}, \tag{4.7}$$

где $i(\theta)$ — информация Фишера одного наблюдения (4.1) или (4.2).

Из (4.7) видно, что $\mathbf{D}\Big\{\widehat{\theta}_n\Big\} = O\left(\frac{1}{n}\right)$, т.е. убывает с ростом объема выборки со скоростью, пропорциональной $\frac{1}{n}$. Кроме того, всякая эффективная оценка с.к.-состоятельна, так как $\Delta_n = \mathbf{D}\Big\{\widehat{\theta}_n\Big\} \to 0$, $n \to \infty$.

Определение 4.6. Если выборка соответствует регулярному распределению, $\theta \in \Theta \subseteq \mathbb{R}^1$, а для некоторой несмещенной оценки $\widehat{\theta}_n$ выполнено $\dfrac{\mathbf{D}\Big\{\widehat{\theta}_n\Big\}}{\Delta_n^{\min}} \to 1$, $n \to \infty$, то $\widehat{\theta}_n$ называется *асимптотически эффективной* оценкой.

Для случая $\theta \in \Theta \subseteq \mathbb{R}^m$, $m > 1$ условия регулярности R.1 и R.2 принимают следующий вид:

R.1') $\sqrt{p(x;\theta)}$ непрерывно дифференцируема по θ_j, $j = 1, \ldots, m$ на Θ для почти всех x;

R.2') матрица $I(\theta) = \{I_{ij}(\theta)\}$ с элементами

$$I_{ij}(\theta) = \int\limits_{-\infty}^{\infty} \frac{\partial \ln p(x;\theta)}{\partial \theta_i} \cdot \frac{\partial \ln p(x;\theta)}{\partial \theta_j}\, p(x;\theta)dx \tag{4.8}$$

непрерывна по θ на Θ и положительно определена.

В этом случае неравенство Рао—Крамера (4.5) принимает вид

$$\mathbf{M}\left\{(\widehat{\theta}_n - \theta)(\widehat{\theta}_n - \theta_0)^\top\right\} \geqslant \frac{1}{n} I^{-1}(\theta), \tag{4.9}$$

где $\widehat{\theta}_n$ — произвольная несмещенная оценка параметра θ. Знак неравенства в (4.9) имеет следующий смысл: если матрицы A и B симметричны и неотрицательно определены, то $A \geqslant B$ означает, что $A - B$ неотрицательно определена. Матрица $I(\theta)$ называется *информационной матрицей Фишера.*

4.2. Примеры

Пример 4.1. Выборка Z_n соответствует распределению $\mathcal{N}(\theta; \sigma^2)$, $\sigma > 0$. Докажите, что выборочное среднее $\overline{X}_n$ является эффективной оценкой математического ожидания θ.

Решение. По условию $p(x;\theta) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\dfrac{(x-\theta)^2}{2\sigma^2}\right\}$, поэтому условие R.1 очевидно выполнено. Проверим условие R.2. Пусть $X \sim \mathcal{N}(\theta; \sigma^2)$, тогда

$$l(X;\theta) = \ln p(X;\theta) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(X-\theta)^2}{2\sigma^2}.$$

Отсюда $\varphi(X;\theta) = \dfrac{\partial l(X;\theta)}{\partial \theta} = \dfrac{X-\theta}{\sigma^2}$, и, следовательно,

$$i(\theta) = \mathbf{M}_\theta\left\{\varphi^2(X;\theta)\right\} = \mathbf{M}_\theta\left\{\frac{(X-\theta)^2}{\sigma^4}\right\} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Итак, информация Фишера для гауссовского распределения $i(\theta) = \dfrac{1}{\sigma^2}$ удовлетворяет R.2 при любом $\sigma \in (0, +\infty)$.

Теперь видно, что нижняя граница в неравенстве (4.5) Рао—Крамера $\Delta_n^{\min} = \dfrac{1}{n\, i(\theta)} = \dfrac{\sigma^2}{n}$ и не зависит от θ.

Так как $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ по определению, то $\mathbf{M}\{\overline{X}_n\} = = \frac{1}{n}\sum_{k=1}^{n}\mathbf{M}\{X_k\} = \frac{n\theta}{n} = \theta$, т.е. $\overline{X}_n$ — несмещенная оценка. При этом

$$\Delta_n = \mathbf{D}\{\overline{X}_n\} = \mathbf{D}\left\{\frac{1}{n}\sum_{k=1}^{n} X_k\right\} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Итак, $\Delta_n = \Delta_n^{\min}$, поэтому $\overline{X}_n$ — эффективная оценка для θ при любом $\sigma > 0$. ■

Пример 4.2. Покажите, что распределение Бернулли $Bi(1;\theta)$, $\theta \in (0;1)$ является регулярным, и найдите информацию Фишера $i(\theta)$.

Решение. По условию $p(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0,1$, а $\theta \in \Theta = (0;1)$. Обозначим $f(x;\theta) = \frac{\partial\sqrt{p(x;\theta)}}{\partial\theta} = -\frac{\partial p(x;\theta)}{\partial\theta}\frac{1}{2\sqrt{p(\theta;x)}}$. Если $x = 0$, то $p(x;\theta) = 1-\theta$, и, следовательно, $f(x;\theta) = -\frac{1}{2\sqrt{1-\theta}}$ непрерывна по θ на $(0;1)$. Аналогично для $x = 1$ $p(x;\theta) = \theta$, т.е. $f(x;\theta) = -\frac{1}{2\sqrt{\theta}}$ также непрерывна по θ на Θ. Таким образом, условие регулярности R.1 выполнено.

Теперь найдем $i(\theta)$. Если $X \sim Bi(1;\theta)$, то $p(X;\theta) = \theta^X(1-\theta)^{1-X}$. Отсюда $l(X;\theta) = \ln p(X;\theta) = X\ln\theta + (1-X)\ln(1-\theta)$. Поэтому $\varphi(X;\theta) = \frac{\partial l(X;\theta)}{\partial\theta} = \frac{X}{1-\theta} - \frac{1-X}{\theta} = \frac{X-\theta}{\theta(1-\theta)}$. Теперь

$$i(\theta) = \mathbf{M}_\theta\left\{\varphi^2(X;\theta)\right\} = \frac{\mathbf{M}_\theta\left\{(X-\theta)^2\right\}}{\theta^2(1-\theta)^2} = $$
$$= \frac{\mathbf{D}_\theta\{X\}}{\theta^2(1-\theta)^2} = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.$$

Видно, что $0 < i(\theta) < \infty$ при любом $\theta \in \Theta$ и $i(\theta)$ непрерывна по θ на Θ, т.е. условие R.2 также выполнено. ■

Пример 4.3. Докажите, что частота $\widehat{\theta}_n = P_n^*(A)$ случайного события A является эффективной оценкой вероятности $\theta = \mathbf{P}(A)$ этого события.

Решение. По определению *частота* $P_n^*(A) = \frac{1}{n}\sum_{k=1}^{n} X_k$, где $X_k \sim Bi(1;\theta)$ — независимые бернуллиевские СВ. Поэтому $\mathbf{M}\{P_n^*(A)\} = \theta$, а $\mathbf{D}\{P_n^*(A)\} = \frac{\mathbf{D}\{X_1\}}{n} = \frac{\theta(1-\theta)}{n}$. Из примера 4.2 следует, что количество информации Фишера в выборке $Z_n = \{X_k,$ $k = 1,\ldots,n\}$ о параметре θ равно $I_n(\theta) = n\,i(\theta) = \frac{n}{\theta(1-\theta)}$. Поэтому $\mathbf{D}\left\{\widehat{\theta}_n\right\} = \mathbf{D}\{P_n^*(A)\} = \dfrac{1}{I_n(\theta)}$, т.е. в неравенстве Рао—Крамера достигается нижняя граница. Таким образом, $\widehat{\theta}_n = P_n^*(A)$ эффективно оценивает $\theta = \mathbf{P}(A)$. Применимость теоремы Рао—Крамера в данном случае обосновывается регулярностью распределения $Bi(1;\theta)$ для всех $\theta \in (0;1)$, что было доказано в примере 4.2. ■

Следующий пример показывает, что выборочное среднее отнюдь не всегда является эффективной оценкой математического ожидания.

Пример 4.4. Выборка $\{X_k,\ k = 1, \ldots, n\}$ соответствует распределению Лапласа с параметрами (θ, λ), где $\lambda > 0$, т.е.

$$p(x;\theta) = \frac{1}{2\lambda}\exp\left\{-\frac{|x-\theta|}{\lambda}\right\}, \quad \theta \in \mathbb{R}^1. \tag{4.10}$$

Докажите, что $\overline{X}_n$ является несмещенной, но не эффективной оценкой среднего θ при любом известном λ.

Решение. Можно показать, что в условиях примера неравенство (4.5) выполнено, причем $I_n(\theta) = \dfrac{n}{\lambda^2}$.

Если СВ X имеет распределение (4.10), тогда $\mathbf{M}\{X\} = \dfrac{1}{2\lambda}\displaystyle\int_{-\infty}^{\infty} x\exp\left\{-\frac{|x-\theta|}{\lambda}\right\}dx = \theta + \frac{\lambda}{2}\int_{-\infty}^{\infty} y\exp\{-|y|\}\,dy = \theta$, поэтому $\mathbf{M}\left\{\overline{X}_n\right\} = \mathbf{M}\{X\} = \theta$ (см. решение примера 4.1). Далее $\mathbf{D}\left\{\overline{X}_n\right\} = \dfrac{\mathbf{D}\{X\}}{n} = \dfrac{2\lambda^2}{n}$, так как $\mathbf{D}\{X\} = \dfrac{1}{2\lambda}\displaystyle\int_{-\infty}^{\infty}(x-\theta)^2\exp\left\{-\frac{|x-\theta|}{\lambda}\right\}dx = 2\lambda^2$.

Отсюда видно, что $\Delta_n = \mathbf{D}\left\{\overline{X}_n\right\} = \dfrac{2\lambda^2}{n} > \dfrac{1}{I_n(\theta)} = \dfrac{\lambda^2}{n} = \Delta_n^{\min}$. Таким образом, с.к.-погрешность Δ_n оценки $\overline{X}_n$ параметра θ в 2 раза больше нижней границы Рао—Крамера $\Delta_n^{\min}$ при любом объеме выборки n и любом $\theta \in \mathbb{R}^1$. Последнее означает, что $\overline{X}_n$ не может быть эффективной оценкой для θ. Более того, $\overline{X}_n$ не является даже асимптотически эффективной, так как $\dfrac{\Delta_n}{\Delta_n^{\min}} \not\to 1$, $n \to \infty$. ■

Приведем пример нерегулярного распределения и рассмотрим точность МП-оценки параметра этого распределения.

Пример 4.5. Покажите, что распределение $R[0;\theta]$, $\theta > 0$ нерегулярно. Исследуйте поведение с.к.-погрешности МП-оценки параметра θ при $n \to \infty$.

Решение. Зафиксируем любое $x > 0$. По условию

$$p(x;\theta) = \begin{cases} 0, & \text{если } \theta < x, \\ \dfrac{1}{\theta}, & \text{если } \theta \geqslant x. \end{cases}$$

Таким образом, $\sqrt{p(x;\theta)}$ терпит разрыв в точке $\theta = x$ и, естественно, не является непрерывно дифференцируемой при любом $x > 0$. Итак, условие R.1 нарушено.

Пусть $\widehat{\theta}_n$ — МП-оценка параметра θ, тогда $\widehat{\theta}_n = X_{(n)} = \max\{X_1, \ldots, X_n\}$ (см. пример 3.2). В примере 1.2 было показано, что $X_{(n)} \sim F_{(n)}(x) = \dfrac{x^n}{\theta^n}$, если $x \in [0;\theta]$, поэтому

$$p(x;\theta) = \begin{cases} \dfrac{nx^{n-1}}{\theta^n}, & \text{если } x \in [0;\theta], \\ 0, & \text{если } x \notin [0;\theta]. \end{cases}$$

Отсюда немедленно следует, что для любого $\theta \in \Theta$

$$\mathbf{M}\left\{\widehat{\theta}_n\right\} = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{n+1}\theta.$$

Поэтому «подправленная» оценка $\widetilde{\theta}_n = \dfrac{n+1}{n}\widehat{\theta}_n$ — несмещенная. Найдем теперь дисперсию оценки $\widetilde{\theta}_n$:

$$\mathbf{M}\left\{(\widetilde{\theta}_n)^2\right\} = \left(\frac{n+1}{n}\right)^2 \mathbf{M}\left\{\left(\widehat{\theta}_n\right)^2\right\} = \left(\frac{n+1}{n}\right)^2 \int_0^\theta x^2 \frac{nx^{n-1}}{\theta^n} dx =$$
$$= \frac{(n+1)^2}{n\theta^n}\int_0^\theta x^{n+1}dx = \frac{(n+1)^2}{n(n+2)}\theta^2.$$

Поэтому $\mathbf{D}\left\{\widetilde{\theta}_n\right\} = \mathbf{M}\left\{(\widetilde{\theta}_n)^2\right\} - \theta^2 = \left[\dfrac{(n+1)^2}{n(n+2)} - 1\right]\theta^2 = \dfrac{\theta^2}{n(n+2)} = O\left(\dfrac{1}{n^2}\right)$.

Итак, мы видим, что для θ найдена несмещенная оценка, с.к.-погрешность которой убывает существенно быстрее, чем $O\left(\dfrac{1}{n}\right)$, что «разрешено» неравенством Рао—Крамера. Указанный эффект вызван нерегулярностью распределения $R[0;\theta]$ и известен как «сверхэффективность» оценки $\widetilde{\theta}_n$. ■

4.3. Задачи для самостоятельного решения

1. Выборка соответствует распределению $Bi(N;\theta)$, $\theta \in (0;1)$. Проверьте условия регулярности, найдите $i(\theta)$ и докажите эффективность МП-оценки параметра θ.

Указание. $\widehat{\theta}_n = \dfrac{\overline{X}_n}{N}$.

Ответ: $i(\theta) = \dfrac{N}{\theta(1-\theta)}$.

2. Покажите, что распределение Пуассона $\Pi(\theta)$, $\theta > 0$ регулярно. Найдите $i(\theta)$. Докажите эффективность МП-оценки $\widehat{\theta}_n$ параметра θ.

Указание. $\widehat{\theta}_n = \overline{X}_n$.

Ответ: $i(\theta) = \dfrac{1}{\theta}$.

3. Для распределения $\mathcal{N}(\mu;\theta^2)$, $\theta > 0$, где μ — известно, найдите информацию Фишера $i(\theta)$.

Ответ: $i(\theta) = \dfrac{2}{\theta^2}$.

4. Проверьте регулярность распределения $E(\theta)$, $\theta > 0$, вычислите $I_n(\theta)$. Докажите, что оценка $\widetilde{\theta}_n = \dfrac{n-1}{n}\widehat{\theta}_n$ асимптотически эффективна, если $\widehat{\theta}_n$ — МП-оценка для θ.

Указание. $\widehat{\theta}_n = \dfrac{1}{\overline{X}_n}$.

Ответ: $I_n(\theta) = \dfrac{n}{\theta^2}$; $\mathbf{D}\left\{\widetilde{\theta}_n\right\} = \dfrac{\theta^2}{n-2}$.

5. Покажите, что информация $I_n(\theta)$, содержащаяся в выборке Z_n, соответствующей распределению Лапласа (4.10), равна $\dfrac{n}{\lambda^2}$.

6. Сравните по точности оценку θ_n^* параметра θ распределения $R[0;\theta]$, $\theta > 0$, полученную методом моментов, с оценкой $\widetilde{\theta}_n$, рассмотренной в примере 4.5.

Указание. $\theta_n^* = 2\overline{X}_n$.

Ответ: $\dfrac{\mathbf{D}\{\theta_n^*\}}{\mathbf{D}\left\{\widetilde{\theta}_n\right\}} = \dfrac{n+2}{3}$.

7. Пусть выборка соответствует распределению $\mathcal{N}(\mu;\theta)$, $\theta > 0$, μ — известно. Докажите, что МП-оценка дисперсии θ эффективна.

Указание. $\widehat{\theta}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n}(X_k-\mu)^2$.

Ответ: $\mathbf{D}\left\{\widehat{\theta}_n\right\} = \dfrac{2\theta^2}{n}$; $i(\theta) = \dfrac{1}{2\theta^2}$.

8. Выборка Z_n соответствует распределению $\mathcal{N}(\theta_1;\theta_2^2)$. Найдите информационную матрицу Фишера $I(\theta_1,\theta_2)$.

Ответ: $I(\theta_1,\theta_2) = \begin{bmatrix} \dfrac{1}{\theta_2^2} & 0 \\ 0 & \dfrac{2}{\theta_2^2} \end{bmatrix}$.

9. Для оценок, построенных в задачах 1, 2 и 7, найдите их представление (4.6) через вклад выборки.

Ответ: Во всех случаях $a(\theta) = \dfrac{1}{I_n(\theta)}$.

5. Интервальные оценки параметров

В разделе 2 были рассмотрены точечные оценки неизвестных параметров. Неменьший интерес представляют процедуры оценивания параметров, связанные с построением интервала, который накрывает неизвестный параметр с заданной вероятностью. Важность построения таких интервалов связана с тем, что результаты экспериментов случайны, и оценка, как функция случайной величины, также является случайной величиной. А следовательно, реализация любой обладающей самыми лучшими статистическими свойствами, оценки, вообще говоря, не совпадает с оцениваемым параметром. Имея лишь точечную оценку параметра, мы не можем судить о том, каково отклонение построенной оценки от истинного значения параметра. Если же удается указать интервал, внутри которого с достаточно высокой вероятностью находится истинное значение параметра, то длина этого интервала будет характеризовать точность оценивания. Например, в разделе 1 была построена точечная оценка математического ожидания роста человека по выборке конечного объема. Насколько точна оценка, построенная по этой выборке? Какие отклонения от истинного параметра допускаются с вероятностью 0,95 или 0,99? Ответ на эти вопросы мы получим, построив интервальные оценки математического ожидания с заданным уровнем надежности.

5.1. Теоретические положения

Пусть выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$ соответствует распределению $F(x; \theta)$, где $\theta \in \Theta \subseteq \mathbb{R}^1$ — неизвестный параметр. Выберем некоторое малое положительное число p и предположим, что найдутся статистики $T_1 = T_1(Z_n)$ и $T_2 = T_2(Z_n)$, $T_1 < T_2$, такие, что для любого $\theta \in \Theta$

$$\mathbf{P}(T_1 \leqslant \theta \leqslant T_2) = 1 - p. \tag{5.1}$$

Определение 5.1. Промежуток $[T_1, T_2]$ называется *доверительным интервалом для* θ *надежности* $q = 1 - p$. Доверительный интервал также называют *интервальной оценкой* параметра θ.

Число $p = 1 - q$ называют *уровнем значимости*, и обычно на практике полагают $p = 0{,}05$ или $p = 0{,}01$.

Выбор уровня значимости в значительной степени зависит от той цели, которую мы перед собой ставим. Например, если оценивается вероятность посадки самолета на посадочную полосу, то неприемлемым может оказаться даже уровень 0,01, так как он означает, что в среднем в одном случае из ста самолет будет вынужден уйти на второй круг или вообще садиться на запасной аэродром. С другой стороны, при статистических исследованиях в биологии и медицине имеется так много дополнительных источников ошибок (недостоверность теоретических предположений, упрощающие допущения и т.д.), что дополнительная ошибка от применения статистики, соответствующей уровню значимости 0,01, представляется сравнительно безобидной.

Пусть ξ — СВ, имеющая непрерывную функцию распределения $F_\xi(x)$.

Определение 5.2. Для любого $\alpha \in (0;1)$ число

$$x_\alpha = \min\{x : F_\xi(x) \geqslant \alpha\} \tag{5.2}$$

называется *квантилью уровня* α распределения $F_\xi(x)$.

Из (5.2) следует, что

$$\mathbf{P}(\xi \leqslant x_\alpha) = \alpha, \quad \mathbf{P}(\xi \geqslant x_\alpha) = 1 - \alpha. \tag{5.3}$$

Понятие квантили имеет существенное значение для построения доверительных интервалов и проверки статистических гипотез.

Центральный доверительный интервал. Пусть $G(Z_n;\theta)$ — такая СВ, что ее функция распределения $F_G(x) = \mathbf{P}\left(G(Z_n;\theta) \leqslant x\right)$ не зависит от θ. Пусть также для каждой реализации z_n выборки Z_n числовая функция $G_n(\theta) = G(z_n;\theta)$ непрерывна и строго монотонна по θ на Θ.

Определение 5.3. СВ $G(Z_n;\theta)$ называется *центральной статистикой* для θ.

Пусть задан уровень значимости p и выбраны произвольно $p_1 > 0$ и $p_2 > 0$ такие, что $p = p_1 + p_2$ (например, $p_1 = p_2 = \frac{p}{2}$). Если g_1 и g_2 — квантили распределения $F_G(x)$ уровней соответственно p_1 и $1 - p_2$, то для любого $\theta \in \Theta$

$$\mathbf{P}\left(g_1 \leqslant G(Z_n;\theta) \leqslant g_2\right) = 1 - p.$$

Найдем решения t_1 и t_2 уравнений $G(Z_n;\theta) = g_i$, $i = 1, 2$ и положим $T_1 = \min\{t_1, t_2\}$, $T_2 = \max\{t_1, t_2\}$. Тогда

$$\mathbf{P}\left(T_1 \leqslant \theta \leqslant T_2\right) = 1 - p = q,$$

т.е. $[T_1, T_2]$ — доверительный интервал для θ надежности q.

В силу произвола в выборе p_1 и p_2 интервал $[T_1, T_2]$ определен неоднозначно. Если при построении T_1 и T_2 с помощью $G(Z_n;\theta)$ дополнительно предположить, что $p_1 = p_2 = \frac{p}{2}$, то $[T_1, T_2]$ называют *центральным доверительным интервалом.*

В общем случае выбор p_1 и p_2 осуществляется так, чтобы длина интервала $T_2 - T_1$ была минимальной при неизменной надежности q (в этом случае интервальная оценка будет самой точной среди всех оценок надежности q).

Следующее утверждение дает общий способ построения центральной статистики.

Теорема 5.1. *Пусть выборка* $Z_n = \{X_k,\ k = 1, \ldots, n\}$ *соответствует функции распределения* $F(x;\theta)$, *удовлетворяющей следующим требованиям:*

1) $F(x;\theta)$ *непрерывна по* x *для любого* $\theta \in \Theta$;
2) $F(x;\theta)$ *непрерывна и монотонна по* θ *для любого* x.

Тогда $G(Z_n;\theta) = -\sum\limits_{k=1}^{n} \ln F(X_k;\theta)$ *является центральной статистикой для* $\theta \in \Theta$.

Асимптотический доверительный интервал. При больших объемах выборки ($n \gg 1$) для построения доверительного интервала можно воспользоваться любой асимптотически нормальной оценкой $\widehat{\theta}_n$ параметра θ. Пусть

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \xi \sim \mathcal{N}(0; d(\theta)), \quad n \to \infty, \tag{5.4}$$

где $d(\theta)$ — *асимптотическая дисперсия* оценки $\widehat{\theta}_n$.

Зададим надежность q, уровень значимости $p = 1 - q$ и определим (по табл. 22.2) квантиль u_α уровня $\alpha = 1 - \frac{p}{2}$ распределения $\mathcal{N}(0;1)$. Так как функция Лапласа строго монотонна, $\Phi(u_\alpha) = 1 - \frac{p}{2}$. Кроме того, если $\beta = \frac{p}{2}$, то $u_\beta = -u_\alpha$.

Если $d(\theta)$ непрерывна по $\theta \in \Theta$, то из (5.4) следует:

$$\mathbf{P}\left(\widehat{\theta}_n - u_\alpha\sqrt{\frac{d(\widehat{\theta}_n)}{n}} \leqslant \theta \leqslant \widehat{\theta}_n + u_\alpha\sqrt{\frac{d(\widehat{\theta}_n)}{n}}\right) \to \Phi(u_\alpha) - \Phi(-u_\alpha) = q.$$

Последнее означает, что интервал

$$\widehat{I} = \left[\widehat{\theta}_n - u_\alpha\sqrt{\frac{d(\widehat{\theta}_n)}{n}};\ \widehat{\theta}_n + u_\alpha\sqrt{\frac{d(\widehat{\theta}_n)}{n}}\right], \quad \alpha = 1 - \frac{p}{2} = \frac{1+q}{2}$$

при $n \gg 1$ накрывает оцениваемый параметр θ с вероятностью, близкой к $q = 1 - p$.

Если $\widehat{\theta}_n$ — МП-оценка параметра θ, то в условиях теоремы 3.2 $d(\theta) = \frac{1}{i(\theta)}$, где $i(\theta)$ — информация Фишера одного наблюдения. Пусть распределение, определяющее выборку, регулярно (см. определение 4.1), тогда $i(\theta) > 0$, $d(\theta) = \frac{1}{i(\theta)}$ непрерывна по θ, причем $\widetilde{d}(\theta) \geqslant d(\theta)$, если $\widetilde{d}(\theta)$ — асимптотическая дисперсия любой другой асимптотически нормальной оценки $\widetilde{\theta}_n$ параметра θ. Поэтому интервал $\widehat{I}$, построенный с использованием МП-оценки $\widehat{\theta}_n$, будет асимптотически наикратчайшим.

Специальные вероятностные распределения. Рассмотрим теперь некоторые специальные вероятностные распределения, необходимые для построения доверительных интервалов и проверки статистических гипотез.

Определение 5.4. Пусть $\{X_k,\ k = 1, \ldots, n\}$ — независимые СВ с распределением $\mathcal{N}(0;1)$. Тогда СВ

$$\chi_n^2 = \sum_{k=1}^{n} (X_k)^2$$

имеет χ^2-*распределение («хи-квадрат»-распределение) с n степенями свободы.*

Обозначение: $\chi_n^2 \sim \mathcal{H}_n$.

СВ χ_n^2 имеет плотность вероятности

$$p_{\chi_n^2}(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma\left(\dfrac{n}{2}\right)}\, x^{\frac{n-2}{2}} \exp\Big\{-\dfrac{x}{2}\Big\}, & x \geqslant 0, \\ 0, & x < 0, \end{cases}$$

где $\Gamma(\lambda) = \int\limits_0^\infty t^{\lambda-1}e^{-t}dt$ — гамма-функция.

Моментные характеристики: $\mathbf{M}\{\chi_n^2\} = n$, $\mathbf{D}\{\chi_n^2\} = 2n$.

Распределение $\mathcal{H}_n$ является асимптотически нормальным (по числу степеней свободы n): $\dfrac{\chi_n^2 - n}{\sqrt{2n}} \xrightarrow{d} \xi \sim \mathcal{N}(0;1)$, $n \to \infty$.

Определение 5.5. Пусть $X_k \sim \mathcal{N}(m_k;\sigma^2)$, $k = 1,\ldots,n$ — независимые СВ. Тогда СВ

$$\chi_{n,\delta}^2 = \frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k)^2$$

имеет *нецентральное распределение «хи-квадрат» с n степенями свободы и параметром нецентральности* $\delta = \dfrac{1}{\sigma^2} \sum\limits_{k=1}^{n} {m_k}^2$.

Обозначение: $\chi_{n,\delta}^2 \sim \mathcal{H}_{n,\delta}$.

Моментные характеристики: $\mathbf{M}\Big\{\chi_{n,\delta}^2\Big\} = n + \delta$, $\mathbf{D}\Big\{\chi_{n,\delta}^2\Big\} = 2(n + 2\delta)$.

Определение 5.6. Пусть $X \sim \mathcal{N}(0;1)$, $Y_n \sim \mathcal{H}_n$, X и Y — независимы. Тогда СВ

$$\tau_n = \frac{X}{\sqrt{\dfrac{1}{n}Y_n}}$$

имеет *распределение Стьюдента с n степенями свободы.*

Обозначение: $\tau_n \sim \mathcal{T}_n$.

СВ τ_n имеет плотность вероятности

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

Свойства распределения $\mathcal{T}_n$:

1) если $n > 2$, то $\mathbf{M}\{\tau_n\} = 0$, $\mathbf{D}\{\tau_n\} = \dfrac{n}{n-2}$;

2) если $n = 1$, то τ_n имеет *распределение Коши*: $p_{\tau_n}(x) = \dfrac{1}{\pi(1+x^2)}$;

3) асимптотическая нормальность: $\tau_n \xrightarrow{d} \xi \sim \mathcal{N}(0;1)$, $n \to \infty$.

Определение 5.7. Пусть СВ $X \sim \mathcal{H}_m$, $Y \sim \mathcal{H}_n$ независимы. Тогда СВ

$$f_{m,n} = \frac{\frac{1}{m}X}{\frac{1}{n}Y}$$

имеет *F-распределение Фишера с m и n степенями свободы.*

Обозначение: $f_{m,n} \sim F(m;n)$.

СВ $f_{m,n}$ имеет плотность вероятности

$$p_{f_{m,n}}(x) = \begin{cases} \dfrac{\Gamma\left(\frac{m+n}{2}\right) m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} x^{\frac{m}{2}-1}(n+mx)^{-\frac{m+n}{2}}, & x > 0, \\ 0, \quad x \leqslant 0. \end{cases}$$

Моментные характеристики: $\mathbf{M}\Big\{f_{m,n}\Big\} = \dfrac{n}{n-2}$, если $n > 2$; $\mathbf{D}\Big\{f_{m,n}\Big\} = \dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$, если $n > 4$.

Определение 5.8. Пусть СВ $X \sim \mathcal{H}_{m,\delta}$, $Y \sim \mathcal{H}_n$ независимы. Тогда СВ

$$f_{m,n,\delta} = \frac{\frac{1}{m}X}{\frac{1}{n}Y}$$

имеет *нецентральное F-распределение Фишера с m и n степенями свободы и параметром нецентральности* δ.

Обозначение: $f_{m,n,\delta} \sim F(m;n;\delta)$.

Пусть теперь $Z_n = \{X_k,\ k = 1,\ldots,n\}$ — выборка, соответствующая распределению $\mathcal{N}(\theta;\sigma^2)$, $\sigma > 0$, $\overline{X}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n} X_k$ — выборочное среднее, $\overline{S}_n^2 = \dfrac{1}{n}\sum\limits_{k=1}^{n}(X_k - \overline{X}_n)^2$ — выборочная дисперсия.

Теорема 5.2. *Статистики $\overline{X}_n$ и $\overline{S}_n^2$ независимы и обладают следующими свойствами:*

1) $\overline{X}_n \sim \mathcal{N}\left(\theta; \dfrac{\sigma^2}{n}\right)$;

2) $g_n = \dfrac{n\,\overline{S}_n^2}{\sigma^2} \sim \mathcal{H}_{n-1}$;

3) $\tau_{n-1} = \dfrac{\sqrt{n-1}(\overline{X}_n - \theta)}{\overline{S}_n} \sim \mathcal{T}_{n-1}$.

Утверждения теоремы 5.2 существенно облегчают построение доверительных интервалов для параметров гауссовского распределения.

5.2. Примеры

Пример 5.1. Выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$ соответствует распределению $\mathcal{N}(\theta; \sigma^2)$; $\sigma^2 > 0$ — известная дисперсия. Постройте для θ доверительный интервал надежности $q = 1 - p$.

Решение. Пусть $G(Z_n; \theta) = \dfrac{\sqrt{n}\,(\overline{X}_n - \theta)}{\sigma}$. По теореме 5.2 $G(Z_n; \theta) \sim \mathcal{N}(0; 1)$. При фиксированном $\overline{X}_n$ статистика $G(Z_n; \theta)$ монотонно убывает по θ. Следовательно, $G(Z_n; \theta)$ — центральная статистика. Пусть $p_1 + p_2 = p$, $p_1 > 0$, $p_2 > 0$. Найдем квантили g_1 и g_2 из соответствующих уравнений $\Phi(g_1) = p_1$ и $\Phi(g_2) = 1 - p_2$. Тогда $\mathbf{P}\left(g_1 \leqslant \dfrac{\sqrt{n}\,(\overline{X}_n - \theta)}{\sigma} \leqslant g_2\right) = q$. Отсюда

$$\mathbf{P}\left(\overline{X}_n - g_2\frac{\sigma}{\sqrt{n}} \leqslant \theta \leqslant \overline{X}_n - g_1\frac{\sigma}{\sqrt{n}}\right) = q. \tag{5.5}$$

Найдем g_1 и g_2 посредством минимизации длины полученного доверительного интервала: $\dfrac{\sigma}{\sqrt{n}}(g_2 - g_1) \to \min$ при условии $\Phi(g_2) - \Phi(g_1) = q$. Для этого рассмотрим функцию Лагранжа:

$$\mathcal{L}(g_1, g_2, \lambda) = \frac{\sigma}{\sqrt{n}}(g_2 - g_1) + \lambda\left(\Phi(g_2) - \Phi(g_1) - q\right), \ \ \lambda > 0.$$

Найдем стационарные точки функции $\mathcal{L}(g_1, g_2, \lambda)$:

$$\frac{\partial\mathcal{L}(g_1, g_2, \lambda)}{\partial g_1} = -\frac{\sigma}{\sqrt{n}} - \lambda p_G(g_1) = 0,$$

$$\frac{\partial\mathcal{L}(g_1, g_2, \lambda)}{\partial g_2} = \frac{\sigma}{\sqrt{n}} + \lambda p_G(g_2) = 0,$$

где $p_G(x)$ — плотность распределения $\mathcal{N}(0; 1)$. Отсюда следует, что $p_G(g_1) = p_G(g_2)$. Так как $p_G(x) = p_G(-x)$ для всех $x \in \mathbb{R}^1$, то либо $g_1 = g_2$, либо $g_1 = -g_2$. Первый случай не подходит, так как $\Phi(g_2) - \Phi(g_1) = 0 \neq q$. Отсюда заключаем, что $\Phi(g_2) - \Phi(-g_2) = q$. Таким образом, $g_2 = u_\alpha$ — квантиль уровня $\alpha = 1 - \frac{p}{2}$, а $g_1 = -u_\alpha$. Подставляя найденные g_1 и g_2 в (5.5), окончательно имеем

$$\mathbf{P}\left(\overline{X}_n - u_\alpha \frac{\sigma}{\sqrt{n}} \leqslant \theta \leqslant \overline{X}_n + u_\alpha \frac{\sigma}{\sqrt{n}}\right) = q. \qquad (5.6)$$

Заметим, что из $g_2 = -g_1 = u_\alpha$ следует, что $p_1 = p_2 = \frac{p}{2}$. Таким образом, доверительный интервал (5.6) является центральным. ■

Пример 5.2. Дана реализация z_n выборки Z_n объема $n = 9$, порожденной гауссовской СВ $X \sim \mathcal{N}(\theta; \sigma^2)$:

$$z_n = \{1{,}23; -1{,}384; -0{,}959; 0{,}731; 0{,}717; -1{,}805; -1{,}186; 0{,}658; -0{,}439\}.$$

Постройте для θ доверительные интервалы надежности $q = 0{,}95$, если а) $\sigma^2 = 1$; б) σ^2 неизвестна.

Решение. а) По условию $p = 1 - q = 0{,}05$, поэтому $\alpha = 1 - \frac{p}{2} = 0{,}975$. По табл. 22.2 находим: $u_\alpha = 1{,}96$. По реализации выборки z_n вычисляем реализацию $\overline{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k = -0{,}271$ выборочного среднего $\overline{X}_n$. Теперь из (5.6) следует, что искомый доверительный интервал $I_1 = \left[\overline{x}_n - u_\alpha \frac{\sigma}{\sqrt{n}}; \overline{x}_n + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$. Подставляя $\overline{x}_n$, $n = 9$, $\sigma = 1$ и $u_\alpha = 1{,}96$, находим, что $I_1 = [-0{,}924;\ 0{,}382]$.

б) Теперь дисперсия σ^2 неизвестна. Воспользуемся статистикой $G_n(Z_n; \theta) = \frac{\sqrt{n-1}\,(\overline{X}_n - \theta)}{\overline{S}_n}$, которая является центральной. Действительно, по теореме 5.2 $G_n(Z_n; \theta) \sim \mathcal{T}_{n-1}$, а монотонность по θ очевидна. Повторяя практически дословно рассуждения, приведенные в примере 5.1, находим доверительный интервал наименьшей длины:

$$\mathbf{P}\left(\overline{X}_n - t_\alpha(r)\frac{\overline{S}_n}{\sqrt{n-1}} \leqslant \theta \leqslant \overline{X}_n + t_\alpha(r)\frac{\overline{S}_n}{\sqrt{n-1}}\right) = q, \qquad (5.7)$$

где $t_\alpha(r)$ — квантиль уровня $\alpha = 0{,}975$ распределения Стьюдента $\mathcal{T}_r$ с $r = n-1 = 8$ степенями свободы. По табл. 22.4 находим, что $t_\alpha(8) = 2{,}306$.

По реализации z_n вычисляем реализацию $\overline{s}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x}_n)^2 = 1{,}115$ выборочной дисерсии $\overline{S}_n^2$. Теперь из (5.7) с учетом $n = 9$, найденного $t_\alpha(8)$ и того, что $\overline{s}_n = 1{,}056$, следует

$$I_2 = \left[\overline{x}_n - t_\alpha(r)\frac{\overline{s}_n}{\sqrt{n-1}}; \overline{x}_n + t_\alpha(r)\frac{\overline{s}_n}{\sqrt{n-1}}\right] = [-1{,}132;\ 0{,}59].$$

Итак, $I_2 = [-1{,}132;\ 0{,}59]$ — искомый доверительный интервал. Реализация выборка, приведенная в условии, в действительности соответствует распределению с параметрами $\theta_0 = 0$ и $\sigma_0^2 = 1$. Видим, что оба полученных интервала «накрывают» истинное значение θ_0 параметра θ.

Заметим, что I_1 и I_2, конечно, являются лишь *реализациями доверительных интервалов*, соответствующими конкретной реализации z_n выборки Z_n. ■

П р и м е р 5.3. В условиях примера 5.2 постройте доверительный интервал надежности $q = 0{,}95$ для неизвестной дисперсии σ^2.

Р е ш е н и е. Статистика $g_n(\sigma^2) = \dfrac{n\overline{S}_n^2}{\sigma^2}$ является центральной для σ^2, так как $g_n(\sigma^2) \sim \mathcal{H}_{n-1}$ и монотонно убывает по σ^2. Пусть $k_\alpha(n-1)$ и $k_\beta(n-1)$ — квантили χ^2-распределения $\mathcal{H}_{n-1}$ уровней соответственно $\alpha = \dfrac{p}{2}$ и $\beta = 1 - \dfrac{p}{2}$. Тогда $\mathbf{P}\left(k_\alpha(n-1) \leqslant \dfrac{n\overline{S}_n^2}{\sigma^2} \leqslant k_\beta(n-1)\right) = q$. Отсюда $\mathbf{P}\left(\dfrac{n\overline{S}_n^2}{k_\beta(n-1)} \leqslant \sigma^2 \leqslant \dfrac{n\overline{S}_n^2}{k_\alpha(n-1)}\right) = 0{,}95$, если $p = 0{,}05$.

Итак, искомый интервал для σ^2 имеет вид

$$I = \left[\frac{n\overline{S}_n^2}{k_\beta(n-1)};\ \frac{n\overline{S}_n^2}{k_\alpha(n-1)}\right].$$

Для $n = 9$, $\alpha = 0{,}025$, $\beta = 0{,}975$ по табл. 22.3 находим $k_\alpha(8) = 2{,}18$, $k_\beta(8) = 17{,}5$. Реализация интервала I с учетом данных примера 5.2 и того, что $\overline{s}_n^2 = 1{,}115$, имеет вид $\left[\dfrac{n\overline{s}_n^2}{k_\beta(8)};\ \dfrac{n\overline{s}_n^2}{k_\alpha(8)}\right] = \left[\dfrac{9\cdot 1{,}115}{17{,}5};\ \dfrac{9\cdot 1{,}115}{2{,}18}\right] = [0{,}58; 4{,}69]$.

Заметим, что истинное значение $\sigma_0^2 = 1$ накрывается найденным интервалом I. ■

П р и м е р 5.4. По данным примера 1.1, считая, что рост мужчины является СВ с гауссовским распределением $\mathcal{N}(m, \sigma^2)$, постройте реализации доверительных интервалов надежности $q = 0{,}95$ для математического ожидания m и дисперсии σ^2.

Р е ш е н и е. Для построения доверительных интервалов воспользуемся результатами примеров 5.2 и 5.3. Доверительный интервал для математического ожидания гауссовской СВ при неизвестной дисперсии имеет вид

$$I_1 = \left[\overline{X}_n - t_\alpha(r)\frac{\overline{S}_n}{\sqrt{n-1}} \leqslant m \leqslant \overline{X}_n + t_\alpha(r)\frac{\overline{S}_n}{\sqrt{n-1}}\right].$$

Из результатов примера 1.1 имеем: $n = 8585$, реализации выборочного среднего $\overline{x}_n = 67{,}46$, выборочной дисперсии $\overline{s}_n^2 = 6{,}6049$. Теперь с учетом $t_{0,975}(8584) = 1{,}96$ (по табл. 22.4) и того, что $\overline{s}_n = 2{,}57$, следует

$$I_1 = \left[67{,}46 - 1{,}96\frac{2{,}57}{\sqrt{8584}};\ 67{,}46 - 1{,}96\frac{2{,}57}{\sqrt{8584}}\right] = [67{,}406;\ 67{,}514].$$

Итак, $I_1 = [67{,}414;\ 67{,}514]$ — искомая реализация доверительного интервала для неизвестного математического ожидания.

Теперь построим реализацию интервала для σ^2, который согласно примеру 5.3 имеет вид

$$I_2 = \left[\frac{n\overline{S}_n^2}{k_\beta(n-1)};\ \frac{n\overline{S}_n^2}{k_\alpha(n-1)}\right].$$

Для $n = 8585$, $\alpha = 0{,}025$, $\beta = 0{,}975$ находим $k_\alpha(n-1) = 8329$, $k_\beta(n-1) = 8843$. Реализация интервала I_2 с учетом данных примера 1.1 имеет вид

$$I_2 = \left[\frac{8585 \cdot 2{,}57}{8843};\ \frac{8585 \cdot 2{,}57}{8329}\right] = [2{,}495; 2{,}649].$$

■

Пример 5.5. Выборка $\{X_k,\ k = 1, \ldots, n\}$, где $n \gg 1$, соответствует распределению $Bi(N;\theta)$, $\theta > 0$. Постройте асимптотический доверительный интервал для θ.

Решение. Известно (см. задачу 1 из раздела 4.3), что оценка $\widehat{\theta}_n = \dfrac{\overline{X}_n}{N}$ эффективна. Так как $\mathbf{M}\{X_k\} = N\theta$, то из центральной предельной теоремы следует: $\sqrt{n}\,(\overline{X}_n - N\theta) \xrightarrow{d} \xi \sim \mathcal{N}(0; N\theta(1-\theta))$, где $N\theta(1-\theta) = \mathbf{D}\{X_k\}$. Отсюда заключаем, что $\sqrt{n}\,(\widehat{\theta}_n - \theta) \xrightarrow{d} \eta \sim \mathcal{N}(0; d(\theta))$, где $d(\theta) = \dfrac{\theta(1-\theta)}{N}$ — асимптотическая дисперсия. Теперь, если u_α — квантиль уровня $\alpha = 1 - \dfrac{p}{2}$ распределения $\mathcal{N}(0;1)$, то искомый интервал имеет вид

$$\widehat{I}(n) = \left[\widehat{\theta}_n - u_\alpha\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n\,N}};\ \widehat{\theta}_n + u_\alpha\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n\,N}}\ \right],$$

где $\widehat{\theta}_n = \dfrac{\overline{X}_n}{N}$. При этом $\mathbf{P}\left(\theta \in \widehat{I}(n)\right) \to q$, $n \to \infty$. ■

5.3. Задачи для самостоятельного решения

1. Выборка $\{X_1, \ldots, X_n\}$, $n \gg 1$ соответствует распределению Пуассона $\Pi(\theta)$, $\theta > 0$. Постройте асимптотический доверительный интервал для θ надежности q.

Ответ: $\left[\overline{X}_n - u_\alpha\sqrt{\dfrac{\overline{X}_n}{n}};\ \overline{X}_n + u_\alpha\sqrt{\dfrac{\overline{X}_n}{n}}\right]$, $\alpha = \dfrac{1+q}{2}$.

2. Выборка $\{X_1, \ldots, X_n\}$ соответствует распределению $\mathcal{N}(\mu; \theta)$, μ — известно. Постройте центральный доверительный интервал надежности q для дисперсии θ.

Указание. Покажите, что $\dfrac{\sum\limits_{k=1}^{n} (X_k - \mu)^2}{\theta}$ — центральная статистика с распределением $\mathcal{H}_n$. Воспользуйтесь примером 5.3.

3. Выборка $\{X_1, \ldots, X_n\}$, $n \gg 1$ соответствует распределению $E\left(\dfrac{1}{\theta}\right)$, $\theta > 0$. Постройте асимптотический доверительный интервал надежности q для параметра θ.

Указание. Воспользуйтесь МП-оценкой $\widehat{\theta}_n$ для θ.

Ответ: $\left[\left(1 - \dfrac{u_\alpha}{\sqrt{n}}\right)\overline{X}_n;\ \left(1 + \dfrac{u_\alpha}{\sqrt{n}}\right)\overline{X}_n\right]$; $\alpha = \dfrac{1+q}{2}$.

4. По выборке $z_n = \{-0{,}26; -0{,}36; 1{,}83; 0{,}54; -2{,}06\}$, соответствующей распределению $\mathcal{N}(\theta_1; \theta_2^2)$, найдите доверительные интервалы надежности $q = 0{,}9$ для θ_1 и θ_2.

Ответ: $[-1{,}41; 1{,}29]$ — для θ_1; $[0{,}94; 3{,}43]$ — для θ_2.

5. Выборка $Z_n = \{X_k,\ k = 1, \ldots, n\}$, $n \gg 1$ соответствует равномерному распределению $R[0; a]$, $a > 0$. Постройте асимптотический доверительный интервал надежности q для параметра $\theta = \mathbf{M}\{X_1\}$.

Указание. Используйте оценку $\widehat{\theta}_n = \overline{X}_n$.

Ответ: $\left[\left(1 - \dfrac{u_\alpha}{\sqrt{3n}}\right)\overline{X}_n;\ \left(1 + \dfrac{u_\alpha}{\sqrt{3n}}\right)\overline{X}_n\right]$; $\alpha = \dfrac{1+q}{2}$.

6. Выборка $Z_n = \{X_1, \ldots, X_n\}$, $n \gg 1$ соответствует распределению с плотностью вероятности $p(x; \theta) = \exp\{\theta - x\}$, $x \geqslant \theta$, где $\theta > 0$. Покажите, что доверительным интервалом надежности q для θ является $\left[X_{(1)} + \dfrac{\ln(1-q)}{n};\ X_{(1)}\right]$.

Указание. Покажите, что $G(Z_n; \theta) = n(X_{(1)} - \theta)$ — центральная статистика.

7. В условиях задачи 6 постройте центральный доверительный интервал надежности q для θ.

Ответ: $\left[X_{(1)} + \dfrac{1}{n}\ln\left(\dfrac{1-q}{2}\right);\ X_{(1)} + \dfrac{1}{n}\ln\left(\dfrac{1+q}{2}\right)\right]$.

8. Пусть выборка $\{X_1, \ldots, X_n\}$, $n \gg 1$ соответствует распределению $R[0; \theta]$, $\theta > 0$. Покажите, что $\left(\dfrac{X_{(n)}}{\theta}\right)^n$ — центральная статистика, и постройте для θ доверительный интервал минимальной длины и надежности q.

Ответ: $\left[X_{(n)}; \dfrac{X_{(n)}}{(1-q)^{1/n}}\right]$.

9. Пусть $Z_n^{(1)} = \{X_1, \ldots, X_n\}$ и $Z_n^{(2)} = \{Y_1, \ldots, Y_n\}$ — две независимые выборки, причем $Z_n^{(1)}$ соответствует распределению $\mathcal{N}(\theta_1; \sigma_1^2)$, а $Z_n^{(2)}$ — $\mathcal{N}(\theta_2; \sigma_2^2)$ (σ_1 и σ_2 — известны). Требуется построить доверительный интервал надежности q для параметра $\theta = \theta_1 - \theta_2$.

Указание. Используйте $G(Z_n^{(1)}; Z_n^{(2)}; \theta) = \dfrac{\overline{X}_n - \overline{Y}_n - \theta}{\sigma}$, где $\sigma^2 = \dfrac{\sigma_1^2 + \sigma_2^2}{n}$.

Ответ: $\left[\overline{X}_n - \overline{Y}_n - u_\alpha\sigma;\ \overline{X}_n - \overline{Y}_n + u_\alpha\sigma\right]$, $\alpha = \dfrac{1+q}{2}$.

6. Проверка параметрических гипотез

В практических задачах часто требуется не только оценить значение неизвестного параметра, но и проверить некоторое предположение относительно этого параметра. Например, пройдет ли партия А на ближайших выборах в парламент, если для этого требуется получить поддержку не менее семи процентов избирателей? Пусть для разрешения этого вопроса социологи опросили 1000 респондентов, и 68 из них высказались в поддержку партии А. Точечная оценка уровня поддержки партии А оказалась чуть меньше требуемых 7 процентов. Можно ли приписать это отклонение статистической изменчивости, связанной со случайным выбором респондентов, или наблюдаемое отличие следует считать значимым, и гипотезу о том, что партия А наберет 7 процентов голосов следует отвергнуть? Какие отклонения от уровня 7 процентов допустимы, чтобы предположение о прохождении партии в парламент считать верным?

Математическая формализация и алгоритм проверки параметрических гипотез будут рассмотрены в этом параграфе.

6.1. Теоретические положения

Пусть СВ X имеет закон распределения, заданный функцией распределения $F(x; \theta)$ или плотностью вероятности $p(x; \theta)$, где θ — некоторый скалярный или векторный параметр.

Определение 6.1. *Статистической гипотезой* называется любое априорное предположение о законе распределения СВ.

Определение 6.2. Любое предположение о возможных значениях параметра θ называется *параметрической гипотезой.*

Определение 6.3. Параметрическая гипотеза, состоящая в том, что $\theta = \theta_0$, где θ_0 — фиксированная величина, называется *простой гипотезой.*

Определение 6.4. Параметрическая гипотеза называется *сложной*, если она состоит в том, что $\theta \in \Theta_0$, где Θ_0 — некоторое фиксированное подмножество, принадлежащее множеству Θ возможных значений параметра θ и содержащее более одной точки.

Статистическая гипотеза, подлежащая проверке, называется *основной* (или *нулевой*) и обозначается H_0. Гипотеза, которая конкурирует с H_0, называется *альтернативой* по отношению к H_0 и обозначается H_1 или H_A. Для сложных параметрических гипотез основной гипотезой является $H_0 : \theta \in \Theta_0$, а альтернативной $H_1 : \theta \in \in \Theta_1$, где $\Theta_1 \in \Theta \setminus \Theta_0$.

Определение 6.5. *Статистическим критерием* называется алгоритм проверки гипотезы H_0 по выборке Z_n.

Определение 6.6. Будем называть *статистикой критерия* некоторую числовую функцию $T(Z_n)$ выборки Z_n, обладающую тем свойством, что ее закон распределения полностью известен, если H_0 верна.

Рассмотрим общую структуру статистического критерия. Пусть V_0 — множество всех возможных значений вектора Z_n в предположении, что H_0 — верна. Выберем малое положительное число $p \in (0; 1)$ и область $S_p \in V_0$ такую, что

$$P_0(S_p) = \mathbf{P}\left(Z_n \in S_p \mid H_0 \text{— верна}\right) = p.$$

Определение 6.7. Число p называется *уровнем значимости* (*размером*) критерия, а множество S_p — *критической областью уровня* p.

Пусть $z_n = [x_1, \dots, x_n]^\top$ — конкретная реализация выборки Z_n. Предположим, что $z_n \in S_p$, тогда гипотеза H_0 *отвергается на уровне значимости* p. Если же $z_n \in \overline{S}_p = V_0 \setminus S_p$, то H_0 — принимается. Область $\overline{S}_p$ называют *доверительной областью*. Очевидно, что $P_0(\overline{S}_p) = 1 - p = q$. Вероятность q называют *уровнем доверия* или *надежностью* критерия.

Определение 6.8. Факт отклонения гипотезы H_0 в случае, когда она верна, называется *ошибкой первого рода*. Принятие гипотезы H_0 при условии, что в действительности верна альтернатива H_1, называется *ошибкой второго рода*.

Поясним смысл ошибок первого и второго рода. Типичным примером является вынесение судебного решения. Если за нулевую гипотезу принять то, что подсудимый невиновен, то ошибка первого рода происходит, когда суд признает его виновным. Ошибка второго рода имеет место в том случае, когда суд ошибочно оправдывает виновного подсудимого.

Очевидно, что вероятность ошибки первого рода равна $P_0(S_p) = p$, т.е. совпадает с уровнем значимости критерия.

Вероятность ошибки второго рода имеет вид

$$\beta = P_1(\overline{S}_p) = \mathbf{P}\left(Z_n \in \overline{S}_p \mid H_1 \text{— верна}\right).$$

Определение 6.9. Пусть $H_0 : \theta = \theta_0$, а альтернатива H_1: $\theta = \gamma$, где $\gamma \neq \theta_0$. Тогда функция

$$W(S_p, \gamma) = \mathbf{P}\{Z_n \in S_p \mid H_1 \text{— верна}\}$$

называется *мощностью критерия* при альтернативе H_1.

Понятно, что критерий будет «хорошо» различать H_0 и H_1, если p близко к нулю, а S_p выбрана так, что $W(\Delta_p, \gamma)$ близка к единице.

Определение 6.10. Статистический критерий называется *состоятельным* против альтернативы H_1: $\theta \in \Theta_1$, если при $p > 0$ и $n \to \infty$ мощность $W(S_p, \gamma)$ стремится к единице для любого $\gamma \in \Theta_1$.

Если альтернатива H_1: $\theta = \theta_1$ — простая, то вероятность ошибки второго рода β связана с мощностью критерия очевидным соотношением $\beta = 1 - W(S_p, \theta_1)$ и состоятельность критерия означает, что β стремится к нулю.

Определение 6.11. Пусть уровень значимости критерия равен $p > 0$. *Наиболее мощным критерием* для проверки простой гипотезы H_0: $\theta = \theta_0$ против H_1: $\theta = \theta_1$ называется критерий с такой критической областью S_p^*, что

$$W(S_p^*, \theta_1) = \max_{S_p \in I_p} W(\Delta_p, \theta_1), \tag{6.1}$$

где I_p — множество всех критических областей уровня p.

В некоторых случаях область S_p^* существует и может быть найдена аналитически (см. далее теорему 6.1).

Обычно на практике критическую область S_p задают неявно с помощью некоторой *статистики критерия* $T(Z_n)$. Пусть Δ_p — область на $\mathbb{R}^1$ такая, что

$$\mathbf{P}\Big(T(Z_n) \in \Delta_p \mid H_0 \text{— верна}\Big) = p.$$

Тогда критическая область S_p определяется так:

$$S_p = \{z : z \in V_0 \text{ и } T(z) \in \Delta_p\}. \tag{6.2}$$

Как правило, описать явно одномерную область Δ_p существенно проще, чем n-мерную область S_p. Например, $T(z)$ и Δ_p достаточно просто определяются с помощью метода доверительных интервалов (см. пример 6.1).

Пусть α и β — вероятности ошибок первого и второго рода соответственно. Тогда

$$\alpha = \mathbf{P}\{T(Z_n) \in \Delta_p \mid H_0 \text{— верна}\} = p,$$

$$\beta = \mathbf{P}\{T(Z_n) \in \overline{\Delta}_p \mid H_1 \text{— верна}\}.$$

Таким образом, Δ_p имеет смысл совокупности маловероятных значений статистики $T(Z_n)$ в случае, когда гипотеза H_0 верна. При этом вероятность попадания статистики $T(Z_n)$ в доверительную область $\overline{\Delta}_p$ близка к единице.

Далее мы будем говорить, что H_0 отвергается на уровне значимости p всякий раз, когда $T(Z_n) \in \Delta_p$. Заметим, что отклонение H_0 означает, что основная гипотеза плохо согласуется с имеющимися экспериментальными данными Z_n. При этом мы, естественно, не можем в общем случае утверждать, что отвергнутая гипотеза H_0 неверна с вероятностью единица.

Рассмотрим общий способ выбора статистики $T(Z_n)$, приводящий к наиболее мощному критерию для проверки простой гипотезы H_0: $\theta = \theta_0$ против простой альтернативы H_1: $\theta = \theta_1$.

Пусть $Z_n = \{X_k, k = 1, \ldots, n\}$ — выборка, соответствующая распределению с плотностью $p(x; \theta) > 0$, где $\theta = \theta_j$, $j = 0, 1$, $\theta_0 \neq \theta_1$, а $z_n = [x_1, \ldots, x_n]^T$ — реализация Z_n. Введем *статистику отношения правдоподобия*:

$$T(Z_n) = \frac{\prod\limits_{k=1}^{n} p(X_k; \theta_1)}{\prod\limits_{k=1}^{n} p(X_k; \theta_0)}. \tag{6.3}$$

Теорема 6.1 (Нейман—Пирсон). *Наиболее мощный критерий для проверки H_0 на уровне значимости p против H_1 существует и задается оптимальной в смысле* (6.1) *критической областью $S_p^* = \{z_n : T(z_n) \geqslant \delta\}$, где параметр δ определяется из условия*

$$\mathbf{P}\left(T(Z_n) \geqslant \delta \mid H_0 \text{— верна}\right) = p,$$

в котором $T(Z_n)$ задается формулой (6.3).

Заметим, что в условиях теоремы 6.1 критическая область Δ_p для $T(Z_n)$ имеет простой вид: $\Delta_p = [\delta, +\infty)$.

Замечание. Аналогичный результат можно получить и для выборки, соответствующей дискретному распределению. Однако в силу дискретности распределения выборки не всегда можно выбрать параметр δ так, чтобы уровень значимости критерия равнялся p (подробнее смотри раздел 4.2 в [16]).

Алгоритм проверки статистической гипотезы:

1) сформулировать основную гипотезу H_0 и альтернативу H_1;

2) выбрать уровень значимости критерия p;

3) выбрать статистику $T(Z_n)$ и найти ее закон распределения в предположении, что H_0 верна;

4) построить критическую Δ_p и доверительную $\overline{\Delta}_p$ области;

5) если H_1 — простая гипотеза, то вычислить мощность критерия и убедиться в том, что выбранная область Δ_p обеспечивает приемлемую вероятность β ошибки второго рода. Если H_1 не является простой, то перейти сразу к п. 6);

6) по реализации $z_n = \{x_1, \ldots, x_n\}^T$ выборки Z_n вычислить реализацию $T(z_n)$ статистики критерия $T(Z_n)$;

7) принять решение о справедливости (не справедливости) гипотезы H_0:

— если $T(z_n) \in \Delta_p$, то H_0 отвергается на уровне значимости p;

— если $T(z_n) \in \overline{\Delta}_p$, то H_0 принимается на уровне значимости p.

Аналогично определению 21.24 введем понятие асимптотической нормальности статистики критерия. Пусть $T(Z_n)$ — статистика некоторого критерия, предназначенного для проверки гипотезы H_0.

Определение 6.12. Будем называть статистику $T(Z_n)$ *асимптотически нормальной*, если

$$T(Z_n) \xrightarrow{d} X, \qquad \text{при } n \to \infty,$$

где $X \sim \mathcal{N}(0;1)$.

6.2. Примеры

Пример 6.1. По выборке $Z_n = \{X_k, k = 1, \ldots, n\}$, соответствующей распределению $\mathcal{N}(\theta;\sigma^2)$, где $\sigma^2 > 0$ — известна, проверьте гипотезу H_0: $\theta = \theta_0$ на уровне значимости p против альтернативы H_1: $\theta \neq \theta_0$.

Решение. Для проверки H_0 воспользуемся методом доверительных интервалов. Рассмотрим статистику $T(Z_n) = \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Из примера 5.1 следует, что при $\theta = \theta_0$ (т.е. H_0 — верна) и $\alpha = 1 - \frac{p}{2}$

$$\mathbf{P}\left(\overline{X}_n - u_\alpha \frac{\sigma}{\sqrt{n}} \leqslant \theta_0 \leqslant \overline{X}_n + u_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - p,$$

если u_α — квантиль уровня α распределения $\mathcal{N}(0;1)$. Отсюда

$$\mathbf{P}\left(\overline{X}_n \in \left[\theta_0 - u_\alpha \frac{\sigma}{\sqrt{n}}, \theta_0 + u_\alpha \frac{\sigma}{\sqrt{n}}\right]\right) = 1 - p.$$

Таким образом, критическая область Δ_p для статистики критерия $T(Z_n) = \overline{X}_n$ принимает вид

$$\Delta_p = \left\{x : |x - \theta_0| > u_\alpha \frac{\sigma}{\sqrt{n}}\right\},$$

а доверительная область $\overline{\Delta}_p = \mathbb{R}^1 \setminus \Delta_p = \left\{x : |x - \theta_0| \leqslant u_\alpha \frac{\sigma}{\sqrt{n}}\right\}$.

Итак, если $z_n = \{x_1, \ldots, x_n\}^\top$ — реализация выборки Z_n, а $\overline{x}_n = T(z_n) = \frac{1}{n}\sum_{k=1}^{n} x_k$ — соответствующая реализация выборочного среднего $\overline{X}_n$ (т.е. статистики критерия), то гипотезу H_0 на уровне значимости p следует отвергнуть, если $\overline{x}_n \in \Delta_p$, т.е. $|\overline{x}_n - \theta_0| > u_\alpha \frac{\sigma}{\sqrt{n}}$. Если же $\overline{x}_n \in \overline{\Delta}_p$, то H_0 следует принять. ■

Пример 6.2. В условиях примера 6.1 вычислите мощность критерия и вероятность ошибки второго рода, если H_1: $\theta = \gamma$, $\gamma \neq \theta_0$.

Решение. По определению 6.9 с учетом имеем

$$\begin{aligned} W(S_p, \gamma) &= \mathbf{P}\left(\overline{X}_n \in \Delta_p \mid H_1\text{— верна}\right) = \\ &= \mathbf{P}\left(\frac{\sqrt{n}|\overline{X}_n - \theta_0|}{\sigma} > u_\alpha \mid H_1\text{— верна}\right) = \\ &= 1 - \mathbf{P}\left(\theta_0 - u_\alpha \frac{\sigma}{\sqrt{n}} \leqslant \overline{X}_n \leqslant \theta_0 + u_\alpha \frac{\sigma}{\sqrt{n}} \mid H_1\text{— верна}\right). \end{aligned}$$

Если верна альтернатива H_1, то $\overline{X}_n \sim \mathcal{N}\left(\gamma; \frac{\sigma^2}{n}\right)$, поэтому

$$\begin{aligned} W(S_p, \gamma) = 1 - &\left[\Phi\left(\frac{\theta_0 + u_\alpha \frac{\sigma}{\sqrt{n}} - \gamma}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{\theta_0 - u_\alpha \frac{\sigma}{\sqrt{n}} - \gamma}{\frac{\sigma}{\sqrt{n}}}\right)\right] = \\ = 1 - &\left[\Phi\left(\frac{\sqrt{n}(\theta_0 - \gamma)}{\sigma} + u_\alpha\right) - \Phi\left(\frac{\sqrt{n}(\theta_0 - \gamma)}{\sigma} - u_\alpha\right)\right] = 1 - \varphi_n(\theta_0, \gamma). \end{aligned}$$

Очевидно, что $W(S_p, \theta_0) = 1 - [\Phi(u_\alpha) - \Phi(-u_\alpha)] = 1 - (1 - p) = p$ — вероятность ошибки первого рода.

По определению 6.9 вероятность ошибки второго рода $\beta = 1 - W(S_p, \gamma) = \varphi_n(\theta_0, \gamma)$.

Сделаем некоторые выводы о зависимости $\varphi_n(\theta_0, \gamma)$ от величины γ и объема выборки n (θ_0 — фиксировано).

1. Если $n = \text{const}$, а $|\theta_0 - \gamma| \to \infty$, то $\varphi_n(\theta_0, \gamma) \to 0$. Поэтому $W(S_p, \theta_0) \to 1$, а $\beta \to 0$. Последнее означает, что при фиксированном объеме выборки n хорошо различаются «далекие» гипотезы H_0 и H_1 (т.е. $|\theta_0 - \gamma| \gg 0$). Если же $\theta_0 \approx \gamma$, то $\beta \approx 1 - W(S_p, \theta_0) = 1 - p$, т.е. близка к единице, так как p мало по условию.

2. Если же $\theta_0 \neq \gamma$, но $n \to \infty$, то $\frac{\sqrt{n}|\theta_0 - \gamma|}{\sigma} \to \infty$. Поэтому $\varphi_n(\theta_0, \gamma) \to 0$ при $n \to \infty$, θ_0, γ — фиксированы. Последнее означает,

что критерий будет хорошо различать даже «близкие» гипотезы ($\theta_0 \approx \approx \gamma$), если объем выборки n достаточно велик. Следовательно, критерий является состоятельным против любой простой альтернативы H_1. ■

Пример 6.3. По реализации z_n выборки Z_n объема $n = 100$, соответствующей распределению $\mathcal{N}(\theta; 1)$, вычислена реализация выборочного среднего $\overline{x}_n = 0{,}153$. На уровне значимости $p = 0{,}05$ проверьте гипотезу H_0: $\theta = 0$ против альтернативы H_1: $\theta = 0{,}5$. Вычислить мощность критерия и вероятность ошибки второго рода β.

Решение. Воспользуемся результатами примеров 6.1 и 6.2. По условию $n = 100$, $\sigma = 1$, $p = 0{,}05$, $\alpha = 1 - \frac{p}{2} = 0{,}975$, $u_\alpha = 1{,}96$. Доверительная область $\overline{\Delta}_p$ имеет вид

$$\overline{\Delta}_p = \left[\theta_0 - u_\alpha \frac{\sigma}{\sqrt{n}}; \theta_0 + u_\alpha \frac{\sigma}{\sqrt{n}}\right] = [-0{,}196; 0{,}196],$$

где учтено, что $\theta_0 = 0$ по условию. Так как $\overline{x}_n = 0{,}153 \in \overline{\Delta}_p$, гипотеза H_0 принимается. Заметим, что, проводя аналогичные выкладки для гипотезы H_1, мы получили бы доверительную область $\overline{\Delta}_p^{(1)} = [-0{,}196 + 0{,}5; 0{,}196 + 0{,}5] = [0{,}304; 0{,}696]$. Так как $\overline{x}_n \notin \overline{\Delta}_p^{(1)}$, то гипотезу H_1 следует отвергнуть.

Из примера 6.2 следует, что при $\theta_0 = 0$ и $\gamma = 0{,}5$

$$W(S_p, \gamma) = 1 - [\Phi(5 + 1{,}96) - \Phi(5 - 1{,}96)] \approx \Phi(3{,}04) = 0{,}9987.$$

Поэтому вероятность ошибки второго рода весьма мала: $\beta = = 1 - W(S_p, \gamma) = 0{,}0013$. ■

Пример 6.4. В условиях примера 6.1 постройте наиболее мощный критерий для проверки гипотезы H_0: $\theta = \theta_0$ против альтернативы H_1: $\theta = \gamma$, $\gamma > \theta_0$.

Решение. Воспользуемся теоремой 6.1 Неймана—Пирсона. Статистика критерия (6.3) с учетом гауссовости выборки принимает вид

$$T(Z_n) = \frac{\left(\sqrt{2\pi}\sigma\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(X_k - \gamma)^2\right\}}{\left(\sqrt{2\pi}\sigma\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(X_k - \theta_0)^2\right\}} =$$

$$= \exp\left\{\frac{\sum_{k=1}^{n} X_k(\gamma - \theta_0)}{\sigma^2} - \frac{n}{2\sigma^2}\left(\gamma^2 - \theta_0^2\right)\right\}.$$

Поэтому неравенство $T(Z_n) \geqslant \delta$ эквивалентно $\ln(T(Z_n)) \geqslant \ln\delta$, т.е. $\overline{X}_n \geqslant \delta_1$, где $\overline{X}_n = \frac{1}{n}\sum\limits_{k=1}^{n} X_k$, а $\delta_1 = \frac{1}{2}(\theta_0+\gamma) + \frac{\sigma^2 \ln\delta}{(\gamma-\theta_0)n}$.

Найдем теперь δ_1 с учетом того, что $\overline{X}_n \sim \mathcal{N}\left(\theta_0; \frac{\sigma^2}{n}\right)$, если H_0 — верна. Из теоремы 6.1 следует:

$$p = \mathbf{P}\left(T(Z_n) \geqslant \delta \mid H_0 \text{— верна}\right) = \mathbf{P}\left(\overline{X}_n \geqslant \delta_1 \mid H_0 \text{— верна}\right) = \\ = 1 - \Phi\left(\frac{\sqrt{n}(\delta_1-\theta_0)}{\sigma}\right).$$

Отсюда следует, что $\Phi\left(\frac{\sqrt{n}(\delta_1-\theta_0)}{\sigma}\right) = 1-p$, т.е. $\frac{\sqrt{n}(\delta_1-\theta_0)}{\sigma} = u_\alpha$, где u_α — квантиль уровня $\alpha = 1-p$ распределения $\mathcal{N}(0;1)$. Таким образом, $\delta_1 = \theta_0 + u_\alpha \frac{\sigma}{\sqrt{n}}$.

Итак, если реализация выборочного среднего $\overline{X}_n$ удовлетворяет неравенству $\overline{x}_n \geqslant \theta_0 + u_\alpha \frac{\sigma}{\sqrt{n}}$, то гипотезу H_0 следует отвергнуть. В заключение заметим, что граница δ_1 зависит от θ_0, но не зависит от конкретного значения γ (учтено лишь, что $\gamma > \theta_0$). ■

Пример 6.5. Опрошено 1000 респондентов, из них 68 высказались в поддержку партии А. Пусть θ — вероятность того, что случайным образом выбранный человек проголосует за партию А. Проверьте гипотезу H_0: $\theta = 0{,}07$ на уровне значимости $p = 0{,}05$ против альтернативы H_1: $\theta \neq 0{,}07$.

Решение. Рассмотрим СВ X, которая принимает значение 1, если человек голосует за партию А, и 0, если не голосует. Тогда выборка $\{X_k,\ k = 1, \ldots, n\}$, где $n = 1000$, соответствует распределению $Bi(1;\theta)$, $\theta > 0$. Для проверки H_0 воспользуемся методом доверительных интервалов. Рассмотрим статистику $T(Z_n) = \overline{X}_n = \frac{1}{n}\sum\limits_{k=1}^{n} X_k$. Из примера 5.5 следует, что при $n \gg 1$, $N = 1$, $\theta = \theta_0$ (т.е. H_0 — верна) и $\alpha = 1 - \frac{p}{2}$

$$\mathbf{P}\left(\overline{X}_n - u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leqslant \theta_0 \leqslant \overline{X}_n + u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right) = 1-p,$$

где u_α — квантиль уровня α распределения $\mathcal{N}(0;1)$. Отсюда

$$\mathbf{P}\left(\overline{X}_n \in \left[\theta_0 - u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \theta_0 + u_\alpha\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right]\right) = 1-p.$$

Таким образом, критическая область Δ_p для статистики критерия $T(Z_n) = \overline{X}_n$ принимает вид

$$\Delta_p = \left\{ x : |x - \theta_0| > u_\alpha \sqrt{\frac{\overline{X}_n(1 - \overline{X}_n)}{n}} \right\}.$$

Итак, если $\overline{x}_n = T(z_n) = \dfrac{1}{n}\sum\limits_{k=1}^{n} x_k = \dfrac{68}{1000}$ — соответствующая реализация выборочного среднего $\overline{X}_n$ (т.е. статистики критерия), то гипотезу H_0 на уровне значимости $p = 0{,}05$ следует отвергнуть, если $|\overline{x}_n - \theta_0| > u_\alpha \sqrt{\dfrac{\overline{x}_n(1 - \overline{x}_n)}{n}}$. Подставляя в это выражение $\overline{x}_n = 0{,}0068$ и $u_{0,975} = 1{,}96$, получаем $|0{,}068 - 0{,}007| \leqslant 0{,}0156$. Следовательно, гипотеза H_0 принимается на уровне значимости 0,05, т.е. можно считать, что 7 процентов избирателей будет голосовать за партию A. ■

6.3. Задачи для самостоятельного решения

1. Обобщите результат примера 6.1 на случай, когда дисперсия σ^2 неизвестна.

Указание. См. примеры 5.2 (б) и 6.1.

Ответ: H_0 принимается, если $|\overline{X}_n - \theta_0| \leqslant t_\alpha \sqrt{\dfrac{\overline{S}^2_n}{n-1}}$, где t_α — квантиль уровня $\alpha = 1 - \dfrac{p}{2}$ распределения Стьюдента с $n - 1$ степенью свободы, $\overline{S}^2_n$ — выборочная дисперсия.

2. Выборка Z_n соответствует распределению $\mathcal{N}(\theta_1; \theta_2)$. Проверьте на уровне значимости p гипотезу H_0: $\theta_2 = \sigma^2$ против $H_1 : \theta_2 \neq \sigma^2$.

Указание. См. пример 5.3.

Ответ: H_0 принимается, если $\overline{S}^2_n \in \left[k_1 \dfrac{\sigma^2}{n}; k_2 \dfrac{\sigma^2}{n} \right]$, где k_1 и k_2 — квантили распределения $\mathcal{H}_{n-1}$ уровней $\dfrac{p}{2}$ и $1 - \dfrac{p}{2}$ соответственно.

3. Пусть $\{X_k, k = 1, \ldots, n\}$ и $\{Y_m, m = 1, \ldots, n\}$ — независимые выборки, порожденные СВ $X \sim \mathcal{N}(\theta_1; \sigma_1^2)$ и $Y \sim \mathcal{N}(\theta_2; \sigma_2^2)$, σ_1 и σ_2 — известны. На уровне значимости p проверьте гипотезу $H_0 : \theta_1 = \theta_2$ против $H_1 : \theta_1 \neq \theta_2$.

Указание. См. задачу 9 из раздела 5.3.

Ответ: H_0 принимается, если $|\overline{X}_n - \overline{Y}_n| \leqslant u_\alpha \sqrt{\dfrac{\sigma_1^2 + \sigma_2^2}{n}}$, u_α — квантиль распределения $\mathcal{N}(0; 1)$ уровня $\alpha = 1 - \dfrac{p}{2}$.

4. В условиях примера 6.4 найдите вероятность β ошибки второго рода.

Ответ: $\beta = \Phi\left(\sqrt{n} \dfrac{(\theta_0 - \gamma)}{\sigma} + u_\alpha \right)$.

5. В условиях предыдущей задачи определите, при каком минимальном объеме выборки n_0 величина β будет не больше 0,01, если $\theta_0 = 0$, $\gamma = 1$, $\sigma^2 = 1$, $p = 0{,}05$.

О т в е т: $n_0 = 9$.

6. В условиях примера 6.4 найдите минимальный объем выборки n_0, при котором вероятности ошибок первого и второго рода не больше заданных значений соответственно $a > 0$ и $b > 0$.

О т в е т: $n_0 = \left[\dfrac{\sigma^2(u_a + u_b)}{(\theta_0 - \gamma)^2}\right] + 1$, где $[\cdot]$ — целая часть числа, u_a и u_b — квантили распределения $\mathcal{N}(0;1)$ уровней a и b соответственно.

7. Выборка $Z_n = \{X_k, k = 1, \ldots, n\}$ соответствует распределению $\mathcal{N}(0;\theta^2)$, $\theta > 0$. Постройте наиболее мощный критерий для проверки на уровне значимости p гипотезы $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta = \sigma > \theta_0$.

О т в е т: H_0 отвергается, если $\sum\limits_{k=1}^{n} X_k^2 \geqslant k_\alpha \theta_0^2$, где $k_\alpha(n-1)$ — квантиль уровня $\alpha = 1 - p$ распределения $\mathcal{H}_{n-1}$.

ГЛАВА II

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Ознакомившись с первичной обработкой данных, перейдем теперь к более сложным статистическим моделям, в которых фигурирует не одна, а две и большее число выборок. В разделе 7 будет рассмотрена задача проверки гипотезы об одинаковой распределенности (однородности) двух выборок. Справедливость основной гипотезы в такой задаче будет означать, что две выборки можно объединить и рассматривать как единую выборку из одной совокупности. Отметим, что при формализации практической задачи исключительное внимание следует обратить на выбор альтернативной гипотезы, так как неправильная формулировка альтернативы может привести к неверному выводу. Выбор альтернативы определяется содержанием конкретной практической задачи и целью исследования. Вопрос о том, как правильно выбрать критерий для проверки однородности и сформулировать основную и альтернативную гипотезы будет подробно обсуждаться в примерах 7.6—7.8. Раздел 8 посвящён однофакторному дисперсионному анализу. Задача однофакторного дисперсионного анализа является обобщением задачи проверки однородности двух выборок, которые могут различаться сдвигом, на случай произвольного конечного числа K $(2 < K < \infty)$ выборок.

В разделе 9 изучается проблема выявления зависимости (независимости) двух (или нескольких) показателей, измеренных в количественной, порядковой или номинальной шкале. Если зависимость между показателями обнаружена, то силу связи можно оценить. Для каждой шкалы будут описаны коэффициенты, характеризующие силу связи изучаемых признаков. Для решения указанных задач будут рассмотрены не только традиционные классические методы, но и интенсивно развивающиеся последние 50–60 лет непараметрические методы. Непараметрические методы не предназначены для какого-либо специального параметрического семейства распределений, а могут применяться к широкому классу распределений (как правило, это класс непрерывных распределений).

В разделах 7 и 8 будет показано, что рассмотренные непараметрические процедуры лишь немного менее эффективны, чем их конкуренты из классической (нормальной) теории, если наблюдения имеют нормальное распределение, зато могут оказаться значительно эффективнее классических процедур, если распределение наблюдений отлично от нормального.

Применение всех описанных методов будет проиллюстрировано экономическими, медицинскими, социологическими и техническими примерами.

7. Проверка гипотезы об однородности двухвыборочной модели

На практике часто встречаются задачи, в которых требуется выяснить, привело ли некоторое воздействие, усовершенствование или обработка к изменению наблюдаемого показателя. Например, привело ли предложенное усовершенствование производственного процесса к увеличению выпуска продукции; увеличивается ли урожайность пшеницы при внесении в почву удобрений; является ли разработанная биодобавка эффективным средством для снижения (увеличения) массы тела; повышается ли точность измерительного прибора после проведённой наладки. Для решения таких задач сначала требуется провести эксперимент (измерение нужного показателя) и получить две независимые выборки. Если при проведении эксперимента воздействие (усовершенствование производства, внесение удобрений, прием биодобавки, ремонт аппаратуры) к измеряемой величине не применяется, то полученные экспериментальные данные представляют выборку, которую принято называть контрольной. Наблюдения, полученные в ходе эксперимента с воздействием, составляют опытную выборку. Понятно, что полученные выборки будут различаться. Вопрос состоит в следующем: можно ли приписать эти различия случайной изменчивости наблюдаемого показателя или контрольная и опытная выборки имеют значимое различие? Если удаётся установить, что законы распределения контрольной и опытной выборок одинаковы, то это означает, что применяемое воздействие не изменяет наблюдаемый показатель. Если же законы распределения этих выборок различаются, то это различие можно приписать эффекту произведённого воздействия.

Отметим, что задача выявления различий может возникать и в ситуациях, когда выборки являются измерениями однотипных показателей, полученных в результате двух различных способов «обработки». Например, нужно выяснить различается ли по прочности сталь, производимая двумя различными методами; различается ли урожайность пшеницы, при применении двух различных удобрений; различается ли некоторый экономический показатель деятельности предприятий в двух регионах страны. Под «обработкой» в этих примерах понимают метод изготовления стали, сорт удобрения, региональный фактор. Различие законов распределения выборок можно приписать эффекту влияния «обработки» на измеряемый показатель.

Все описанные задачи сводятся к решению проблемы проверки одинаковой распределенности (однородности) двух случайных величин, порождающих две выборки. Для решения таких задач применяют статистические критерии проверки гипотезы об однородности в двухвыборочной модели. В этом разделе будут представлены пять наиболее важных и распространенных критериев. Почему же существует так много критериев для решения этой проблемы? Это обстоятельство, в основном, связано с двумя аспектами.

Первый — это характер неоднородности, который может определяться физическим смыслом задачи. Например, применение удобрений производится с целью повышения урожайности, употребление биодобавки — для снижения или наоборот увеличения веса. Понятно, что в этих случаях СВ, порождающие две выборки, различаются лишь средними значениями, а соответствующие им распределения — сдвигом. Если же требуется выяснить, одинакова ли точность двух однотипных не имеющих систематической ошибки измерительных приборов, то распределения выборок, соответствующих показаниям двух приборов, могут различаться только параметром масштаба. Случайные величины, порождающие эти выборки, могут различаться лишь дисперсиями. В случаях, когда известно, что неоднородность выборок связана со сдвигом или растяжением (сжатием), применяют специальные критерии, предназначенные только для таких типов неоднородности. Если характер неоднородности априори неизвестен, то существуют критерии, которые позволяют проверять однородность двух выборок против любых возможных альтернатив.

Второй аспект связан с ограничениями, накладываемыми на статистическую модель. Так, если априори известно или может быть проверено, что наблюдения имеют нормальное распределение, то оптимальными критериями проверки однородности являются критерии, называемые классическими. К сожалению, на практике закон распределения выборок редко бывает известен. Для этих ситуаций разработаны непараметрические критерии, которые не основаны на предположении о том, что выборки имеют некоторое определённое параметрическое распределение. К таким критериям относятся, представленные в этом разделе, ранговые критерии.

Здесь у читателя может возникнуть естественный вопрос о том, можно ли сравнить разные критерии проверки однородности? Какой из многочисленных критериев следует выбрать для решения конкретной задачи? В разделе 7.7 обсуждается проблема сравнения асимптотических эффективностей критериев при разных распределениях наблюдаемых величин.

7.1. Теоретические положения

Пусть выборка $\mathbb{X}_m = [X_1, \ldots, X_m]^\top$ соответствует распределению $F_X(t)$, а выборка $\mathbb{Y}_n = [Y_1, \ldots, Y_n]^\top$ распределению $F_Y(t)$.

Определение 7.1. Выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ называются *однородными*, если $F_X(t) = F_Y(t)$ для любого $t \in \mathbb{R}^1$.

Статистическую гипотезу вида

$$H_0: \ F_X(t) = F_Y(t), \ \ \forall t \in \mathbb{R}^1 \tag{7.1}$$

называют *гипотезой об однородности* выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$.

Альтернативной гипотезой общего вида для H_0 является гипотеза

$$H_1: \ \exists t \in \mathbb{R}^1, \ \text{такое что} \ F_X(t) \neq F_Y(t). \tag{7.2}$$

Неоднородность выборок может быть обусловлена разными причинами. Рассмотрим два важнейших типа неоднородности, которые можно описать, используя понятия сдвига и сжатия (растяжения) распределений $F_X(t)$ и $F_Y(t)$.

Пусть неоднородность выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$ состоит в том, что распределения $F_X(t)$ и $F_Y(t)$ различаются лишь сдвигом на некоторую величину θ, а именно:

$$F_Y(t) = F_X(t - \theta), \quad \forall t \in \mathbb{R}^1. \tag{7.3}$$

В этом случае будем говорить, что неоднородность выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$ вызвана наличием сдвига. Если неоднородность выборок обусловлена лишь сдвигом (7.3), то гипотеза об однородности формулируется следующим образом:

$$H_0: \ \theta = 0, \tag{7.4}$$

а альтернативные гипотезы имеют вид: $H_1 : \theta < 0$, $H_2 : \theta > 0$, $H_3 : \theta \neq 0$.

Отметим (см. пример 7.1), что если справедлива гипотеза H_1, и СВ X, порождающая выборку $\mathbb{X}_m$, имеет конечное математическое ожидание, то $\mathbf{M}\{Y\} < \mathbf{M}\{X\}$. Аналогично, из справедливости H_2 и $\mathbf{M}\{X\} < \infty$ следует, что $\mathbf{M}\{Y\} > \mathbf{M}\{X\}$, а из H_3 и $\mathbf{M}\{X\} < \infty$ следует, что $\mathbf{M}\{Y\} \neq \mathbf{M}\{X\}$.

Пусть выборка $\mathbb{X}_m$ соответствует распределению $F_X(t - \mu)$, а выборка $\mathbb{Y}_n$ распределению $F_Y(t) = F_X\left(\dfrac{t-\mu}{\Delta}\right)$, $\Delta \neq 0$, где функция $F(t)$ удовлетворяет условию $\int\limits_{-\infty}^{+\infty} t dF(t) = 0$, а μ — мешающий параметр сдвига. В этом случае математические ожидания СВ X и Y, порождающих выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$, совпадают (см. пример 7.2), а неоднородность выборок вызвана растяжением (сжатием). Гипотеза об однородности в этом случае имеет вид

$$H_0:\ \Delta = 1, \tag{7.5}$$

а альтернативные к ней гипотезы вид $H_1 : \Delta < 1$, $H_2 : \Delta > 1$, $H_3 : \Delta \neq 1$.

Пусть $\mathbf{D}\{X\} < \infty$, тогда из справедливости гипотезы H_1 следует (см. пример 7.2), что $\mathbf{D}\{X\} > \mathbf{D}\{Y\}$, из H_2 следует, что $\mathbf{D}\{X\} < \mathbf{D}\{Y\}$ и из H_3 следует, что $\mathbf{D}\{X\} \neq \mathbf{D}\{Y\}$.

Для проверки гипотезы об однородности используются различные критерии, применимость которых обусловлена различными требованиями, предъявляемыми к выборкам.

Определение 7.2. *Рангом* элемента выборки называют номер места, которое занимает этот элемент в вариационном ряду.

Процедуру определения рангов всех элементов выборки называют ранжированием.

Определение 7.3. Совокупность совпадающих наблюдений называется *связкой*. Количество наблюдений в связке называют *размером связки*. При ранжировании всем элементам связки присваивается *средний ранг*. Средний ранг связки определяется следующим образом: если связке предшествует k элементов вариационного ряда, и связка имеет размер m, то средний ранг этой связки равен $\frac{1}{m}\sum\limits_{i=k+1}^{k+m} i$. Заметим, что средний ранг может принимать как целые, так и дробные значения.

Критерии, базирующиеся на предположении о гауссовости выборок, принято называть *классическими*. Критерии, статистики которых являются функциями рангов наблюдений, называются *ранговыми критериями*.

Для проверки гипотезы об однородности вида (7.4) можно использовать, например, критерий Стьюдента и ранговый критерий Вилкоксона, а для проверки гипотезы вида (7.5) — критерий Фишера и ранговый критерий Ансари—Брэдли. Подробно рассмотрим эти критерии.

7.2. Критерий Стьюдента

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует распределению $\mathcal{N}\left(m_X; \sigma_X^2\right)$, а выборка $\mathbb{Y}_n$ распределению $\mathcal{N}\left(m_Y; \sigma_Y^2\right)$;

2) дисперсии σ_X^2 и σ_Y^2 одинаковы $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ и неизвестны;

3) выборки независимы.

Гипотеза об однородности в рамках этой модели имеет вид (7.4):

$$H_0:\ \theta = m_Y - m_X = 0.$$

Статистика критерия Стьюдента вычисляется следующим образом:

$$T(\mathbb{X}_m, \mathbb{Y}_n) = T(\mathbb{Z}_N) = \frac{\overline{Y}_n - \overline{X}_m}{S_N\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}, \tag{7.6}$$

где $S_N^2 = \dfrac{1}{N-2}\left[\sum\limits_{i=1}^{m}(X_i - \overline{X}_m)^2 + \sum\limits_{i=1}^{n}(Y_i - \overline{Y}_m)^2\right]$ — оценка неизвестной дисперсии σ^2 по объединенной выборке $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^\top$ объема $N = m + n$.

При справедливости гипотезы H_0 статистика $T(\mathbb{Z}_N)$ имеет распределение Стьюдента $\mathcal{T}_r$ с $r = N-2$ степенями свободы. В зависимости от типа задачи имеет смысл рассматривать альтернативные гипотезы H_A различного вида. Критические области уровня значимости $\alpha \in (0;1)$ критерия Стьюдента, соответствующие различным H_A, приведены в табл. 7.1, где $t_\gamma(r)$ — квантиль уровня γ распределения Стьюдента $\mathcal{T}_r$.

Таблица 7.1

H_A	Критические области для $T(\mathbb{Z}_N)$
$\theta < 0$	$(-\infty; t_\alpha(r))$
$\theta > 0$	$\big(t_{1-\alpha}(r); +\infty\big)$
$\theta \neq 0$	$\Big(-\infty; t_{\frac{\alpha}{2}}(r)\Big) \cup \Big(t_{1-\frac{\alpha}{2}}(r); +\infty\Big)$

Заметим, что при $N^* = \min(m,n) \to \infty$ статистика (7.6) асимптотически нормальна.

7.3. Критерий Вилкоксона

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует неизвестному непрерывному распределению $F(t)$, а выборка $\mathbb{Y}_n$ распределению $F(t-\theta)$;

2) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ независимы.

Критерий Вилкоксона позоляет проверить гипотезу вида (7.4)

$$H_0:\ \theta = 0.$$

Статистика критерия имеет вид

$$T(\mathbb{Z}_N) = W_{m,n} = \sum_{j=1}^{n} R_j, \tag{7.7}$$

где R_j — ранг элемента Y_j в объединенной выборке $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^\top$ объема $N = m + n$.

Для выборок небольшого объема можно найти точное распределение статистики $W_{m,n}$ при справедливости гипотезы H_0 вида (7.4). Квантили этого распределения табулированы в [7] для $1 \leqslant n \leqslant m \leqslant 25$.

Можно показать [12], что при справедливости H_0 для любых m и n

$$\mathbf{M}\Big\{W_{m,n}\Big\} = \frac{n}{2}(N+1), \quad \mathbf{D}\Big\{W_{m,n}\Big\} = \frac{m \cdot n}{12}(N+1),$$

а стандартизованная статистика

$$W^*_{m,n} = \frac{W_{m,n} - \mathbf{M}\Big\{W_{m,n}\Big\}}{\sqrt{\mathbf{D}\Big\{W_{m,n}\Big\}}}$$

при $N^* = \min(m, n) \to \infty$ асимптотически нормальна.

Если в выборке имеются связки, то при вычислении статистики $W^*_{m,n}$ следует заменить дисперсию $\mathbf{D}\Big\{W_{m,n}\Big\}$ на выражение

$$\widetilde{\mathbf{D}}\Big\{W_{m,n}\Big\} = \mathbf{D}\Big\{W_{m,n}\Big\} - \frac{m \cdot n \sum\limits_{k=1}^{l} t_k(t_k^2 - 1)}{12N(N-1)},$$

где l — количество связок в выборке $\mathbb{Z}_N$, а t_k — размер k-й связки, $k = 1, \ldots, l$.

Важно отметить, что распределение статистик $W_{m,n}$ и $W^*_{m,n}$ критерия Вилкоксона при справедливости H_0 не зависит от распределения $F(t)$. Критерии, статистики которых обладают таким свойством, принято называть *свободными от распределения.*

Отметим также, что для применения критерия Вилкоксона не требуется выполнения условия $\mathbf{M}\{X\} < \infty$.

Критические области уровня значимости α критерия Вилкоксона, основанного на статистике $W^*_{m,n}$, соответствующие различным альтернативам H_A, указаны в табл. 7.2. Через u_γ обозначена квантиль уровня γ распределения $\mathcal{N}(0; 1)$.

Таблица 7.2

H_A	Критические области для $W^*_{m,n}$
$\theta < 0$	$(-\infty; u_\alpha]$
$\theta > 0$	$\big(u_{1-\alpha}; +\infty\big)$
$\theta \neq 0$	$\Big(-\infty; u_{\frac{\alpha}{2}}\Big) \cup \Big(u_{1-\frac{\alpha}{2}}; +\infty\Big)$

Критерии Стьюдента и Вилкоксона при выполнении указанных выше предположений являются состоятельными для альтернатив вида $H_1: \theta < 0$, $H_2: \theta > 0$, $H_3: \theta \neq 0$.

7.4. Критерий Фишера

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует распределению $\mathcal{N}(m_X; \sigma_X^2)$, выборка $\mathbb{Y}_n$ — распределению $\mathcal{N}(m_Y; \sigma_Y^2)$, причем параметры m_X, m_Y, σ_X^2, σ_Y^2 неизвестны;

2) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ независимы.

Критерий Фишера позволяет проверить гипотезу

$$H_0: \ \Delta = \frac{\sigma_Y}{\sigma_X} = 1. \tag{7.8}$$

Если последняя гипотеза верна, и при этом $m_X = m_Y = m$, то верна гипотеза H_0 вида (7.5). Мешающий параметр сдвига в этом случае совпадает с математическим ожиданием m.

Статистика критерия Фишера вычисляется следующим образом:

$$T(\mathbb{X}_m, \mathbb{Y}_n) = F_{n,m} = \frac{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2}{\dfrac{1}{m-1}\sum\limits_{i=1}^{m}\left(X_i - \overline{X}_m\right)^2} = \frac{\widetilde{S}_Y^2}{\widetilde{S}_X^2}, \tag{7.9}$$

где $\widetilde{S}_X^2$, $\widetilde{S}_Y^2$ — несмещенные выборочные дисперсии, построенные по выборкам $\mathbb{X}_m$ и $\mathbb{Y}_n$ соответственно.

Если верна гипотеза H_0 вида (7.8), то статистика $F_{n,m}$ имеет распределение Фишера $F(n-1; m-1)$ с $n-1$ и $m-1$ степенями свободы.

В таблицах (см., например, [7]) представлены квантили распределения Фишера только высоких (не менее 0,5) уровней. Это обстоятельство связано с тем, что если СВ $T \sim F(n; m)$, то СВ $\dfrac{1}{T} \sim F(m; n)$. Тогда (см. пример 7.5) число $\dfrac{1}{f_\beta(m; n)}$, где $f_\beta(m; n)$ — квантиль уровня β распределения $F(m; n)$, является квантилью уровня $1-\beta$ распределения $F(n; m)$.

При различных альтернативах процедуру проверки гипотезы H_0 вида (7.8) с помощью критерия Фишера целесообразно организовать следующим образом.

Если альтернативная гипотеза имеет вид $H_1: \dfrac{\sigma_Y}{\sigma_X} = \Delta < 1$, и $\widetilde{S}_Y^2 < \widetilde{S}_X^2$, то следует рассмотреть статистику $F_{m,n} = \dfrac{1}{F_{n,m}} = \dfrac{\widetilde{S}_X^2}{\widetilde{S}_Y^2}$.

Эта статистика при справедливости H_0 вида (7.8) имеет распределение $F(m-1;n-1)$, поэтому критическая область имеет вид: $\big(f_{1-\alpha}(m-1;n-1);+\infty\big)$, где α — уровень значимости критерия, а $f_{1-\alpha}(m-1;n-1)$ — квантиль уровня $(1-\alpha)$ распределения $F(m-1;n-1)$. Если же $\widetilde{S}^2_Y > \widetilde{S}^2_X$, то принимается гипотеза H_0.

Если альтернативная гипотеза имеет вид $H_2 : \dfrac{\sigma_Y}{\sigma_X} = \Delta > 1$, и $\widetilde{S}^2_Y > \widetilde{S}^2_X$, то статистика $F_{n,m}$ имеет распределение $F(n-1;m-1)$, а критическая область критерия имеет вид: $\big(f_{1-\alpha}(n-1;m-1);+\infty\big)$, где $f_{1-\alpha}(n-1;m-1)$ — квантиль уровня $(1-\alpha)$ распределения $F(n-1;m-1)$. В случае когда $\widetilde{S}^2_Y < \widetilde{S}^2_X$, принимается гипотеза H_0.

Если альтернативная гипотеза имеет вид $H_3 : \dfrac{\sigma_Y}{\sigma_X} = \Delta \neq 1$, и $\widetilde{S}^2_Y > \widetilde{S}^2_X$, то статистикой критерия Фишера будет $F_{n,m}$. Если же $\widetilde{S}^2_Y < \widetilde{S}^2_X$, то статистикой критерия будет $F_{m,n} = \dfrac{1}{F_{n,m}}$. Соответствующие критические области будут иметь вид: $\Big(f_{1-\frac{\alpha}{2}}(n-1;m-1);+\infty\Big)$ в первом случае или $\Big(f_{1-\frac{\alpha}{2}}(m-1;n-1);+\infty\Big)$ во втором случае.

7.5. Критерий Ансари—Брэдли

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует неизвестному непрерывному распределению $F(t-\mu)$, а выборка $\mathbb{Y}_n$ — распределению $F\left(\dfrac{t-\mu}{\Delta}\right)$, $\Delta \neq 0$, причем параметры μ и Δ неизвестны, и $F(0) = 0{,}5$;

2) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ независимы.

Замечание. Если СВ X, порождающая выборку $\mathbb{X}_m$, имеет распределение $F(t-\mu_1)$, а СВ Y, порождающая выборку $\mathbb{Y}_n$, имеет распределение $F\left(\dfrac{t-\mu_2}{\Delta}\right)$, и параметры μ_1 и μ_2 — неизвестны, то рекомендуется (см. [39]) оценить параметры положения μ_1 и μ_2 выборочными медианами $\widehat{\mu}_X$ и $\widehat{\mu}_Y$. А затем построить преобразованные выборки $\widetilde{\mathbb{X}}_m = [X_1 - \widehat{\mu}_X, \ldots, X_m - \widehat{\mu}_X]^\top$ и $\widetilde{\mathbb{Y}}_n = [Y_1 - \widehat{\mu}_Y, \ldots, Y_n - \widehat{\mu}_Y]^\top$, и применить критерий Ансари—Брэдли к полученным выборкам $\widetilde{\mathbb{X}}_m$ и $\widetilde{\mathbb{Y}}_n$.

Гипотеза об однородности выборок имеет вид (7.5), т.е. H_0: $\Delta = 1$.

Статистика критерия Ансари—Брэдли имеет вид

$$A_{m,n} = \sum_{i=1}^{m} \left(\frac{N+1}{2} - \left| R_i - \frac{N+1}{2} \right| \right), \tag{7.10}$$

где R_i — ранг элемента X_i в объединенной выборке $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^\top$ объема $N = m + n$.

Точное распределение статистики $A_{m,n}$ при справедливости H_0 табулировано для $2 \leqslant m \leqslant n$ при $m + n \leqslant 20$ в [39].

Известно, что при справедливости H_0 вида (7.5) и любых $m, n \geqslant 1$

$$\mathbf{M}\Big\{A_{m,n}\Big\} = \begin{cases} \dfrac{m(N+2)}{4}, & \text{если } N \text{ — четное,} \\ \dfrac{m(N+1)^2}{4N}, & \text{если } N \text{ — нечетное;} \end{cases}$$

$$\mathbf{D}\Big\{A_{m,n}\Big\} = \begin{cases} \dfrac{mn(N+2)(N-2)}{48(N-1)}, & \text{если } N \text{ — четное,} \\ \dfrac{mn(N^2+3)(N+1)}{48N^2}, & \text{если } N \text{ — нечетное,} \end{cases}$$

а стандартизованная статистика

$$A^*_{m,n} = \frac{A_{m,n} - \mathbf{M}\Big\{A_{m,n}\Big\}}{\sqrt{\mathbf{D}\Big\{A_{m,n}\Big\}}} \tag{7.11}$$

асимптотически нормальна при $N^* = \min(m, n) \to \infty$.

Если в выборке $\mathbb{Z}_N$ имеются связки, то дисперсию $\sqrt{\mathbf{D}\Big\{A_{m,n}\Big\}}$ статистики $A_{m,n}$ следует заменить выражением

$$\widetilde{\mathbf{D}}\{A_{m,n}\} = \begin{cases} \dfrac{mn\left(16\sum\limits_{j=1}^{k} t_j R_j^2 - N(N+2)^2\right)}{16N(N-1)}, & \text{если } N \text{ — четное,} \\ \dfrac{mn\left(16N\sum\limits_{j=1}^{k} t_j R_j^2 - (N+1)^4\right)}{16N^2(N-1)}, & \text{если } N \text{ — нечетное,} \end{cases}$$

где k — количество связок в выборке $\mathbb{Z}_N$, t_j — размер j-й связки, R_j — средний ранг элементов j-й связки, $j = 1, \ldots, k$.

Отметим, что для применения критерия Ансари—Брэдли не требуется условия конечности дисперсии $\mathbf{D}\{X\} < \infty$. Критерий Ансари—Брэдли является свободным от распределения и состоятельным для альтернатив вида $H_1 : \Delta < 1$, $H_2 : \Delta > 1$, $H_3 : \Delta \neq 1$.

Критические области критерия Ансари—Брэдли, основанного на статистике $A^*_{m,n}$ для этих альтернатив, указаны в табл. 7.3, где α — уровень значимости, u_γ — квантиль стандартного гауссовского распределения.

Таблица 7.3

H_A	Критические области для $A^*_{m,n}$
$\Delta < 1$	$(-\infty; u_\alpha)$
$\Delta > 1$	$(u_{1-\alpha}; +\infty)$
$\Delta \neq 1$	$\left(-\infty; u_{\frac{\alpha}{2}}\right) \cup \left(u_{1-\frac{\alpha}{2}}; +\infty\right)$

7.6. Критерий Колмогорова—Смирнова

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует неизвестному непрерывному распределению $F_X(t)$, а выборка $\mathbb{Y}_n$ — непрерывному распределению $F_Y(t)$;

2) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ независимы.

Критерий Колмогорова—Смирнова позволяет проверить гипотезу об однородности вида (7.1) против альтернативной гипотезы общего вида (7.2).

Статистика Колмогорова—Смирнова имеет вид:

$$D_{m,n} = \sup_{t \in \mathbb{R}^1} |\widehat{F}_{X,m}(t) - \widehat{F}_{Y,n}(t)|,$$

где $\widehat{F}_{X,m}(t)$ и $\widehat{F}_{Y,n}(t)$ — выборочные функции распределения, построенные по выборкам $\mathbb{X}_m$ и $\mathbb{Y}_n$ соответственно.

Так как выборочная функция распределения монотонна и изменяется в конечном числе точек, то

$$D_{m,n} = \max_{1 \leqslant i \leqslant m+n} |\widehat{F}_{X,m}(Z_i) - \widehat{F}_{Y,n}(Z_i)|, \tag{7.12}$$

где $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^T$ — объединенная выборка объема $N = m + n$.

Точное распределение статистики $D_{m,n}$ при справедливости гипотезы H_0 вида (7.1) табулировано в [7] для $2 \leqslant m \leqslant n \leqslant 20$.

Если $N^* = \min\{m, n\} \to \infty$, то при справедливости H_0 статистика

$$D^*_{m,n} = \sqrt{\frac{nm}{n+m}} D_{m,n} \tag{7.13}$$

асимптотически имеет распределение Колмогорова с функцией распределения

$$K(t) = \sum_{k=-\infty}^{\infty} (-1)^k \exp\{-2k^2 t^2\}.$$

Квантили распределения $K(t)$ также табулированы в [7].

Критическая область уровня значимости α критерия Колмогорова–Смирнова, основанного на статистике (7.13), имеет вид: $\left(K_{1-\alpha}; +\infty\right)$, где $K_{1-\alpha}$ — квантиль уровня $1 - \alpha$ распределения Колмогорова.

7.7. Асимптотическая относительная эффективность по Питмену

Из вышеизложенного следует, что для проверки гипотезы об однородности двух выборок могут применяться различные статистические критерии. Например, для проверки гипотезы вида (7.4) можно использовать критерий Стьюдента, Вилкоксона и Колмогорова—Смирнова. Для проверки гипотезы вида (7.5) — критерий Фишера, Ансари—Брэдли и Колмогорова—Смирнова. В связи с этим встает вопрос о том, какой из критериев предпочтительнее применять в каждой конкретной задаче. Одним из инструментов, позволяющих производить сравнение критериев, является асимптотическая относительная эффективность (АОЭ) по Питмену одного критерия по отношению к другому.

Прежде чем дать определение АОЭ по Питмену, введем необходимые обозначения и допущения.

Пусть выборка $\mathbb{X}_n = [X_1, \ldots, X_n]^\top$ соответствует распределению $F(t-\theta)$. Функция распределения $F(t)$ обладает плотностью распределения $p(t)$ и удовлетворяет условию $F(0) = 0{,}5$.

Обозначим $T_n^{(1)}$ и $T_n^{(2)}$ — статистики двух состоятельных критериев для проверки гипотезы $H_0 : \theta = 0$ против альтернативы $H_A : \theta > 0$.

Пусть при справедливости гипотезы H_0 статистика $T_n^{(i)}$, $i = 1, 2$, имеет асимптотическое распределение $G_i(t)$, т.е.

$$T_n^{(i)} \xrightarrow{d} T^{(i)} \sim G_i(t) \text{ при } n \to \infty.$$

Обозначим $z_{1-\alpha}^{(i)}$, $i = 1, 2$ — квантиль уровня $1-\alpha$ распределения $G_i(t)$.

Пусть уровень значимости обоих критериев асимптотически равен $\alpha \in (0; 1)$, т.е.

$$\mathbf{P}\{T_n^{(i)} \geqslant z_{1-\alpha}^{(i)}\} \to \alpha \text{ при } n \to \infty.$$

Зафиксируем для обоих критериев значение функции мощности равным $\beta \in (\alpha; 1)$.

Рассмотрим последовательность альтернатив $H_A^{(j)} : \theta = \theta_j > 0$, таких что $\{\theta_j\} \to 0$ при $j \to \infty$, $j \in \mathbb{N}$.

Обозначим через $\{n_j^{(i)}\}$, $i = 1, 2$, соответствующие последовательности объемов выборок, при которых

$$\mathbf{P}_{\theta_j, F}\left\{T_{n_j^{(i)}}^{(i)} \geqslant z_{1-\alpha}^{(i)}\right\} \to \beta, \quad n_j^{(i)} \to \infty,$$

где $\mathbf{P}_{\theta_j, F}\{\cdot\}$ — вероятность, зависящая от значения параметра θ_j и распределения выборки $F(t)$. Понятно, что в силу состоятельности критериев $T^{(1)}$ и $T^{(2)}$ соответствующие $n_j^{(1)}$ и $n_j^{(2)}$ существуют.

Определение 7.4. Если предел

$$e\left(T^{(1)}, T^{(2)}\right) = \lim_{j\to\infty} \frac{n_j^{(2)}}{n_j^{(1)}}$$

существует и не зависит от выбора последовательности $\{\theta_j\}$, α и β, то он называется *асимптотической относительной эффективностью критерия* $T^{(1)}$ *по отношению к критерию* $T^{(2)}$.

Это определение было введено Питменом и опубликовано в [49].

Отметим, что ситуация, когда $e\left(T^{(1)}, T^{(2)}\right) > 1$, означает, что критерию со статистикой $T^{(2)}$ требуется больший объем выборки, чем критерию со статистикой $T^{(1)}$, чтобы при справедливости альтернативы достичь заданного уровня мощности. И, следовательно, критерий $T^{(2)}$ менее эффективен, чем критерий $T^{(1)}$.

Приведенное определение не дает конструктивного способа вычисления АОЭ $e\left(T^{(1)}, T^{(2)}\right)$. Использование следующей теоремы позволит вычислить значения АОЭ.

Теорема 7.1. *Пусть для критериев со статистиками* $T_n^{(1)}$ *и* $T_n^{(2)}$ *выполнены следующие условия регулярности*:

1) *критерии со статистиками* $T_n^{(i)}$, $i = 1, 2$ *являются состоятельными*;

2) *существуют последовательности* $\{m_n^{(i)}(\theta)\}$ *и* $\{\sigma_n^{(i)}(\theta)\}$ *такие, что случайные последовательности* $\dfrac{T_n^{(i)} - m_n^{(i)}(\theta)}{\sigma_n^{(i)}(\theta)}$, $i = 1, 2$ *асимптотически нормальны, причем равномерно по* θ *в окрестности* $\theta = 0$;

3) *существует* $\left.\dfrac{dm_n^{(i)}(\theta)}{d\theta}\right|_{\theta=0} = \left(m_n^{(i)}(0)\right)'$, $i = 1, 2$;

4) *для последовательности* $\{\theta_n\} \to 0$ *при* $n \to \infty$

$$\frac{\sigma_n^{(i)}(\theta_n)}{\sigma_n^{(i)}(0)} \to 1 \;\; \text{и} \;\; \frac{\left(m_n^{(i)}(\theta_n)\right)'}{\left(m_n^{(i)}(0)\right)'} \to 1, \; i = 1, 2;$$

5) $\dfrac{\left(m_n^{(i)}(0)\right)'}{\sqrt{n}\sigma_n^{(i)}(0)} \to c_i > 0$ *при* $n \to \infty$, $i = 1, 2$.

Тогда АОЭ критерия $T^{(1)}$ *относительно* $T^{(2)}$ *равна*

$$e\left(T^{(1)}, T^{(2)}\right) = \lim_{j\to\infty} \frac{n_j^{(2)}}{n_j^{(1)}} = \frac{c_1^2}{c_2^2}.$$

Величина c_i называется *эффективностью* критерия $T^{(i)}$.

Можно сказать, что величина c_i характеризует нормированную скорость изменения асимптотического среднего статистики $T^{(i)}$ в окрестности точки $\theta = 0$. Таким образом, чем больше значение эффективности c_i, тем быстрее критерий «реагирует» на альтернативу.

Определение 7.4 АОЭ и сформулированная теорема 7.1 естественным образом переносятся на критерии, предназначенные для проверки гипотез в двухвыборочных задачах.

Проведем сравнение асимптотических эффективностей критериев Вилкоксона и Стьюдента предназначеных для проверки гипотезы H_0 вида (7.4) при различных распределениях $F(t)$, обладающих плотностью распределения $p(t)$.

В качестве альтернативы рассмотрим $H_A\colon \theta > 0$.

Будем считать, что объемы выборок $m, n \to \infty$ и $\dfrac{m}{m+n} \to \lambda$, $0 < \lambda < 1$.

Можно показать (см. [38] и пример 7.9), что эффективность критерия Стьюдента $c_T = \dfrac{\sqrt{\lambda(1-\lambda)}}{\sigma}$.

Заметим, что эффективность критерия Стьюдента обратно пропорциональна среднеквадратическому отклонению $\sigma = \sigma_X = \sigma_Y$ случайных величин X и Y, порождающих выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$. В случае, когда выборки порождены распределениями с бесконечными дисперсиями, этот критерий имеет нулевую эффективность.

При фиксированном распределении $F(t)$ с конечной дисперсией критерий Стьюдента достигает максимальной эффективности при $\lambda = 0{,}5$, т.е. критерий наиболее эффективен при выборках одинакового объема.

Известно (см. [38]), что эффективность критерия Вилкоксона

$$c_W = \sqrt{12\lambda(1-\lambda)} \int\limits_{-\infty}^{\infty} p^2(t)dt.$$

Обозначим $e(W,T)$ — АОЭ по Питмену критерия Вилкоксона по отношению к критерию Стьюдента. Тогда, согласно теореме 7.1,

$$e(W,T) = \frac{c_W^2}{c_T^2} = 12\sigma^2 \left(\int\limits_{-\infty}^{\infty} p^2(t)dt \right)^2. \tag{7.14}$$

Приведем примеры значений АОЭ $e(W,T)$ для гауссовского $\mathcal{N}(0;\sigma^2)$, равномерного $R[-1;1]$, Лапласа $\mathcal{L}(1)$ и логистического $Lg(0;1)$ распределений.

Таблица 7.4

Распределение	$\mathcal{N}(0;\sigma^2)$	$R[-1;1]$	$\mathcal{L}(1)$	$Lg(0;1)$
$e(W,T)$	$\dfrac{3}{\pi} \approx 0{,}955$	1	1,5	$\dfrac{\pi^2}{9}$

Если $F(t)$ есть распределение Тьюки с функцией распределения

$$F_{\gamma}(t) = (1-\gamma)\Phi(t) + \gamma\Phi\left(\frac{t}{3}\right),$$

где $\Phi(t)$ — функция Лапласа, а $\gamma \in [0;1)$ — параметр «засорения», то значения $e(W,T)$ при различных значениях параметра γ представлены в табл. 7.5.

Таблица 7.5

γ	0	0,01	0,05	0,1	0,15
$e(W,T)$	$\frac{3}{\pi}$	1,009	1,196	1,373	1,496

Можно сказать, что даже при небольшом «засорении» наблюдений асимптотическая эффективность критерия Вилкоксона существенно больше, чем у критерия Стьюдента.

Здесь возникает вопрос о нижней границе АОЭ $e(W,T)$. Как показали Ходжес и Леман (см., например, [38] или [21]), если распределение $F(t)$ принадлежит классу симметричных распределений $\mathcal{F}_S = \{F(t) : F(t) = 1 - F(-t)\}$, то

$$\inf_{F(t)\in\mathcal{F}_S} e(W,T) = 0{,}864.$$

Последнее означает, что вне зависимости от типа симметричного распределения $F(t)$ АОЭ критерия Вилкоксона по отношению к критерию Стьюдента не будет ниже чем 0,864.

7.8. Примеры

Пример 7.1. Пусть выборка $\mathbb{X}_m$ порождена СВ X с непрерывным распределением $F(t)$, а выборка $\mathbb{Y}_n$ — СВ Y с распределением $F(t-\theta)$, и $\mathbf{M}\{X\} < \infty$. Докажите, что из справедливости гипотезы $H_1 : \theta < 0$ следует, что $\mathbf{M}\{Y\} < \mathbf{M}\{X\}$.

Решение. Пусть $p_Y(t)$ — плотность распределения $F_Y(t)$. Тогда

$$\mathbf{M}\{Y\} = \int_{-\infty}^{+\infty} tp_Y(t)dt = \int_{-\infty}^{+\infty} tp_X(t-\theta)dt = \int_{-\infty}^{+\infty} (\theta + z)p_X(z)dz =$$

$$= \theta + \int_{-\infty}^{+\infty} zp_X(z)dz = \theta + \mathbf{M}\{X\}.$$

Если справедлива гипотеза $H_1 : \theta < 0$, то $\mathbf{M}\{Y\} - \mathbf{M}\{X\} = \theta < 0$. Следовательно, $\mathbf{M}\{Y\} < \mathbf{M}\{X\}$. ■

Пример 7.2. Пусть выборка $\mathbb{X}_m$ порождена СВ X с непрерывным распределением $F(t-\mu)$, а выборка $\mathbb{Y}_n$ — СВ Y с распределением $F\left(\dfrac{t-\mu}{\Delta}\right)$, $\Delta \neq 0$, и функция $F(t)$ удовлетворяет условию $\int\limits_{-\infty}^{+\infty} tF'(t)dt = 0$. Предполагается, что $\mathbf{D}\{X\} < \infty$. Докажите, что из справедливости гипотезы $H_1 : \Delta < 1$ следует, что $\mathbf{D}\{X\} > \mathbf{D}\{Y\}$.

Решение. Пусть $p(t)$ — плотность распределения $F(t)$. Тогда, согласно примеру 7.1, $\mathbf{M}\{X\} = \mu$. Покажем, что и $\mathbf{M}\{Y\} = \mu$.

$$\mathbf{M}\{Y\} = \int\limits_{-\infty}^{+\infty} tp_Y(t)dt = \int\limits_{-\infty}^{+\infty} t\frac{1}{\Delta}p\left(\frac{t-\mu}{\Delta}\right)dt = \int\limits_{-\infty}^{+\infty} (\Delta z + \mu)p(z)dz =$$

$$= \Delta \int\limits_{-\infty}^{+\infty} zp(z)dz + \mu \int\limits_{-\infty}^{+\infty} p(z)dz.$$

Первый интеграл равен нулю по условию. Следовательно, $\mathbf{M}\{Y\} = \mu$.

Вычислим дисперсии $\mathbf{D}\{X\}$ и $\mathbf{D}\{Y\}$.

$$\mathbf{D}\{X\} = \int\limits_{-\infty}^{+\infty} (t-\mu)^2 p(t-\mu)dt = \int\limits_{-\infty}^{+\infty} z^2 p(z)dz.$$

$$\mathbf{D}\{Y\} = \int\limits_{-\infty}^{+\infty} (t-\mu)^2 p_Y(t)dt = \int\limits_{-\infty}^{+\infty} (t-\mu)^2 \frac{1}{\Delta}p\left(\frac{t-\mu}{\Delta}\right)dt =$$

$$= \int\limits_{-\infty}^{+\infty} z^2\Delta^2 p(z)dz = \Delta^2\mathbf{D}\{X\}.$$

Таким образом, $\dfrac{\mathbf{D}\{Y\}}{\mathbf{D}\{X\}} = \Delta^2$. Следовательно, если справедлива гипотеза $H_1 : \Delta < 1$, то $\mathbf{D}\{X\} > \mathbf{D}\{Y\}$. ■

Пример 7.3. Докажите, что при справедливости гипотезы H_0 вида (7.4) статистика (7.6) имеет распределение Стьюдента с $r = N-2$ степенями свободы.

Решение. Найдем математическое ожидание и дисперсию числителя статистики (7.6).

$$\mathbf{M}\{\overline{Y}_n - \overline{X}_m\} = \mathbf{M}\{\overline{Y}_n\} - \mathbf{M}\{\overline{X}_m\} = m_Y - m_X = \theta.$$

$$\mathbf{D}\{\overline{Y}_n - \overline{X}_m\} = \mathbf{D}\{\overline{Y}_n\} + \mathbf{D}\{\overline{X}_m\} = \frac{1}{n}\mathbf{D}\{Y_1\} + \frac{1}{m}\mathbf{D}\{X_1\} = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right).$$

Так как выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ соответствуют гауссовскому распределению, то СВ $\overline{Y}_n - \overline{X}_m$, являющаяся линейной комбинацией гауссовских СВ, имеет гауссовское распределение (см. разд. 21.5).

Тогда

$$\frac{\overline{Y}_n - \overline{X}_m - \theta}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim \mathcal{N}(0; 1).$$

Представим статистику (7.6) в виде

$$T(\mathbb{X}_m, \mathbb{Y}_n) = \frac{\overline{Y}_n - \overline{X}_m}{S_N\sqrt{\frac{1}{n} + \frac{1}{m}}} =$$
$$= \frac{\overline{Y}_n - \overline{X}_m}{\sigma^2\sqrt{\frac{1}{n} + \frac{1}{m}}} \frac{1}{\sqrt{\frac{1}{N-2}\left(\sum\limits_{i=1}^{m}\left(\frac{X_i - \overline{X}_m}{\sigma}\right)^2 + \sum\limits_{i=1}^{n}\left(\frac{Y_i - \overline{Y}_n}{\sigma}\right)^2\right)}}.$$

Покажем, что СВ $\dfrac{(m-1)\widetilde{S}_X^2}{\sigma^2} = \sum\limits_{i=1}^{m}\left(\dfrac{X_i - \overline{X}_m}{\sigma}\right)^2$ имеет распределение хи-квадрат с $(m-1)$-й степенью свободы и случайные величины $\overline{X}_m$ и $\widetilde{S}_X^2$ независимы.

Рассмотрим ортогональное преобразование $Z = B\mathbb{X}_m$, выбрав в качестве первой строки $m \times m$ матрицы B строку $\left[\dfrac{1}{\sqrt{m}}, \ldots, \dfrac{1}{\sqrt{m}}\right]$. Тогда $Z_1 = \sqrt{m}\,\overline{X}_m$, а $Z_2, \ldots, Z_m$ можно найти, используя процесс ортогонализации.

В силу ортогональности матрицы B имеем

$$\sum_{i=1}^{m} Z_i^2 = Z^\top Z = \mathbb{X}_m^\top B^\top B\mathbb{X}_m = \mathbb{X}_m^\top\mathbb{X}_m = \sum_{i=1}^{m} X_i^2.$$

Нетрудно видеть, что

$$\sum_{i=1}^{m} X_i^2 = \sum_{i=1}^{m}(X_i - \overline{X}_m)^2 + m\overline{X}_m^2.$$

Так как $\sum\limits_{i=1}^{m}(X_i - \overline{X}_m)^2 = (m-1)\widetilde{S}_X^2$, а $m\overline{X}_m^2 = Z_1^2$, имеем

$$Z_2^2 + \ldots + Z_m^2 = (m-1)\widetilde{S}_X^2.$$

Заметим также, что

$$(Z_1 - \sqrt{m}m_X)^2 + Z_2^2 + \ldots + Z_m^2 = (X_1 - m_X)^2 + \ldots + (X_m - m_X)^2,$$

где m_X — математическое ожидание СВ X, порождающей выборку $\mathbb{X}_m$.

Тогда совместная плотность независимых СВ $X_1, \ldots, X_m$

$$c \exp\left\{ -\frac{\sum\limits_{i=1}^{m}(x_i - m_X)^2}{2\sigma^2} \right\}$$

преобразуется к виду

$$c \exp\left\{ -\frac{\left((z_1 - m_X\sqrt{m})^2 + z_2^2 + \ldots + z_m^2\right)}{2\sigma^2} \right\}.$$

Следовательно, СВ $Z_1, \ldots, Z_m$ — независимы, и

$$\sqrt{m}\,\overline{X}_m = Z_1 \sim \mathcal{N}(\sqrt{m}m_X; \sigma^2),$$

а $\dfrac{(m-1)\widetilde{S}_X^2}{\sigma^2} = \dfrac{Z_2^2 + \ldots + Z_m^2}{\sigma^2} \sim \mathcal{H}_{m-1}$.

Аналогично, случайные величины $\overline{Y}_n$ и $\widetilde{S}_Y^2$ — независимы, а СВ $\dfrac{(n-1)\widetilde{S}_Y^2}{\sigma^2} \sim \mathcal{H}_{n-1}$.

Тогда в силу независимости выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$ случайная величина

$$\xi = \sum_{i=1}^{m}\left(\frac{X_i - \overline{X}_m}{\sigma}\right)^2 + \sum_{j=1}^{n}\left(\frac{Y_j - \overline{Y}_n}{\sigma}\right)^2 = \frac{(N-2)S_N^2}{\sigma^2}$$

имеет распределение хи-квадрат с $r = (m-1) + (n-1) = N - 2$ степенями свободы, и не зависима с $\overline{X}_m$ и $\overline{Y}_n$.

Таким образом, при справедливости гипотезы H_0 вида (7.4) СВ

$$\eta = \frac{\overline{Y}_n - \overline{X}_m}{\sqrt{\sigma^2\left(\dfrac{1}{n} + \dfrac{1}{m}\right)}} \sim \mathcal{N}(0; 1),$$

а статистика $T(\mathbb{X}_m, \mathbb{Y}_n) = \dfrac{\eta}{\sqrt{\dfrac{1}{N-2}\xi}}$ согласно определению 5.6 имеет распределение Стьюдента с $r = N - 2$ степенями свободы. ■

Пример 7.4. Постройте функцию распределения статистики Вилкоксона $W_{m,n}$ при справедливости гипотезы H_0 вида (7.3) — (7.4) для $m = 4$, $n = 2$. Найдите квантили распределения статистики $W_{4,2}$ уровня 0,9 и 0,1.

Решение. При справедливости гипотезы H_0 вида (7.3) — (7.4) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ являются однородными, и вероятности появления любого набора рангов игреков $(R_1, \ldots, R_n)$ в объединенной выборке $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^\top$ объема $N = m + n$ одинаковы. Поскольку существует C_N^n различных способов размещения элементов $Y_1, \ldots, Y_n$ среди $n + m$ элементов объединенной выборки, то вероятность появления любого набора рангов $(R_1, \ldots, R_n)$ будет равна $\dfrac{1}{C_N^n}$. Выпишем все возможные комбинации рангов игреков (R_1, R_2) для случая $m = 4$, $n = 2$ и вычислим соответствующие им значения статистики Вилкоксона $W_{4,2} = \sum\limits_{i=1}^{2} R_i$. Таких комбинаций будет $C_6^2 = 15$ (табл. 7.6).

Таблица 7.6

(R_1, R_2)	$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$	$(2,3)$	$(2,4)$	$(2,5)$
$W_{4,2}$	3	4	5	6	7	5	6	7
(R_1, R_2)	$(2,6)$	$(3,4)$	$(3,5)$	$(3,6)$	$(4,5)$	$(4,6)$	$(5,6)$	
$W_{4,2}$	8	7	8	9	9	10	11	

Ряд распределения статистики $W_{4,2}$ при справедливости гипотезы H_0 имеет вид:

$W_{4,2}$	3	4	5	6	7	8	9	10	11
P	1/15	1/15	2/15	2/15	3/15	2/15	2/15	1/15	1/15

По определению 5.2 квантилью уровня γ, где $\gamma \in (0; 1)$, распределения $F(t)$ называется такое число z_γ, что

$$z_\gamma = \min\{t : F(t) \geqslant \gamma\}.$$

Зная функцию распределения $F_W(t)$ статистики $W_{4,2}$, найдем квантили $W_{0,1}(4; 2)$ уровня 0,1 и $W_{0,9}(4; 2)$ уровня 0,9.

Поскольку в данном случае $F_W(3) = \dfrac{1}{15} \approx 0{,}067$, а $F_W(4) = \dfrac{2}{15} \approx$ $\approx 0{,}13$, имеем $W_{0,1}(4; 2) = \min\{t : F_W(t) \geqslant 0{,}1\} = 4$.

Аналогично, $F_W(9) = \dfrac{13}{15} \approx 0{,}87$, $F_W(10) = \dfrac{14}{15} \approx 0{,}93$. Значит, $W_{0,9}(4; 2) = \min\{t : F_W(t) \geqslant 0{,}9\} = 10$.

Заметим, что минимальное значение статистики $W_{m,n}$ соответствует ситуации, при которой все элементы второй выборки меньше всех элементов первой выборки, т.е. $(R_1, \ldots, R_n) = (1, \ldots, n)$, и $\min W_{m,n} = 1 + \ldots + n = \dfrac{n(n+1)}{2}$. Максимальное значение статистики $W_{m,n}$ соответствует ситуации, при которой все элементы второй выборки больше всех элементов первой выборки, т.е. $(R_1, \ldots, R_n) = (n+1, \ldots, n+m)$, и $\max W_{m,n} = (n+1) + \ldots + (n+m) = \dfrac{n(m+2n+1)}{2}$. Распределение статистики $W_{m,n}$ при справедливости H_0 симметрично относительно своего среднего значения $\mathbf{M}\Big\{W_{m,n}\Big\} = \dfrac{n(m+n+1)}{2}$. Поэтому статистические таблицы содержат квантили либо высоких (не менее 0,5) либо низких (менее 0,5) уровней. Таким образом, если известна квантиль уровня γ распределения $W_{m,n}$, то, используя симметрию распределения $F_W(t)$, можно найти квантиль уровня $1-\gamma$ распределения статистики $W_{m,n}$ из соотношения

$$W_{1-\gamma}(m;n) - \mathbf{M}\Big\{W_{m,n}\Big\} = \mathbf{M}\Big\{W_{m,n}\Big\} - W_{\gamma}(m;n).$$

Этим свойством можно было воспользоваться и в данном примере. Зная $W_{0,9}(4;2) = 10$ и вычисляя $\mathbf{M}\Big\{W_{4,2}\Big\} = \dfrac{2(2+4+1)}{2} = 7$, найдем $W_{0,1}(4;2) = 2\mathbf{M}\Big\{W_{4,2}\Big\} - W_{0,9}(4;2) = 14 - 10 = 4$. ■

П р и м е р 7.5. Пусть z_β — квантиль уровня β распределения Фишера $F(m;n)$. Докажите, что число $\delta = \dfrac{1}{z_\beta}$ является квантилью уровня $1-\beta$ распределения $F(n;m)$.

Р е ш е н и е. Пусть СВ $X \sim F(m;n)$. Тогда по определению 5.7 СВ $Y = \dfrac{1}{X}$ имеет распределение $F(n;m)$. По определению 5.2 число z_β есть квантиль уровня β непрерывного строго монотонного распределения, если

$$\beta = \mathbf{P}\{X \leqslant z_\beta\}.$$

Тогда

$$\beta = \mathbf{P}\Big\{X \leqslant z_\beta\Big\} = \mathbf{P}\left\{\frac{1}{X} > \frac{1}{z_\beta}\right\} = \mathbf{P}\left\{Y > \frac{1}{z_\beta}\right\} = 1 - \mathbf{P}\left\{Y \leqslant \frac{1}{z_\beta}\right\},$$

и $\mathbf{P}\left\{Y \leqslant \dfrac{1}{z_\beta}\right\} = 1 - \beta$.

Таким образом, $\dfrac{1}{z_\beta}$ является квантилью уровня $1-\beta$ распределения $F(n;m)$. ■

П р и м е р 7.6. Имеются данные Федеральной службы государственной статистики о среднедушевых денежных доходах населения (рублей в месяц) в 2008 г. по некоторым областям Центрального и Приволжского федеральных округов. Данные представлены в табл. 7.7.

Таблица 7.7

Центральный федеральный округ	Доход, руб. в месяц	Приволжский федеральный округ	Доход, руб. в месяц
Брянская область	10 043	Республика Башкортостан	14 253
Владимирская область	9 596	Республика Марий Эл	7 843
Воронежская область	10 305	Удмуртская Республика	9 581
Ивановская область	8 354	Чувашская Республика	8 594
Костромская область	9 413	Пермский край	16 119
Московская область	19 776	Кировская область	10 112
Орловская область	9 815	Пензенская область	10 173
Рязанская область	11 311	Ульяновская область	9 756
Тамбовская область	11 253		
Тверская область	10 856		
Тульская область	11 389		

Выясните, одинаковы ли в среднем среднедушевые доходы населения в этих округах. Уровень значимости считайте равным 0,05.

Решение. Пусть среднедушевые доходы по Центральному федеральному округу (ЦФО) являются выборкой $\mathbb{X}_m = [X_1, \ldots, X_m]^\top$ объема $m = 11$, соответствующей некоторому распределению $F_X(t)$, а доходы по Приволжскому федеральному округу (ПФО) — выборкой $\mathbb{Y}_n = [Y_1, \ldots, Y_n]^\top$ объема $n = 8$, соответствующей распределению $F_Y(t)$. В данной задаче естественно предположить, что неоднородность выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$ обусловлена различием средних значений СВ X (показатель среднедушевого дохода в ЦФО) и СВ Y (показатель среднедушевого дохода в ПФО), порождающих выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$. Тогда $F_Y(t) = F_X(t - \theta)$. Для проверки гипотезы $H_0 : \theta = 0$ об однородности выборок против альтернативы сдвига $H_A : \theta \neq 0$ можно применить критерий Вилкоксона.

Причина, по которой выбирается альтернативная гипотеза указанного вида, заключается в том, что мы не имеем априорной информации о том, что в каком-то из рассматриваемых нами округов показатели среднедушевых доходов должны быть больше или меньше, чем в другом округе.

Статистика критерия Вилкоксона имеет вид (7.7)

$$W_{m,n} = \sum_{j=1}^{n} R_j,$$

где R_j — ранг случайной величины Y_j в объединенной выборке $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^\top$. Проведем ранжирование объединенной выборки и вычислим реализацию статистики Вилкоксона

$$W_{11,8} = 1 + 3 + 5 + 7 + 10 + 11 + 17 + 18 = 72.$$

Критическая область, соответствующая уровню значимости 0,05, имеет вид: $\Big[\min W_{11,8}; W_{0,025}(11;8)\Big) \cup \Big(W_{0,975}(11;8); \max W_{11,8}\Big]$. По таблице [7] найдем квантиль $W_{0,025}(11;8) = 55$. Тогда (см. пример 7.4) квантиль

$$W_{0,975}(11;8) = 2\mathbf{M}\Big\{W_{11,8}\Big\} - W_{0,025}(11;8) = \frac{2n(m+n+1)}{2} - 55 = \\ = 160 - 55 = 105.$$

Таким образом, критическая область критерия Вилкоксона, основанного на статистике $W_{m,n}$, имеет вид:

$$\Big[\min W_{11,8};\ 55\Big) \cup \Big(105\,; \max W_{11,8}\Big].$$

Так как реализация статистики $W_{m,n}$ попадает в доверительную область, то гипотеза H_0 принимается на уровне значимости $\alpha = 0{,}05$.

Предположим теперь, что наблюдаемые СВ соответствуют гауссовскому распределению. Такое предположение допустимо, так как каждый элемент выборки является выборочным средним большого количества СВ. Тогда задача может быть формализована следующим образом. Среднедушевые доходы по ЦФО $X_1, \ldots, X_m$ являются выборкой объема $m = 11$, соответствующей распределению $\mathcal{N}(m_X; \sigma_X^2)$, а среднедушевые доходы по ПФО $Y_1, \ldots, Y_n$ — выборкой объема $n = 8$, соответствующей распределению $\mathcal{N}(m_Y; \sigma_Y^2)$.

В рамках такой модели требуется проверить гипотезу

$$H_0: \ \theta = m_Y - m_X = 0$$

против альтернативы $H_1 : \theta \neq 0$.

Так как дисперсии σ_X^2 и σ_Y^2 неизвестны, то сначала следует проверить гипотезу $H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$ против альтернативы $H_1 : \sigma_X^2 \neq \sigma_Y^2$. Применим для этого критерий Фишера.

Вычислим реализации выборочных средних и выборочных несмещенных дисперсий:

$$\overline{X}_m = \frac{1}{11}\sum_{i=1}^{11} X_i = 11\,101{,}0; \quad \widetilde{S}^2_X = \frac{1}{10}\sum_{i=1}^{11}\left(X_i - \overline{X}_m\right)^2 = (3025{,}4)^2;$$

$$\overline{Y}_n = \frac{1}{8}\sum_{i=1}^{8} Y_i = 10\,803{,}9; \quad \widetilde{S}^2_Y = \frac{1}{7}\sum_{i=1}^{8}\left(Y_i - \overline{Y}_m\right)^2 = (2860{,}3)^2.$$

Так как $\widetilde{S}^2_X > \widetilde{S}^2_Y$, то статистика Фишера будет иметь вид

$$T(\mathbb{X}_m, \mathbb{Y}_n) = F_{m,n} = \frac{\widetilde{S}^2_X}{\widetilde{S}^2_Y}.$$

Реализация статистики $F_{m,n} = 1{,}12$.

При справедливости гипотезы $H_0 : \sigma^2_X = \sigma^2_Y = \sigma^2$ статистика $F_{m,n}$ имеет распределение Фишера $F(10; 7)$. Выберем уровень значимости $\alpha = 0{,}05$, тогда критическая область имеет вид $\left(f_{1-\frac{\alpha}{2}}(n-1; m-1); +\infty\right) = (f_{0,975}(10; 7); +\infty)$, где $f_{0,975}(10; 7)$ квантиль распределения $F(10; 7)$. По таблице [7] находим, что $f_{0,975}(10; 7) = 4{,}76$. Следовательно, реализация статистики $F_{m,n}$ попала в доверительную область, и гипотеза H_0 принимается на уровне значимости $\alpha = 0{,}05$.

Теперь предположения, требуемые для применения критерия Стьюдента со статистикой (7.6), выполнены. Вычислим реализацию статистики критерия Стьюдента

$$T(\mathbb{X}_m, \mathbb{Y}_n) = T(\mathbb{Z}_N) = \frac{\overline{Y}_n - \overline{X}_m}{S_N\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}},$$

где

$$S^2_N = \frac{\left[\sum\limits_{i=1}^{m}(X_i - \overline{X}_m)^2 + \sum\limits_{i=1}^{n}(Y_i - \overline{Y}_m)^2\right]}{m+n-2} = \frac{(m-1)\widetilde{S}^2_X + (n-1)\widetilde{S}^2_Y}{m+n-2} =$$
$$= \frac{10\cdot(3025{,}4)^2 + 7\cdot(2860{,}3)^2}{11+8-2} = (2958{,}5)^2.$$

Окончательно получаем $T(\mathbb{Z}_N) = \dfrac{297{,}13}{2958{,}5\sqrt{0{,}22}} = 0{,}216$.

При справедливости $H_0 : \theta = m_Y - m_X = 0$ статистика критерия Стьюдента (7.6) имеет распределение Стьюдента $\mathcal{T}_r$ с

$r = m+n-2 = 17$ степенями свободы. Так как альтернативная гипотеза имеет вид $H_1 : \theta \neq 0$, то критическая область уровня значимости α имеет вид $(-\infty; t_{\frac{\alpha}{2}}(r)) \cup (t_{1-\frac{\alpha}{2}}(r); +\infty)$. Для $\alpha = 0{,}05$ по табл. 22.4 находим $t_{0{,}975}(17) = 2{,}11$ и, учитывая то, что $t_{1-\alpha}(r) = -t_\alpha(r)$, имеем $t_{0{,}025}(17) = -2{,}11$. Так как реализация статистики попала в доверительную область, то принимается гипотеза H_0 на уровне значимости $\alpha = 0{,}05$, и можно считать, что среднедушевые доходы в этих федеральных округах в среднем одинаковы. ■

П р и м е р 7.7. Станок штампует детали, размер которых соответствует заданному нормативу, т.е. вероятность превышения и занижения нормативного размера одинакова. Технологи провели наладку станка для того, чтобы уменьшить отклонения размеров изготовленных деталей от размера, требуемого стандартом. До и после наладки случайным образом было выбрано по 11 деталей. Оказалось, что размер деталей, выбранных до наладки, составил (в мм):

52,4; 56,1; 48,6; 46,5; 46,0; 42,2; 48,8; 56,6; 59,8; 49,7; 51,6.

Размер деталей, изготовленных после наладки станка (в мм):

49,3; 47,7; 52,9; 48,3; 49,1; 46,4; 47,0; 52,0; 51,5; 51,2; 49,8.

Можно ли считать, опираясь на эти данные, что точность изготовления деталей увеличилась после наладки станка? Уровень значимости считать равным 0,05.

Р е ш е н и е. Пусть размеры деталей, проверенные до наладки станка, являются выборкой $\mathbb{X}_m = [X_1, \ldots, X_m]^\top$ объема $m = 11$, порожденной непрерывной СВ X, а размеры деталей, проверенные после наладки, выборкой $\mathbb{Y}_n = [Y_1, \ldots, Y_n]^\top$ объема $n = 11$, порожденной непрерывной СВ Y.

Поскольку размер деталей до и после наладки станка соответствует заданному нормативу, то это означает, что медианы случайных величин X и Y одинаковы $\mu_X = \mu_Y = \mu$, и параметр μ совпадает с нормативным размером. Так как наладка станка производится с целью уменьшения отклонений размеров изготовленных деталей от размера, требуемого стандартом, то можно считать, что распределения случайных величин X и Y различаются лишь параметром масштаба Δ, т.е.

$$F_X(t) = F(t-\mu), \quad F_Y(t) = F\left(\frac{t-\mu}{\Delta}\right),$$

а $F(0) = 0{,}5$.

Тогда гипотеза $H_0 : \Delta = 1$ будет означать, что выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ однородны, и наладка не привела к ожидаемому результату. В качестве альтернативной гипотезы в этой задаче следует выбрать

$H_A : \Delta < 1$, так как справедливость этой альтернативы означает (см. пример 7.2), что $\mathbf{D}\{X\} > \mathbf{D}\{Y\}$, т. е. точность изготовления деталей увеличилась.

Для проверки указанной гипотезы можно применить критерий Ансари–Брэдли со статистикой (7.10)

$$T(\mathbb{Z}_N) = A_{m,n} = \sum_{i=1}^{m} \left(\frac{N+1}{2} - \left| R_i - \frac{N+1}{2} \right| \right),$$

где R_i — ранг элемента X_i в объединенной выборке $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^\top$ объема $N = m + n$.

Проранжировав реализацию объединенной выборки $\mathbb{Z}_N$, получим вектор искомых реализаций рангов

$$[r_1, \ldots, r_{11}]^T = [18, 20, 8, 4, 2, 1, 9, 21, 22, 12, 16]^\top.$$

Реализация статистики

$$A_{11,11} = 53.$$

К сожалению, таблицы точного распределения статистики $A_{m,n}$ при справедливости H_0 составлены только для выборок объема $n + m \leqslant 20$, поэтому для построения критической области придется воспользоваться аппроксимацией.

При справедливости H_0 и $N^* = \min(m, n) \to \infty$, стандартизованная статистика $A^*_{m,n}$ вида (7.11) является асимптотически нормальной.

Так как $N = m + n = 22$ — четное число, то

$$\mathbf{M}\Big\{A_{11,11}\Big\} = \frac{m(N+2)}{4} = \frac{11 \cdot 24}{4} = 66,$$

$$\mathbf{D}\Big\{A_{11,11}\Big\} = \frac{mn(N+2)(N-2)}{48(N-1)} = \frac{11 \cdot 11 \cdot 24 \cdot 20}{49 \cdot 21} = 57{,}62.$$

Следовательно, реализация стандартизованной статистики

$$A^*_{11,11} = \frac{53 - 66}{\sqrt{57{,}62}} = -1{,}71.$$

Критическая область $(-\infty; u_\alpha)$ для уровня значимости $\alpha = 0{,}05$ имеет вид $(-\infty; -1{,}65)$. Таким образом, реализация статистики попала в критическую область и гипотеза H_0 отвергается в пользу альтернативы H_A на уровне значимости 0,05, т.е. точность изготовления деталей увеличилась после наладки станка. ■

П р и м е р 7.8. Известно, что одним из факторов риска сердечно-сосудистых заболеваний является склад психоэмоциональной сферы человека. Медики выделяют две основных модели поведения людей.

Модель типа A характеризуется постоянным острым дефицитом времени и склонностью к соперничеству, модель типа B — спокойствием и размеренностью. Склонность к сердечно-сосудистым заболеваниям характерна для людей с моделью поведения типа A. Высказано предположение о том, что различие в типах поведения индивидуумов обусловлено их физиологическими различиями. Чтобы проверить это предположение, исследователи [46] сравнили максимальные уровни концентрации гормонов роста в плазме крови у испытуемых различных типов поведения. Получены следующие результаты (в мг/мл). Испытуемые с моделью поведения типа A:

3,6; 2,6; 4,7; 8,0; 3,1; 8,8; 4,6; 5,8; 4,0; 4,6.

Испытуемые с моделью поведения типа B:

14,9; 16,6; 15,9; 5,3; 10,5; 16,2; 17,4; 8,5; 15,6; 5,4; 9,8.

Можно ли, опираясь на эти результаты исследования, считать предположение верным?

Решение. Пусть результаты измерений по группе с поведением типа A представляются выборкой $\mathbb{X}_m = [X_1, \ldots, X_m]^\top$ объема $m = 10$, соответствующей непрерывному распределению $F_X(t)$, а результаты измерений по группе с поведением типа B — выборкой $\mathbb{Y}_n = [Y_1, \ldots, Y_n]^\top$ объема $n = 11$, соответствующей непрерывному распределению $F_Y(t)$.

Проверим гипотезу об однородности выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$. Так как медики не дают априорной информации о типе неоднородности, следует проверить гипотезу H_0 вида (7.1) против альтернативной гипотезы H_A общего вида (7.2). Для решения этой задачи можно применить критерий Колмогорова–Смирнова со статистикой (7.12)

$$D_{m,n} = \max_{1 \leqslant i \leqslant m+n} |\widehat{F}_{X,m}(Z_i) - \widehat{F}_{Y,n}(Z_i)|,$$

где $\left[Z_1, \ldots, Z_{m+n}\right]^\top = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^T$ — объединенная выборка. Используя графический метод (см. рис.) или простой перебор, можно видеть, что максимальное расхождение между реализациями выборочных функций распределения $\widehat{F}_{x,10}(t)$ и $\widehat{F}_{y,11}(t)$ достигается в точке $t = 8{,}8$, и реализация статистики

$$D_{10,11} = |\widehat{F}_{x,10}(8{,}8) - \widehat{F}_{y,11}(8{,}8)| = 1 - \frac{3}{11} = \frac{8}{11}.$$

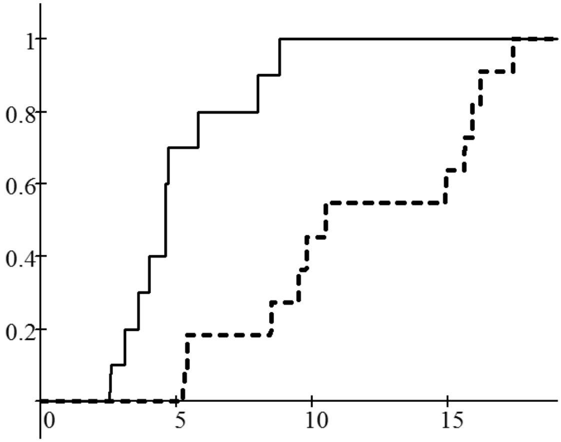


Рис. 7.1. — $\widehat{F}_{x,10}(t)$; --- $\widehat{F}_{y,11}(t)$

Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид $(z_{0,95}; 1)$, где $z_{0,95} = \dfrac{6}{11}$ — квантиль уровня 0,95 распределения статистики $D_{10,11}$ при справедливости гипотезы H_0. Так как реализация статистики $D_{10,11}$ попадает в критическую область, то гипотеза H_0 отвергается в пользу альтернативы H_A на уровне значимости 0,05.

Если считать, что m и n достаточно велики, и использовать статистику

$$D^*_{m,n} = \sqrt{\frac{nm}{n+m}} D_{m,n},$$

то критическая область будет иметь вид $(K_{0,95}; \infty)$, где $K_{0,95}$ — квантиль уровня 0,95 распределения Колмогорова $K(t)$. Согласно таблицам [7], $K_{0,95} = 1{,}36$. Реализация статистики

$$D^*_{m,n} = \sqrt{\frac{10 \cdot 11}{10+11}} \cdot \frac{8}{11} \approx 1{,}66$$

также попадает в критическую область.

Таким образом, гипотеза H_0 об однородности вида (7.1) отвергается на уровне значимости $\alpha = 0{,}05$. Следовательно, можно считать, что различие типов поведения людей обусловлено их физиологическими различиями. ■

Пример 7.9. Пусть выборка $\mathbb{X}_m$ порождена СВ X с непрерывным распределением $F(t)$, а выборка $\mathbb{Y}_n$ — СВ Y с непрерывным распределением $F(t - \theta)$ с конечной дисперсией σ^2. Для проверки гипотезы H_0: $\theta = 0$ против альтернативы H_A: $\theta > 0$ используется критерий со статистикой

$$T(\mathbb{X}_m, \mathbb{Y}_n) = T(\mathbb{Z}_N) = \overline{Y}_n - \overline{X}_m,$$

где $\overline{X}_m$, $\overline{Y}_n$ — выборочные средние, построенные по выборкам $\mathbb{X}_m$ и $\mathbb{Y}_n$, а $\mathbb{Z}_N$ — объединенная выборка объема $N = m + n$. Найдите асимптотическую эффективность c_T этого критерия.

Решение. Для того чтобы вычислить эффективность c_T, требуется проверить условия регулярности, сформулированные в теореме 7.1.

Согласно условию 1) теоремы 7.1 критерий со статистикой $T(\mathbb{Z}_N)$ должен быть состоятельным, т.е. функция мощности $W(S_\alpha, \theta) \to 1$ при $N^* = \min\{m, n\} \to \infty$ для любого фиксированного $\theta > 0$.

Действительно, пусть уровень значимости критерия равен $\alpha \in (0; 1)$, а $z_{1-\alpha}$ — квантиль уровня $1 - \alpha$ асимптотического распределения статистики $T(\mathbb{Z}_N)$ при справедливости гипотезы H_0. Тогда при $N^* \to \infty$ критическая область S_α имеет вид

$$S_\alpha = \left\{\mathbb{Z}_N : T(\mathbb{Z}_N) \geqslant z_{1-\alpha}\right\},$$

а функция мощности — вид $W(S_\alpha, \theta) = \mathbf{P}_\theta \{T(\mathbb{Z}_N) \geqslant z_{1-\alpha}\}$.

Для вычисления этой вероятности нужно найти асимптотическое распределение статистики $T(\mathbb{Z}_N)$ при справедливости гипотезы H_0 и при справедливости гипотезы H_A.

Согласно ЦПТ при $N^* \to \infty$ случайная величина $T(\mathbb{Z}_N)$ имеет гауссовское распределение с параметрами $\mathbf{M}\{T(\mathbb{Z}_N)\}$ и $\mathbf{D}\{T(\mathbb{Z}_N)\}$, где $\mathbf{M}\{T(\mathbb{Z}_N)\} = \mathbf{M}\{\overline{Y}_n\} - \mathbf{M}\{\overline{X}_m\} = \mathbf{M}\{Y\} - \mathbf{M}\{X\}$, а

$$\mathbf{D}\{T(\mathbb{Z}_N)\} = \mathbf{D}\{\overline{Y}_n - \overline{X}_m\} = \mathbf{D}\{\overline{Y}_n\} + \mathbf{D}\{\overline{X}_m\} = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right) = \sigma^2\frac{N}{n \cdot m}.$$

Согласно примеру 7.1, $\mathbf{M}\{Y\} - \mathbf{M}\{X\} = \theta$.

Обозначим $\lambda = \dfrac{m}{m+n} = \dfrac{m}{N}$. Пусть $0 < \lambda < 1$, тогда при достаточно большом N^* и справедливости H_0 статистика $T(\mathbb{Z}_N)$ имеет распределение $\mathcal{N}\left(0; \dfrac{\sigma^2}{\lambda(1-\lambda)N}\right)$, а при справедливости H_A распределение $\mathcal{N}\left(\theta; \dfrac{\sigma^2}{\lambda(1-\lambda)N}\right)$.

Таким образом

$$\begin{aligned} W(S_\alpha, \theta) &= \mathbf{P}_\theta\{T(\mathbb{Z}_N) \geqslant z_{1-\alpha}\} = \mathbf{P}_\theta\left\{T(\mathbb{Z}_N) \geqslant \frac{u_{1-\alpha}\sigma}{\sqrt{\lambda(1-\lambda)}\sqrt{N}}\right\} = \\ &= 1 - \Phi\left(\left(\frac{u_{1-\alpha}\sigma}{\sqrt{\lambda(1-\lambda)}\sqrt{N}} - \theta\right)\sqrt{\frac{\lambda(1-\lambda)N}{\sigma^2}}\right) = \\ &= 1 - \Phi\left(u_{1-\alpha} - \frac{\theta\sqrt{\lambda(1-\lambda)}\sqrt{N}}{\sigma}\right), \end{aligned}$$

где $u_{1-\alpha}$ — квантиль уровня $1 - \alpha$ распределения $\mathcal{N}(0; 1)$.

При любом фиксированном $\theta > 0$ аргумент функции Лапласа

$$u_{1-\alpha} - \frac{\theta\sqrt{\lambda(1-\lambda)}\sqrt{N}}{\sigma} \to -\infty \text{ при } N \to \infty,$$

следовательно, $\Phi\left(u_{1-\alpha} - \frac{\theta\sqrt{\lambda(1-\lambda)}\sqrt{N}}{\sigma}\right) \to 0$ при $N \to \infty$, а

$$W(S_\alpha, \theta) \to 1 \text{ при } N \to \infty.$$

Заметим, что из требования $N^* = \min\{m, n\} \to \infty$ следует, что $N = m + n \to \infty$. Таким образом, критерий со статистикой $T(\mathbb{Z}_N)$ является состоятельным для любой альтернативы H_A: $\theta > 0$.

Проверим теперь условие регулярности 2). Определим последовательности $m_N(\theta) = \theta$ и $\sigma_N(\theta) = \dfrac{\sigma}{\sqrt{\lambda(1-\lambda)}\sqrt{N}}$. Тогда случайная последовательность $\dfrac{T(\mathbb{Z}_N) - m_N(\theta)}{\sigma_N(\theta)}$ асимптотически нормальна, а равномерность по θ в окрестности $\theta = 0$ обеспечивается тем, что распределение случайных величин $\dfrac{T(\mathbb{Z}_N) - m_N(\theta)}{\sigma_N(\theta)}$ не зависит от θ.

Условия 3), 4) и 5) также справедливы, поскольку

$$\left.\frac{dm_N(\theta)}{d\theta}\right|_{\theta=0} = 1, \quad \frac{\sigma_N(\theta)}{\sigma_N(0)} = 1, \quad \frac{m'_N(\theta)}{m'_N(0)} = 1,$$

$$\frac{m'_N(0)}{\sqrt{N}\sigma_N(0)} = \frac{\sqrt{\lambda(1-\lambda)}}{\sigma} = c_T.$$

Таким образом, все условия регулярности выполнены, и величина $c_T = \dfrac{\sqrt{\lambda(1-\lambda)}}{\sigma}$, где $\lambda = \dfrac{m}{m+n}$, является эффективностью критерия со статистикой $T(\mathbb{Z}_N) = \overline{Y}_n - \overline{X}_m$. ■

7.9. Задачи для самостоятельного решения

1. Согласно опросам 29 семей, проводившимся в 1968 г. в юго-западном регионе Англии, выборочное среднее еженедельной арендной платы за мебелированную квартиру составило 2,5£, а выборочная дисперсия 0,67£^{2}. В Уэльсе выборочное среднее арендной платы 16 семей составило 2,06£, а выборочная дисперсия 0,42£^{2}. Проверьте, является ли различие арендной платы в этих регионах статистически значимым, предполагая, что выборки порождены гауссовскими случайными величинами. Примите уровень значимости равным 0,05.

2. Уровень гистамина в мокроте у семи курильщиков, склонных к аллергии, составил (в микрограммах): 102,4; 100,0; 67,5; 65,9; 64,7; 39,6; 31,2, а у десяти курильщиков, не склонных к аллергии: 48,1; 45,5; 41,7;

35,4; 29,1; 18,9; 58,3; 66,8; 71,3; 94,3. Верно ли предположение о том, что уровень гистамина у курильщиков, подверженных аллергии, выше, чем у неаллергиков? Примите уровень значимости равным 0,05.

3. Для определения содержания железистой сыворотки Рамсей использовал метод прямого определения. Этот метод очень трудоемкий, реакции протекают медленно, не исключено помутнение смеси. Для преодоления этих трудностей Jung и Parekh предложили улучшенный метод, основанный на недавно синтезированной присадке. Методика выполнения анализа новым методом явно лучше, чем у метода Рамсея. Однако возникло подозрение, что новый метод имеет меньшую точность, чем процедура Рамсея. Для сравнения точности нового и известного методов было проделано 20 пар анализов, каждый двумя методами. В качестве эталонов служили сыворотки Hyland, содержащие 105 микрограммов железистой сыворотки на 100 миллилитров. Данные представлены в таблице.

Метод Рамсея	111	107	100	99	102	106	109	108	104	99
Метод J.–P.	107	108	106	98	105	103	110	105	104	100
Метод Рамсея	101	96	97	102	107	113	116	113	110	98
Метод J.–P.	96	108	103	104	114	114	113	108	106	99

Можно ли считать на уровне значимости 0,05, что новый метод имеет меньшую точность, чем метод Рамсея?

4. Имеются данные Федеральной службы государственной статистики о среднем размере назначенных пенсий (руб.) по регионам Центрального и Сибирского федеральных округов в 2008 г.

Центральный федеральный округ	Размер пенсии	Сибирский федеральный округ	Размер пенсии
Белгородская область	4297,3	Республика Бурятия	4281,2
Брянская область	4244,3	Республика Тыва	4426,5
Ивановская область	4386,3	Кемеровская область	4571,7
Курская область	4085,4	Красноярский край	4896,0
Смоленская область	4321,5	Омская область	4339,5
Рязанская область	4290,5	Республика Алтай	4278,0
Московская область	4771,7		
г. Москва	4809,9		

Проверьте гипотезу о равенстве средних значений назначенных пенсий в Центральном и Сибирском федеральных округах, предполагая, что представленные наблюдения порождены гауссовскими случайными величинами. Уровень значимости считайте равным 0,05.

5. Используя формулу (7.14) для АОЭ $e(W,T)$ критерия Вилкоксона по отношению к критерию Стьюдента, вычислите значения $e(W,T)$, указанные в табл. 7.4.

6. В ателье по ремонту телевизоров имеются данные о времени (в неделях) безотказной (до первого ремонта) работы 9 телевизоров марки A и 8 телевизоров марки B. При этом для телевизоров марки A указанное время составило: 5; 12; 26; 50; 57; 110; 200; 230; 270 недель, а для телевизоров марки B: 1; 25; 31; 42; 54; 70; 250; 260 недель. Можно ли считать среднее время безотказной работы у телевизоров марки A и B одинаковыми? Примите уровень значимости равным 0,05.

7. Деятельность отделения банка характеризуется некоторым показателем X. Для проверки была случайным образом выбрана группа из 10 однотипных отделений банка. Показатель X у этих отделений составил: 258, 588, 477, 577, 619, 614, 641, 543, 517, 593. После экономического кризиса показатель X у 9 случайным образом выбранных отделений составил: 537, 398, 256, 440, 376, 524, 527, 589, 479. Можно ли считать, опираясь на эти данные, что экономический кризис привел к снижению показателя X? Примите уровень значимости равным 0,05.

8. Для проверки гипотезы H_0 вида (7.3) — (7.4) против альтернатив вида H_1: $\theta < 0$, H_2: $\theta > 0$, H_3: $\theta \neq 0$ Манном и Уитни был предложен критерий, основанный на статистике

$$U_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n} \varphi(X_i, Y_j), \quad \text{где} \quad \varphi(z,v) = \begin{cases} 1, & \text{если} \ z < v, \\ 0, & \text{если} \ z \geqslant v. \end{cases}$$

Покажите, что при отсутствии связок, статистика Манна—Уитни $U_{m,n}$ и статистика Вилкоксона $W_{m,n}$ связаны соотношением $W_{m,n} = U_{m,n} + \dfrac{n(n+1)}{2}$.

9. Два завода изготовляют электролампы одинакового типа. Из продукции завода № 1 было случайным образом выбрано 10 ламп, а из продукции завода № 2 — 12 ламп. Испытания по длительности горения (в часах) этих ламп показали следующие результаты. Для ламп завода № 1: 1243, 1238, 1253, 1243, 1254, 1260, 1251, 1246, 1255, 1237. Для ламп завода № 2: 1244, 1255, 1258, 1266, 1249, 1257, 1260, 1247, 1256, 1271, 1252, 1259. Проверьте гипотезу о равенстве средней продолжительности горения электроламп завода № 1 и завода № 2. Примите уровень значимости равным 0,05.

Указание. При построении доверительной и критической областей используйте точные значения квантилей статистики.

10. Проверьте гипотезу об однородности двух выборок из задачи 9 с помощью критерия Колмогорова—Смирнова. Примите уровень значимости равным 0,05.

Указание. Согласно таблицам [7] квантиль $D_{1-0{,}049}(10;12) = \dfrac{33}{60}$.

8. Однофакторный дисперсионный анализ

В предыдущем разделе были изучены критерии, предназначенные для выявления однородности (неоднородности) двух выборок, которые являлись измерениями однотипных показателей, полученных в результате различных «обработок». В частности, в примере 7.6 мы рассмотрели показатели среднедушевых доходов населения в двух федеральных округах и, используя различные статистические критерии, показали, что среднедушевые доходы населения в Центральном и Приволжском округах можно считать в среднем одинаковыми. Это означает, что способ «обработки», под которым можно понимать экономические условия конкретного региона, не оказывает влияния на измеряемый показатель (среднедушевые доходы). Обобщим теперь эту задачу. Пусть имеется три или более выборок, соответствующих различным обработкам (округам). Требуется выяснить, равны ли средние значения измеряемого показателя для всех обработок. Задачу выявления однородности (или неоднородности) трех или большего числа выборок, которые могут различаться сдвигом, называют задачей дисперсионного анализа. В этом разделе будут рассмотрены классические и непараметрические ранговые критерии для решения задачи однофакторного дисперсионного анализа и проведено сравнение асимптотических эффективностей этих критериев.

8.1. Теоретические положения

Пусть имеется k независимых выборок $Z_1 = [X_{11}, X_{21}, \ldots, X_{n_1 1}]^\top, \ldots, Z_k = [X_{1k}, X_{2k}, \ldots, X_{n_k k}]^\top$, порожденных СВ $X_1, \ldots, X_k$ с распределениями $F(t - \theta_1), \ldots, F(t - \theta_k)$ соответственно.

Требуется проверить гипотезу

$$H_0: \ \theta_1 = \cdots = \theta_k = \theta \tag{8.1}$$

против альтернативы

$$H_A: \ \exists i, j, \ \text{такие что} \ \theta_i \neq \theta_j, \ i \neq j. \tag{8.2}$$

Справедливость гипотезы H_0 означает, что выборки $Z_1, \ldots, Z_k$ однородны, и объединенная выборка $\mathbb{Z}_N = [Z^\top{}_1, \ldots, Z^\top{}_k]^\top = [X_{11}, \ldots, X_{n_1 1}, \ldots, X_{1k}, \ldots, X_{n_k k}]^\top$ объема $N = n_1 + \cdots + n_k$ является однородной выборкой соответствующей распределению $F(t - \theta)$. Если же гипотеза H_0 нарушается, то это означает, что среди k рассматриваемых выборок найдутся выборки, распределения которых различаются сдвигом. Предполагается, что этот сдвиг вызван воздействием (влиянием) одной или нескольких переменных. Такие переменные называют *факторами*. Если предполагается наличие только одного фактора, то задача проверки гипотезы (8.1) называется *задачей однофакторного дисперсионного анализа*. При описании задач однофакторного анализа принято использовать следующие термины:

— *уровень фактора* (или способ обработки) — конкретная реализация фактора;

— *отклик* — значение измеряемой СВ.

Фактор может быть как количественной, так и качественной переменной. Однако при решении задачи однофакторного дисперсионного анализа должно быть выбрано конечное число k различных уровней фактора, и при этом реализации $\left[x_{1j}, \ldots, x_{n_j j}\right]^\top$ выборок Z_j, $j = 1, \ldots, k$, должны быть откликами, соответствующими j-му уровню фактора.

Отметим, что описанная задача проверки гипотезы H_0 вида (8.1) является обобщением задачи проверки гипотезы об однородности двух выборок против альтернативы сдвига на случай $k > 2$ выборок.

Таблица 8.1, в которой в первой строке записаны уровни факторов, а x_{ij}, $j = 1, \ldots, k$, $i = 1, \ldots, n_j$ есть реализации СВ X_{ij}, называется таблицей с одним входом или таблицей однофакторного анализа.

Таблица 8.1

1	2	$\ldots$	k
x_{11}	x_{22}	$\ldots$	x_{1k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{n_1 1}$	$x_{n_2 2}$	$\ldots$	$x_{n_k k}$

Рассмотрим классический F-критерий и ранговые критерии Краскела—Уоллиса и Джонкхиера для проверки гипотезы вида (8.1).

8.2. Критерий Краскела—Уоллиса

Пусть справедливо следующее предположение:

$F(t)$ — непрерывная функция распределения с плотностью распределения $p(t)$.

Обозначим R_{ij} — ранг X_{ij} в объединенной выборке $\mathbb{Z}_N = \left[X_{11}, \ldots, X_{n_1 1}, \ldots, X_{1k}, \ldots, X_{n_k k}\right]^\top$, а $\overline{R}_{\cdot j} = \dfrac{1}{n_j} \sum\limits_{i=1}^{n_j} R_{ij}$ — средний ранг элементов выборки, соответствующий j-му уровню фактора, $j = 1, \ldots, k$.

Статистика критерия Краскела—Уоллиса имеет вид

$$T(\mathbb{Z}_N) = H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{R}_{\cdot j} - \frac{N+1}{2}\right)^2. \qquad (8.3)$$

Для удобства вычислений можно использовать другую форму этой статистики:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{1}{n_j} \left(\sum_{i=1}^{n_j} R_{ij}\right)^2 - 3(N+1). \qquad (8.4)$$

Если в выборке $\mathbb{Z}_N$ имеются связки, то рекомендуется использовать модифицированную форму статистики H вида

$$H' = \frac{H}{1 - \frac{1}{N^3 - N}\sum\limits_{i=1}^{g}(t^3{}_i - t_i)}, \tag{8.5}$$

где g — количество связок, а t_i — размер i-й связки.

Заметим, что для вычисления статистики (8.3) не обязательно знать количественные реализации откликов, достаточно иметь их совместную ранжировку. Поэтому табл. 8.1 однофакторного анализа заменим на табл. 8.2.

Таблица 8.2

1	2	$\ldots$	k
r_{11}	r_{22}	$\cdots$	r_{1k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$r_{n_1 1}$	$r_{n_2 2}$	$\cdots$	$r_{n_k k}$

Точные квантили статистики (8.3) при справедливости гипотезы H_0 вида (8.1) представлены в [28] для следующих значений

$$k = 3, \quad 2 \leqslant n_1 \leqslant n_2 \leqslant n_3 \leqslant 8;$$

$$k = 4, \quad 2 \leqslant n_1 \leqslant \ldots \leqslant n_4 \leqslant 4;$$

$$k = 5, \quad 2 \leqslant n_1 \leqslant \ldots \leqslant n_5 \leqslant 3.$$

Статистика (8.3) имеет распределение хи-квадрат $\mathcal{H}_r$ с $r = k - 1$ степенями свободы при справедливости гипотезы H_0 вида (8.1) и $\min\{n_1, \ldots, n_k\} \to \infty$.

При нарушении гипотезы (8.1) расхождение между средним рангом $\dfrac{N+1}{2}$ объединенной выборки $\mathbb{Z}_N$ и средними рангами $\overline{R}_{\cdot j}$, $j = 1, \ldots, k$ столбцов, соответствующих j-м уровням фактора, будет большим. Поэтому статистика (8.3) в случае справедливости альтернативы (8.2) будет принимать большие значения, а критическая область уровня значимости α будет иметь вид $(k_{1-\alpha}(k-1); +\infty)$, где $k_{1-\alpha}(k-1)$ — квантиль уровня $1-\alpha$ распределения $\mathcal{H}_{k-1}$.

8.3. Критерий Джонкхиера

Пусть справедливо следующее предположение:

$F(t)$ — непрерывная функция распределения с плотностью распределения $p(t)$.

Критерий Джонкхиера позволяет проверить гипотезу H_0 вида (8.1) против альтернативы

$$H_A: \ \theta_1 \leqslant \theta_2 \leqslant \dots \leqslant \theta_k, \tag{8.6}$$

где хотя бы одно из неравенств строгое.

Альтернативы такого вида принято называть упорядоченными. Упорядоченные альтернативы описывают ситуацию, при которой увеличение уровня фактора вызывает увеличение сдвига распределения соответствующей этому уровню выборки относительно распределения первой выборки.

Если имеется априорное предположение о том, что с увеличением уровня фактора средние значения СВ $X_1, \dots, X_k$ уменьшаются, то следует перенумеровать столбцы табл. 8.2 в обратном порядке.

Введем обозначения

$$\varphi(y,z) = \begin{cases} 1, & \text{если } y < z, \\ 0{,}5, & \text{если } y = z, \\ 0, & \text{если } y > z; \end{cases}$$

$$U_{lm} = \sum_{i=1}^{n_l} \sum_{j=1}^{n_m} \varphi\left(X_{il}, X_{jm}\right).$$

Статистика критерия Джонкхиера имеет вид

$$T(\mathbb{Z}_N) = J = \sum_{1 \leqslant l < m \leqslant k} U_{l,m}. \tag{8.7}$$

Заметим, что $U_{lm} = \sum\limits_{i=1}^{n_l} \sum\limits_{j=1}^{n_m} \varphi\left(X_{il}, X_{jm}\right) = \sum\limits_{i=1}^{n_l} \sum\limits_{j=1}^{n_m} \varphi\left(R_{il}, R_{jm}\right)$, т.е. при вычислении реализации статитстики Джонкхиера (8.7) можно использовать либо реализации выборок $z_1, \dots, z_k$, либо реализации рангов объединенной выборки $\mathbb{Z}_N$.

Можно показать, что при справедливости гипотезы H_0 вида (8.1)

$$\mathbf{M}\{J\} = \frac{1}{4}\left[N^2 - \sum_{j=1}^{k} n_j^2\right]; \ \mathbf{D}\{J\} = \frac{1}{72}\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right],$$

а стандартизованная статистика

$$J^* = \frac{J - \mathbf{M}\{J\}}{\sqrt{\mathbf{D}\{J\}}}$$

при $\min(n_1, \dots, n_k) \to \infty$ асимптотически нормальна.

Точные квантили распределения статистики J представлены в [39] для следующих значений

$$k=3,\ \ 2\leqslant n_1\leqslant n_2\leqslant n_3\leqslant 8;$$
$$k=4,5,6,\ \ 2\leqslant n_1=\ldots=n_k\leqslant 6.$$

Критическая область уровня значимости α критерия Джонкхиера, основанного на статистике J^*, и соответствующая альтернативе (8.6), имеет вид $(u_{1-\alpha};+\infty)$, где $u_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения $\mathcal{N}(0;1)$.

Отметим, что при $k=2$ статистика $J=W_{n_1,n_2}-\dfrac{n_2(n_2+1)}{2}$, где W_{n_1,n_2} — статистика критерия Вилкоксона (7.7), вычисленная для выборок Z_1 и Z_2. Таким образом, статистика Джонкхиера (8.7) с точностью до известной константы представляется в виде суммы $\dfrac{k(k-1)}{2}$ статистик W_{n_l,n_m}, вычисленных для всех возможных $C_k^2=\dfrac{k(k-1)}{2}$ пар выборок Z_l и Z_m, $1\leqslant l<m\leqslant k$.

Для состоятельности критерия Джонкхиера против альтернатив вида (8.6) достаточно, чтобы $n_j\to\infty$ так, что $\dfrac{n_j}{N}\to\lambda_j$, $0<\lambda_j<1$, $j=1,\ldots,k$.

8.4. Классический F-критерий

Для представления классического критерия удобно описать рассматриваемую задачу с помощью следующей статистической модели:

$$X_{ij}=\theta+\tau_j+\varepsilon_{ij},\quad j=1,\ldots,k,\ \ i=1,\ldots,n_j, \tag{8.8}$$

где

θ — неизвестное математическое ожидание;

τ_j — неизвестные отклонения от общего среднего θ, вызванные изменениями уровня фактора (эффект j-й обработки), причем $\sum\limits_{j=1}^{k}\tau_j=0$;

ε_{ij} — независимые ненаблюдаемые погрешности соответствующие распределению $F(t)$, причем $\mathbf{M}\Big\{\varepsilon_{ij}\Big\}=0$.

В рамках такой модели параметры $\theta-\tau_j$, $j=1,\ldots,k$ совпадают с параметрами θ_j из разд. 8.1, и гипотеза (8.1) будет иметь вид

$$H_0:\ \tau_1=\tau_2=\cdots=\tau_k=0, \tag{8.9}$$

а альтернативная гипотеза (8.2) —

$$H_A: \ \exists j \text{ такое, что } \tau_j \neq 0. \tag{8.10}$$

Пусть справедливо следующее предположение:

ε_{ij}, $j = 1, \ldots, k$, $i = 1, \ldots, n_j$ соответствуют распределению $\mathcal{N}(0; \sigma^2)$ с неизвестной дисперсией σ^2.

Статистика F-критерия для проверки гипотезы (8.9) имеет вид

$$T(\mathbb{Z}_N) = \frac{\dfrac{1}{k-1}\displaystyle\sum_{j=1}^{k} n_j(\overline{X}_{\cdot j} - \overline{X}_N)^2}{\dfrac{1}{N-k}\displaystyle\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2}, \tag{8.11}$$

где $\overline{X}_{\cdot j} = \dfrac{1}{n_j}\displaystyle\sum_{i=1}^{n_j} X_{ij}$ — выборочное среднее, построенное по выборке Z_j, $j = 1, \ldots, k$, а $\overline{X}_N = \dfrac{1}{N}\displaystyle\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}$ — выборочное среднее, построенное по объединенной выборке $\mathbb{Z}_N = \left[Z^\top{}_1, \ldots, Z^\top{}_k\right]^\top$ объема $N = n_1 + \cdots + n_k$.

Как показано в примере 8.1, статистика F-критерия при справедливости гипотезы H_0 вида (8.9) имеет F-распределение Фишера $F(k-1; N-k)$ с $k-1$ и $N-k$ степенями свободы, а при справедливости H_A вида (8.10) — нецентральное F-распределение Фишера $F(k-1; N-k; \delta)$ с $k-1$ и $N-k$ степенями свободы и параметром нецентральности $\delta = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{k} n_j\left(\tau_j - \sum_{j=1}^{k}\frac{n_j}{N}\tau_j\right)^2$.

Понятно, что при больших значениях отклонений τ_j параметр нецентральности δ тоже будет принимать большие значения. Следовательно, и статистика (8.11) при нарушении H_0 будет принимать большие значения. Таким образом, критическая область уровня значимости α F-критерия будет иметь вид:

$$(f_{1-\alpha}(k-1; N-k); \infty),$$

где $f_{1-\alpha}(k-1; N-k)$ — квантиль уровня $1-\alpha$ распределения $F(k-1; N-k)$.

Если гипотеза H_0 вида (8.9) принимается, то выборки $Z_1, \ldots, Z_k$, полученные при различных значениях уровня фактора, однородны, и, следовательно, можно считать, что фактор не оказывает влияния на отклик. В этом случае задача исследования влияния фактора на отклик завершена. Если же гипотеза H_0 отвергнута, то это свидетельствует о том, что фактор оказывает влияние на отклик. В связи с этим возникает задача оценивания и сравнения средних значений СВ $X_1, \ldots, X_k$, порождающих выборки $Z_1, \ldots, Z_k$. Этой задаче посвящен следующий раздел.

Теперь проведем сравнение критерия Краскела—Уоллиса и классического F-критерия. Можно показать [38], что АОЭ по Питмену критерия Краскела—Уоллиса по отношению к F-критерию

$$e(H, F) = 12\sigma^2 \left(\int\limits_{-\infty}^{\infty} p^2(t)dt \right)^2 .$$

Выражение для $e(H, F)$ совпадает с определенной в (7.14) АОЭ $e(W, T)$ критерия Вилкоксона по отношению к критерию Стьюдента. Поэтому сравнительный анализ АОЭ для различных плотностей распределений $p(t)$, проведенный в параграфе 7.1 для $e(W, T)$, целиком переносится на АОЭ $e(H, F)$.

8.5. Доверительное оценивание параметров сдвига и контрастов

Обозначим $\theta_j = \theta + \tau_j$, $j = 1, \ldots, k$ математическое ожидание СВ X_j, порождающей выборку $Z_j = \left[X_{1j}, \ldots, X_{n_j j}\right]^\top$.

Тогда модель (8.8) будет иметь вид

$$X_{ij} = \theta_j + \varepsilon_{ij}, \quad j = 1, \ldots, k, \ \ i = 1, \ldots, n_j. \tag{8.12}$$

Предположим, что ε_{ij}, $j = 1, \ldots, k$, $i = 1, \ldots, n_j$ имеют распределение $\mathcal{N}(0; \sigma^2)$ с неизвестной дисперсией σ^2.

Построим доверительные интервалы для параметров θ_j, $j = 1, \ldots, k$. Как показано в примере 8.2, статистика

$$G(\mathbb{Z}_N, \theta_j) = \frac{\sqrt{n_j}(\overline{X}_{\cdot j} - \theta_j)}{\sqrt{\dfrac{1}{N-k} \sum\limits_{j=1}^{k} \sum\limits_{i=1}^{n_j} (X_{ij} - \overline{X}_{\cdot j})^2}}, \quad j = 1, \ldots, k \tag{8.13}$$

является центральной статистикой для θ_j и имеет распределение Стьюдента $\mathcal{T}_{N-k}$, где $N = \sum\limits_{j=1}^{k} n_j$.

Тогда центральный доверительный интервал параметра θ_j, $j = 1, \ldots, k$ надежности $1 - p$ имеет вид:

$$\mathbf{P}\left(\overline{X}_{\cdot j} - \frac{1}{\sqrt{n_j}}\sqrt{\widetilde{S}^2_N}\, t_{1-\frac{p}{2}}(N-k) < \theta_j < \right.$$
$$\left. < \overline{X}_{\cdot j} + \frac{1}{\sqrt{n_j}}\sqrt{\widetilde{S}^2_N}\, t_{1-\frac{p}{2}}(N-k)\right) = 1 - p, \tag{8.14}$$

где $\widetilde{S}^2_N = \dfrac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2$, а $t_{1-\frac{p}{2}}(N-k)$ — квантиль уровня $1 - \dfrac{p}{2}$ распределения $\mathcal{T}_{N-k}$.

На практике при проведении сравнительного анализа бывает важно строить доверительные интервалы не только для средних значений θ_j, но и для разностей средних значений.

Определение 8.1. *Контрастом параметров* θ_j, $j = 1, \ldots, k$ в модели (8.12) называется величина $\gamma = \sum\limits_{j=1}^{k} c_j\,\theta_j$, где c_j — константы, удовлетворяющие условию $\sum\limits_{j=1}^{k} c_j = 0$.

Например, если положить значения $c_l = 1$, $c_m = -1$ и $c_j = 0$ для $j \neq l$ и $j \neq m$, то контраст $\gamma = \theta_l - \theta_m$ представляет разность средних значений откликов, соответствующих l-му и m-му уровням фактора.

Несмещенной оценкой контраста γ является статистика (задача 2)

$$\widehat{\gamma} = \sum_{j=1}^{k} c_j\,\widehat{\theta}_j = \sum_{j=1}^{k} c_j\,\overline{X}_{\cdot j}. \tag{8.15}$$

Можно показать (задача 3), что центральная статистика для γ имеет вид

$$G(\mathbb{Z}_N, \gamma) = \frac{\sum\limits_{j=1}^{k} c_j \overline{X}_{\cdot j} - \gamma}{\sqrt{\sum\limits_{j=1}^{k} \dfrac{c_j^2}{n_j}\left(\dfrac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2\right)}} \tag{8.16}$$

и $G(\mathbb{Z}_N, \gamma) \sim \mathcal{T}_{N-k}$.

Тогда, обозначая $\widetilde{S}^2_N = \dfrac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2$, получим

$$\mathbf{P}\left(-t_{1-\frac{p}{2}}(N-k) < G(\mathbb{Z}_N, \gamma) < -t_{1-\frac{p}{2}}(N-k)\right) = 1 - p,$$

$$\mathbf{P}\left(\sum_{j=1}^{k} c_j\overline{X}_{\cdot j} - t_{1-\frac{p}{2}}(N-k)\sqrt{\widetilde{S}^2_N \sum_{j=1}^{k}\frac{c_j^2}{n_j}} < \gamma < \right.$$

$$\left. < \sum_{j=1}^{k} c_j\overline{X}_{\cdot j} + t_{1-\frac{p}{2}}(N-k)\sqrt{\widetilde{S}^2_N \sum_{j=1}^{k}\frac{c_j^2}{n_j}}\right) = 1-p,$$

и центральный доверительный интервал контраста γ уровня надежности $1-p$ имеет вид

$$\begin{gathered}\left(\sum_{j=1}^{k} c_j\overline{X}_{\cdot j} - t_{1-\frac{p}{2}}(N-k)\sqrt{\widetilde{S}^2_N \sum_{j=1}^{k}\frac{c_j^2}{n_j}}\ ;\right.\\ \left.\sum_{j=1}^{k} c_j\overline{X}_{j} + t_{1-\frac{p}{2}}(N-k)\sqrt{\widetilde{S}^2_N \sum_{j=1}^{k}\frac{c_j^2}{n_j}}\right).\end{gathered} \tag{8.17}$$

8.6. Примеры

П р и м е р 8.1. Пусть случайные величины X_j, $j=1,\ldots,k$, $i = 1,\ldots,n_j$, описываются моделью (8.8), где $\varepsilon_{ij} \sim \mathcal{N}(0;\sigma^2)$. Найдите распределение статистики (8.11) при справедливости гипотезы H_0 вида (8.9) и при справедливости H_A вида (8.10).

Р е ш е н и е. Разложим сумму квадратов SS (sum of squares) отклонений СВ X_{ij} от выборочного среднего $\overline{X}_N$ на составляющие

$$\begin{aligned} SS &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_N)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_N+\overline{X}_{\cdot j}-\overline{X}_{\cdot j})^2 = \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\cdot j})^2 + \sum_{j=1}^{k} n_j(\overline{X}_{\cdot j}-\overline{X}_N)^2 + \\ &+2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\cdot j})(\overline{X}_{\cdot j}-\overline{X}_N). \end{aligned}$$

Нетрудно видеть, что последнее слагаемое равно нулю, так как

$$\begin{aligned} &\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\cdot j})(\overline{X}_{\cdot j}-\overline{X}_N) = \sum_{j=1}^{k}\left[(\overline{X}_{\cdot j}-\overline{X}_N)\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\cdot j})\right] = \\ &= \sum_{j=1}^{k}(\overline{X}_{\cdot j}-\overline{X}_N)(n_j\overline{X}_{\cdot j}-n_j\overline{X}_{\cdot j}) = 0. \end{aligned}$$

Таким образом, квадратичная форма SS от X_{ij} представляется в виде суммы двух компонент

$$SS = SS_1 + SS_2,$$

где $SS_1 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2$ является суммой квадратов отклонений каждого элемента выборки от выборочного среднего соответствующего столбца, а $SS_2 = \sum_{i=1}^{n_j} n_j(\overline{X}_{\cdot j} - \overline{X}_N)^2$ — суммой квадратов отклонений между выборочными средними столбцов и выборочным средним объединенной выборки $\mathbb{Z}_N$.

В связи с этим SS_1 принято называть «суммой квадратов внутри (within) групп» и обозначать SS_w, а SS_2 — «суммой квадратов между (between) группами» и обозначать SS_b. Общую (total) сумму квадратов SS принято обозначать SS_t.

Установим независимость СВ $\frac{1}{\sigma^2}SS_w$ и $\frac{1}{\sigma^2}SS_b$, пользуясь теоремой Фишера—Кочрена. Согласно теореме Фишера—Кочрена (см. [33] п. 3b.4), если квадратичная форма $Y^\top Y$, где $Y = [Y_1, \ldots, Y_r\}]^\top$, а Y_j, $i = 1, \ldots, r$ — независимые случайные величины с распределением $\mathcal{N}(m_i; 1)$, представима в виде

$$Y^\top Y = Q_1 + Q_2,$$

где Q_1 и Q_2 — квадратичные формы рангов r_1 и r_2 соответственно, то необходимым и достаточным условием независимости Q_1 и Q_2 является равенство $r = r_1 + r_2$.

Найдем ранги соответствующих квадратичных форм и укажем распределения СВ $\frac{1}{\sigma^2}SS_t$, $\frac{1}{\sigma^2}SS_b$, $\frac{1}{\sigma^2}SS_w$.

Пусть справедлива гипотеза H_0 вида (8.9), тогда $X_{ij} \sim \mathcal{N}(\theta; \sigma^2)$ для всех $j = 1, \ldots, k$, $i = 1, \ldots, n_j$. Сделаем ортогональное преобразование $Y = B\mathbb{Z}_N$ выборки $\mathbb{Z}_N = \left[X_{11}, \ldots, X_{n_1 1}, \ldots, X_{1k}, \ldots, X_{n_k k}\right]^\top$, выбирая в качестве первой строки ортогональной матрицы B строку $\left(\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}\right)$.

Тогда справедливы следующие соотношения:

$$Y_1 = \sqrt{N}\,\overline{X}_N,$$

$$\sum_{i=1}^{N} Y_i^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_N)^2 + N(\overline{X}_N)^2.$$

Данное преобразование переводит квадратичную форму $SS_t = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_N)^2$ в форму $\sum_{i=2}^{N} Y_i^2$ ранга $r = N - 1$.

Случайная величина $\frac{1}{\sigma^2} SS_t = \frac{1}{\sigma^2}(Y_2^2 + \ldots + Y_N^2)$ имеет распределение хи-квадрат $\mathcal{H}_r$ с $r = N - 1$ степенями свободы.

Рассуждая аналогичным образом, заключаем, что квадратичная форма $\frac{1}{\sigma^2}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2$ для каждого фиксированного $1 \leqslant j \leqslant k$ имеет ранг $n_j - 1$. В силу независимости СВ X_{ij}, $j = 1, \ldots, k$, $i = 1 \ldots, n_j$ квадратичная форма $\frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2$ имеет ранг $r_1 = \sum_{j=1}^{k}(n_j - 1) = = N - k$, а СВ $\frac{1}{\sigma^2} SS_w$ распределение хи-квадрат $\mathcal{H}_{r_1}$ с $r_1 = N - k$ степенями свободы.

Отметим, что и при справедливости гипотезы H_A вида (8.10) СВ $\frac{1}{\sigma^2} SS_w$ также будет иметь распределение $\mathcal{H}_{r_1}$, так как $\mathbf{M}\left\{X_{ij}\right\} = \mathbf{M}\left\{\overline{X}_{\cdot j}\right\} = \theta + \tau_j$ при каждом фиксированном $j = 1, \ldots, k$.

Рассмотрим квадратичную форму $SS_b = \sum_{j=1}^{k} n_j(\overline{X}_{\cdot j} - \overline{X}_N)^2$. Сделаем ортогональное преобразование $\widetilde{Y} = C\widetilde{X}$ вектора $\widetilde{X} = \left(\sqrt{n_1}\overline{X}_{\cdot 1}, \ldots, \sqrt{n_k}\overline{X}_{\cdot k}\right)^T$, выбирая в качестве первой строки ортогональной матрицы C строку $\left(\frac{\sqrt{n_1}}{\sqrt{N}}, \ldots, \frac{\sqrt{n_k}}{\sqrt{N}}\right)$.

Учитывая, что

$$\overline{X}_N = \frac{1}{N}\left(n_1\overline{X}_{\cdot 1} + \ldots + n_k\overline{X}_{\cdot k}\right),$$

$$\sum_{j=1}^{k}\left(\sqrt{n_j}\overline{X}_{\cdot j}\right)^2 = \sum_{j=1}^{k} n_j(\overline{X}_{\cdot j} - \overline{X}_N)^2 + N(\overline{X}_N)^2,$$

$$\widetilde{Y}_1 = \sqrt{N}\overline{X}_N,$$

$$\sum_{j=1}^{k}\widetilde{Y}_j^2 = \sum_{j=1}^{k}\left(\sqrt{n_j}\overline{X}_{\cdot j}\right)^2,$$

заключаем, что квадратичная форма $SS_b = \sum_{j=2}^{k}\widetilde{Y}_j^2$ имеет ранг $k - 1$.

При справедливости гипотезы H_0 вида (8.9) СВ $\frac{1}{\sigma^2}SS_b$ имеет распределение хи-квадрат $\mathcal{H}_{r_2}$ с $r_2 = k-1$ степенью свободы.

Если же справедлива H_A вида (8.10), то $\mathbf{M}\Big\{X_{ij}\Big\} = \theta + \tau_j$, а

$$\mathbf{M}\Big\{\sqrt{n_j}(\overline{X}_{\cdot j} - \overline{X}_N)\Big\} = \sqrt{n_j}\mathbf{M}\left\{\overline{X}_{\cdot j} - \sum_{j=1}^{k}\frac{n_j}{N}\overline{X}_{\cdot j}\right\} =$$

$$= \sqrt{n_j}\left(\theta + \tau_j - \sum_{j=1}^{k}\frac{n_j}{N}(\theta + \tau_j)\right) = \sqrt{n_j}\left(\tau_j - \sum_{j=1}^{k}\frac{n_j}{N}\tau_j\right).$$

Тогда, согласно определению 5.5, СВ $\frac{1}{\sigma^2}SS_b$ имеет нецентральное распределение хи-квадрат $\mathcal{H}_{r_2,\delta}$ с $r_2 = k-1$ степенями свободы и параметром нецентральности $\delta = \frac{1}{\sigma^2}\sum\limits_{j=1}^{k} n_j\left(\tau_j - \sum\limits_{j=1}^{k}\frac{n_j}{N}\tau_j\right)^2$.

Таким образом, мы получили представление квадратичной формы

$$\frac{1}{\sigma^2}SS_t = \frac{1}{\sigma^2}SS_w + \frac{1}{\sigma^2}SS_b,$$

в котором ранг левой части $r = N-1$ равен сумме рангов $r_1 + r_2$ квадратичных форм правой части, где $r_1 = N-k$ и $r_2 = k-1$. Следовательно, $\frac{1}{\sigma^2}SS_w$ и $\frac{1}{\sigma^2}SS_b$ независимы.

Тогда отношение

$$F = \frac{\frac{1}{k-1}\frac{1}{\sigma^2}SS_b}{\frac{1}{N-k}\frac{1}{\sigma^2}SS_w} = \frac{\frac{1}{k-1}\sum\limits_{j=1}^{k} n_j(\overline{X}_{\cdot j} - \overline{X}_N)^2}{\frac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2}$$

при справедливости гипотезы H_0 вида (8.9) имеет, согласно определению 5.7, F-распределение с $k-1$ и $N-k$ степенями свободы. Если справедлива гипотеза H_A вида (8.10), то, согласно определению 5.8, статистика F имеет нецентральное F-распределение с $k-1$ и $N-k$ степенями свободы и параметром нецентральности $\delta = \frac{1}{\sigma^2}\sum\limits_{j=1}^{k} n_j\left(\tau_j - \sum\limits_{j=1}^{k}\frac{n_j}{N}\tau_j\right)^2$. ■

П р и м е р 8.2. Постройте центральную статистику $G(\mathbb{Z}_N;\theta_j)$ для параметра θ_j, определенного в модели (8.12), при фиксированном $1 \leqslant j \leqslant k$.

Р е ш е н и е. Так как выборка $Z_j = \left[X_{1j}, \ldots, X_{n_j j}\right]^T$ соответствует распределению $\mathcal{N}(\theta_j; \sigma^2)$, то по теореме 1.2

$$\overline{X}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} \sim \mathcal{N}\left(\theta_j; \frac{\sigma^2}{n_j}\right),$$

а статистика $\dfrac{\sqrt{n_j}(\overline{X}_{\cdot j} - \theta_j)}{\sigma} \sim \mathcal{N}(0;1)$.

Покажем, что оценка

$$\widetilde{S}^2_N = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2,$$

построенная по объединенной выборке $\mathbb{Z}_N = \left[X_{11}, \ldots, X_{n_1 1}, \ldots, X_{1k}, \ldots, X_{n_k k}\right]^\top$ объема $N = \sum_{j=1}^{k} n_j$, является несмещенной оценкой параметра σ^2.

Действительно,

$$\begin{aligned}
&\mathbf{M}\left\{\widetilde{S}^2_N\right\} = \mathbf{M}\left\{\frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2\right\} = \\
&= \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\mathbf{M}\left\{(X_{ij} - \overline{X}_{\cdot j})^2\right\} = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\mathbf{D}\left\{X_{ij} - \overline{X}_{\cdot j}\right\} = \\
&= \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\mathbf{D}\{X_{ij}\} + \mathbf{D}\left\{\overline{X}_{\cdot j}\right\} - 2\mathbf{cov}(X_{ij}, \overline{X}_{\cdot j})\right) = \\
&= \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\sigma^2 + \frac{\sigma^2}{n_j} - \frac{2}{n_j}\mathbf{cov}\left(X_{ij}, \sum_{i=1}^{n_j} X_{ij}\right)\right) = \\
&= \frac{1}{N-k}\sum_{j=1}^{k}(n_j\sigma^2 + \sigma^2 - 2\sigma^2) = \frac{1}{N-k}(N\sigma^2 - k\sigma^2) = \sigma^2.
\end{aligned}$$

В примере 8.1 было показано, что СВ $\dfrac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2$ имеет распределение хи-квадрат $\mathcal{H}_{N-k}$ с $N-k$ степенями свободы.

Проведя по аналогии с примером 7.3 ортогональное преобразование вектора $\mathbb{Z}_N$, можно показать, что СВ $\overline{X}_{\cdot j}$ и $\widetilde{S}^2_N$ независимы.

Тогда СВ

$$G(\mathbb{Z}_N; \theta_j) = \frac{\sqrt{n_j}(\overline{X}_{\cdot j} - \theta_j)}{\sigma\sqrt{\dfrac{\frac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2}{\sigma^2}}} = \frac{\sqrt{n_j}(\overline{X}_{\cdot j} - \theta_j)}{\sqrt{\frac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2}}$$

имеет, согласно определению 5.6, распределение Стьюдента $\mathcal{T}_{N-k}$ с $N-k$ степенями свободы.

Так как случайная функция $G(\mathbb{Z}_N; \theta_j)$ является монотонной и непрерывной по θ_j для каждой фиксированной реализации z_N выборки $\mathbb{Z}_N$, то, согласно определению 5.3, $G(\mathbb{Z}_N; \theta_j)$ является центральной статистикой для θ_j. ■

Пример 8.3. Имеются данные (в руб.) Федеральной службы государственной статистики о среднедушевых месячных денежных

Таблица 8.3

Центральный федеральный округ	Приволжский федеральный округ	Дальневосточный федеральный округ
Брянская область 10 043	Кировская область 10 112	Камчатский край 19 063
Владимирская область 9 596	Республика Марий Эл 7 843	Магаданская область 19 703
Воронежская область 10 305	Удмуртская Республика 9 581	Сахалинская область 24 552
Ивановская область 8 354	Пермский край 16 119	Хабаровский край 15 705
Костромская область 9 413	Чувашская Республика 8 594	Еврейская автономная область 10 877
Московская область 19 776	Республика Башкортостан 14 253	Чукотский автономный округ 32 140
Орловская область 9 815	Пензенская область 10 173	
Рязанская область 11 311	Ульяновская область 9 756	
Тамбовская область 11 253		
Тверская область 10 856		
Тульская область 11 389		

доходах населения в 2008 г. по некоторым областям Центрального, Приволжского и Дальневосточного федеральных округов. Данные представлены в табл. 8.3. Можно ли считать, что средние значения среднедушевых месячных доходов населения одинаковы во всех трех округах?

Решение. Фактором, т.е. переменной, которая может оказывать влияние на измеряемую величину (среднедушевой доход), является федеральный округ. Фактор в данном случае имеет три уровня: 1 — «Центральный федеральный округ», 2 — «Приволжский федеральный округ», 3 — «Дальневосточный федеральный округ».

Имеющиеся данные представляются тремя выборками $Z_1 = [X_{11}, \dots, X_{n_1 1}]^\top$ объема $n_1 = 11$, $Z_2 = [X_{12}, \dots, X_{n_2 2}]^\top$ объема $n_2 = 8$ и $Z_3 = [X_{13}, \dots, X_{n_3 3}]^\top$ объема $n_3 = 6$, которые соответствуют непрерывным распределениям $F(t - \theta_1)$, $F(t - \theta_2)$ и $F(t - \theta_3)$.

Проверим гипотезу об однородности H_0: $\theta_1 = \theta_2 = \theta_3$ против альтернативы

$$H_A: \text{ не все } \theta_1, \theta_2, \theta_3 \text{ равны между собой.}$$

Для проверки этой гипотезы можно применить критерий Краскела—Уоллиса.

Статистика критерия имеет вид (8.3)

$$T(\mathbb{Z}_N) = H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{R}_{\cdot j} - \frac{N+1}{2}\right)^2 .$$

Для вычисления реализации статистики составим табл. 8.4 реализаций рангов СВ X_{ij}, $j = 1, 2, 3$, $i = 1, \dots, n_j$.

Таблица 8.4

Уровни фактора	Реализации рангов										
1	9	6	12	2	4	23	8	16	15	13	17
2	18	1	5	3	20	10	11	7			
3	21	19	22	24	14	25					

Следовательно, $\overline{R}_{\cdot 1} = \frac{1}{11}\sum_{i=1}^{11} R_{i1} = 11{,}36$, $\overline{R}_{\cdot 2} = \frac{1}{8}\sum_{i=1}^{8} R_{i2} = 9{,}38$, $\overline{R}_{\cdot 3} = \frac{1}{6}\sum_{i=1}^{6} R_{i3} = 20{,}83$ и реализация статистики критерия

$$T(z_N) = \frac{12}{25 \cdot 26}\left(11(11{,}36 - 13)^2 + 8(9{,}38 - 13)^2 + 6(20{,}83 - 13)^2\right) \approx 9{,}281.$$

При справедливости гипотезы H_0 статистика (8.3) критерия Краскела–Уоллиса имеет распределение хи-квадрат $\mathcal{H}_r$ с $r = k-1 = 2$ степенями свободы.

Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид $(k_{0,95}(2); +\infty) = (5{,}99; +\infty)$. Реализация статистики попадает в критическую область, следовательно, гипотеза H_0 отвергается на уровне значимости 0,05. Таким образом, нельзя считать, что средние значения показателей среднедушевых месячных доходов населения одинаковы во всех трех округах.

Если предположить, что рассматриваемые выборки соответствуют гауссовскому распределению (обоснования такого предположения приведены в примере 7.6), то имеющиеся данные можно описать моделью (8.8) и проверить гипотезу H_0 об однородности вида (8.9) против альтернативы H_A вида (8.10).

Для проверки гипотезы (8.9) применим F-критерий со статистикой (8.11):

$$T(\mathbb{Z}_N) = F = \frac{\dfrac{1}{k-1}\sum\limits_{j=1}^{k} n_j(\overline{X}_{\cdot j} - \overline{X}_N)^2}{\dfrac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2}.$$

Вычислим сначала реализации выборочных средних $\overline{X}_{\cdot 1}$, $\overline{X}_{\cdot 2}$, $\overline{X}_{\cdot 3}$ для выборок Z_1, Z_2, Z_3 и выборочного среднего $\overline{X}_N$ объединенной выборки $\mathbb{Z}_N$.

Имеем $\overline{X}_{\cdot 1} = 11\,101{,}0$, $\overline{X}_{\cdot 2} = 10\,803{,}9$, $\overline{X}_{\cdot 3} = 20\,340{,}0$, $\overline{X}_N = 13\,223{,}3$.

Тогда

$$\frac{1}{k-1}SS_b = \frac{1}{2}\sum_{j=1}^{3} n_j(\overline{X}_{\cdot j} - \overline{X}_N)^2 \approx 11(11\,101 - 13\,223{,}3)^2 + 8(10\,803{,}9 - 13\,223{,}3)^2 + 6(20\,340 - 13\,223{,}3)^2 \approx 200\,129\,591,$$

$$\frac{1}{N-k}SS_w = \frac{1}{22}\sum_{j=1}^{3}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2 \approx 19\,038\,613$$

и

$$T(z_N) = \frac{200\,129\,591}{19\,038\,613} \approx 10{,}51.$$

При справедливости гипотезы H_0 вида (8.9) статистика $T(\mathbb{Z}_N)$ имеет F-распределение $F(k-1; N-k)$.

Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид $(f_{0,95}(2; 22); +\infty)$, где $f_{0,95}(2; 22)$ — квантиль уровня 0,95 распределения $F(2; 22)$. По таблице [7] находим $f_{0,95}(2; 22) = 3{,}44$.

Таким образом, реализация статистики F-критерия попадает в критическую область и гипотеза об однородности H_0 вида (8.9) отвергается на уровне значимости 0,05. ■

Пример 8.4. Пусть значения доходов X_{ij} из примера 8.3 описываются моделью

$$X_{ij} = \theta_j + \varepsilon_{ij}, \quad j = 1, 2, 3, i = 1, \ldots, n_j,$$

где ε_{ij} соответствуют гауссовскому распределению с нулевым математическим ожиданием и неизвестной дисперсией σ^2, а $\theta_j \in \mathbb{R}^1$, $j = 1, 2, 3$ — неизвестные параметры. Постройте доверительные интервалы уровня надежности 0,95 для контрастов γ_1, γ_2, γ_3 параметров θ_j, где

$$\gamma_1 = \sum_{j=1}^{3} c_{1j}\theta_j = \theta_1 - \theta_2, \text{ т.е. } c_{11} = 1, \ c_{12} = -1, \ c_{13} = 0;$$

$$\gamma_2 = \sum_{j=1}^{3} c_{2j}\theta_j = \theta_1 - \theta_3, \text{ т.е. } c_{21} = 1, \ c_{22} = 0, \ c_{23} = -1;$$

$$\gamma_3 = \sum_{j=1}^{3} c_{3j}\theta_j = \theta_2 - \theta_3, \text{ т.е. } c_{31} = 0, \ c_{32} = 1, \ c_{33} = -1.$$

Дайте содержательную трактовку контрастов γ_1, γ_2, γ_3.

Решение. Точечной оценкой параметра $\gamma_1 = \theta_1 - \theta_2$ будет, согласно (8.15),

$$\widehat{\gamma}_1 = \sum_{j=1}^{3} c_{1j}\overline{X}_{\cdot j} = \overline{X}_{\cdot 1} - \overline{X}_{\cdot 2},$$

а доверительный интервал (8.17) надежности 0,95

$$I_1(Z_N) = \left(\widehat{\gamma}_1 - t_{0,975}(N-k)\sqrt{\widetilde{S}_N^2 \sum_{j=1}^{3} \frac{c_j^2}{n_j}}\ ;\right.$$

$$\left.\widehat{\gamma}_1 + t_{0,975}(N-k)\sqrt{\widetilde{S}_N^2 \sum_{j=1}^{3} \frac{c_j^2}{n_j}}\right),$$

где $\widetilde{S}_N^2 = \dfrac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2$. Вычислим соответствующие реализации оценки контраста и доверительного интервала $I_1(Z_N)$.

Поскольку (см. пример 8.3) $\overline{X}_{\cdot 1} = 11\,101{,}0$, $\overline{X}_{\cdot 2} = 10\,803{,}9$, $\overline{X}_{\cdot 3} = 20\,340{,}0$, то

$$\widehat{\gamma}_1 = 11\,101 - 10\,803{,}9 = 297{,}1,$$

$$\widetilde{S}^2_N = \frac{1}{N-3}\sum_{j=1}^{3}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2 = 19\,038\,613,$$

$$\sqrt{\sum_{j=1}^{3}\frac{c_j^2}{n_j}} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{1}{11} + \frac{1}{8}} = 0{,}46.$$

По табл. 22.4 находим квантиль $t_{0,975}(22) = 2{,}074$ и получаем реализацию доверительного интервала

$$\begin{aligned} I_1(z_{25}) &= (297{,}1 - 2{,}074\cdot 2027{,}5;\ 297{,}1 + 2{,}074\cdot 2027{,}5) = \\ &= (-3907{,}6;\ 4501{,}8)\,. \end{aligned}$$

Аналогично,

$$\widehat{\gamma}_2 = \sum_{j=1}^{3} c_{2j}\overline{X}_{\cdot j} = \overline{X}_{\cdot 1} - \overline{X}_{\cdot 3};\ \ \widehat{\gamma}_3 = \sum_{j=1}^{3} c_{1j}\overline{X}_{\cdot j} = \overline{X}_{\cdot 2} - \overline{X}_{\cdot 3};$$

$$I_2(Z_N) = \left(\widehat{\gamma}_2 - t_{0,975}(N-k)\sqrt{\frac{1}{n_1} + \frac{1}{n_3}}\sqrt{\widetilde{S}^2_N}\ ;\right.$$
$$\left.\widehat{\gamma}_2 + t_{0,975}(N-k)\sqrt{\frac{1}{n_1} + \frac{1}{n_3}}\sqrt{\widetilde{S}^2_N}\right);$$

$$I_3(Z_N) = \left(\widehat{\gamma}_3 - t_{0,975}(N-k)\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}\sqrt{\widetilde{S}^2_N}\ ;\right.$$
$$\left.\widehat{\gamma}_3 + t_{0,975}(N-k)\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}\sqrt{\widetilde{S}^2_N}\right).$$

Вычисляя соответствующие реализации, получим:

$$\widehat{\gamma}_2 = 11\,101 - 20\,340 = -9239;\ \widehat{\gamma}_3 = 10\,803{,}9 - 20\,340 = -9536{,}1;$$

$$I_2(z_{25}) = (-13\,831{,}5;\ -4646{,}5)\,;\ \ I_3(z_{25}) = (-14\,423{,}1;\ -4649{,}1)\,.$$

Контраст γ_1 представляет разность средних значений СВ, порождающих выборки Z_1 и Z_2, контраст γ_2 — разность средних значений

СВ, порождающих выборки Z_1 и Z_3, контраст γ_3 — разность средних значений СВ, порождающих выборки Z_2 и Z_3.

Важно отметить, что доверительный интервал для γ_1 включает значение ноль. Это означает, что на уровне доверия 0,95 можно считать, что разность параметров $\theta_1 - \theta_2 = \gamma_1$ равна нулю, и гипотеза об однородности выборок Z_1 и Z_2 в данной модели верна. Доверительные интервалы γ_2 и γ_3 не включают значение ноль, это означает, что средние значения θ_1 и θ_3, а также θ_2 и θ_3 различаются. ■

Пример 8.5. После разрыва ахиллова сухожилия травмированному человеку необходимо сделать операцию и последующую иммобилизацию в течении шести недель. Однако восстановление двигательных функций травмированной ноги требует длительного времени. Для сокращения восстановительного периода необходимо, по мнению врачей, пройти реабилитационный курс, включающий физиотерапевтическое лечение и занятие лечебной гимнастикой. Не все пациенты имеют силы и возможности пройти такой курс.

В табл. 8.5 представлены данные о времени (в неделях) восстановительного периода для трех групп успешно прооперированных пациентов примерно одинакового возраста и состояния здоровья. Пациенты первой группы прошли полный реабилитационный курс, пациенты второй группы получили только физиотерапевтическое лечение, а пациенты третьей группы целенаправленно не занимались реабилитацией.

Таблица 8.5

Группа 1	Группа 2	Группа 3
26	41	36
29	34	44
19	44	47
37	23	41
28	28	37
33	45	49
40	33	42
36	35	44
34	40	
31	54	

Можно ли считать, что указанные реабилитационные процедуры способствуют сокращению времени восстановления пациентов после травмы?

Решение. Фактором в данной задаче является наличие реабилитационного лечения. Уровни фактора: 1 — наличие полного реабилитационного курса лечения, 2 — частичное реабилитационное лечение, 3 — отсутствие реабилитационного лечения. Данные представлены тремя выборками $Z_j = \left[X_{1j}, \ldots, X_{n_j j}\right]^\top$, $j = 1, 2, 3$ объемов $n_1 = 10$,

$n_2 = 10$ и $n_3 = 8$, которые соответствуют неизвестным непрерывным распределениям $F(t - \theta_j)$.

Гипотеза H_0: $\theta_1 = \theta_2 = \theta_3$ означает, что выборки Z_1, Z_2 и Z_3 однородны, т.е. реабилитационное лечение не оказывает влияния на срок восстановления после травмы.

В качестве альтернативной гипотезы H_A можно выбрать и гипотезу общего вида H_1: $\exists\, \theta_i \neq \theta_j$ при $i \neq j$, и упорядоченную альтернативу H_2: $\theta_1 \leqslant \theta_2 \leqslant \theta_3$, где хотя бы одно из неравенств строгое. Последняя альтернатива описывает ситуацию, когда более полное реабилитационное лечение обуславливает более быстрое восстановление.

Для проверки гипотезы H_0 против альтернативы H_1 можно использовать критерий Краскела—Уоллиса, для проверки H_0 против H_2 — критерий Джонкхиера.

Чтобы вычислить статистику критерия Краскела—Уоллиса (8.3)

$$T(\mathbb{Z}_N) = H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{R}_{\cdot j} - \frac{N+1}{2}\right)^2,$$

составим таблицу (табл. 8.6) реализаций рангов r_{ij} СВ X_{ij}, $j = 1, 2, 3$, $i = 1, \ldots, n_j$.

Таблица 8.6

Группа 1	Группа 2	Группа 3
3	19,5	13,5
6	10,5	23
1	23	26
15,5	2	19,5
4,5	4,5	15,5
8,5	25	27
17,5	8,5	21
13,5	12	23
10,5	17,5	
7	28	

Следовательно, $\overline{R}_{\cdot 1} = 8{,}7$, $\overline{R}_{\cdot 2} = 15{,}05$, $\overline{R}_{\cdot 3} = 21{,}06$.

Тогда реализация статистики

$$H = \frac{12}{28 \cdot 29}[10(8{,}7 - 14{,}5)^2 + 10(15{,}05 - 14{,}5)^2 + 8(21{,}06 - 14{,}5)^2] \approx 10{,}1.$$

Поскольку в выборке $\mathbb{Z}_N$ имеются связки, то следует использовать модифицированную форму (8.5) статистики H. Так как в выборке есть 8 связок, из которых 7 связок имеют размер 2, а одна — размер 3, то

$$H' = \frac{H}{1 - \frac{1}{N^3 - N} \sum\limits_{i=1}^{g} ({t^3}_i - t_i)} = \frac{10{,}1}{1 - \frac{1}{28^3 - 28}\,(7(2^3 - 2) + (3^3 - 3))} = 10{,}13.$$

При справедливости гипотезы H_0 статистика $T(\mathbb{Z}_N)$ имеет при $N \to \infty$ распределение хи-квадрат $\mathcal{H}_r$ с $r = k - 1 = 2$ степенями свободы. Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид $(k_{0,95}(2); +\infty)$, где $k_{0,95}(2)$ — квантиль уровня 0,95 распределения $\mathcal{H}_2$.

По табл. 22.3 находим $k_{0,95}(2) = 5{,}99$.

Таким образом, реализация статистики попадает в критическую область, и гипотеза H_0 отвергается в пользу альтернативы H_1 на уровне значимости 0,05.

Применим теперь критерий Джонкхиера, статистика которого (8.7) имеет вид

$$T(\mathbb{Z}_N) = J = U_{12} + U_{13} + U_{23},$$

где $U_{lm} = \sum\limits_{i=1}^{n_l} \sum\limits_{j=1}^{n_m} \varphi\left(X_{il}, X_{jm}\right)$.

Для вычисления реализации величины $U_{12} = \sum\limits_{i=1}^{n_1} \sum\limits_{j=1}^{n_2} \varphi\left(X_{i1}, X_{j2}\right)$ необходимо сложить $n_1 \cdot n_2 = 10 \cdot 10$ значений $\varphi\left(x_{i1}, x_{j2}\right)$.

Так, сравнивая каждое значение первого столбца с каждым значением второго столбца получим, $\varphi\left(x_{11}, x_{12}\right) = 1$, поскольку $x_{11} = 26 <$ $< x_{12} = 41$; $\varphi\left(x_{11}, x_{22}\right) = 1$, поскольку $x_{11} = 26 < x_{22} = 34$ и т.д. Таким образом, при $i = 1$ сумма $\sum\limits_{j=1}^{10} \varphi\left(x_{11}, x_{j2}\right) = 9$, при $i = 2$ сумма $\sum\limits_{j=1}^{10} \varphi\left(x_{21}, x_{j2}\right) = 8$, и т.д. для $i = 3, \ldots, 10$.

В итоге получаем

$$u_{12} = 9 + 8 + 10 + 5 + 8{,}5 + 7{,}5 + 4{,}5 + 5 + 6{,}5 + 8 = 72,$$

$$u_{13} = 8 + 8 + 8 + 7 + 8 + 8 + 6 + 7 + 8 + 8 = 76,$$

$$u_{23} = 5 + 8 + 3 + 8 + 8 + 2 + 8 + 8 + 6 + 0 = 56.$$

Тогда реализация статистики J

$$J = T(z_N) = 72 + 76 + 56 = 204.$$

Найдем математическое ожидание и дисперсию статистики J при справедливости гипотезы H_0

$$\mathbf{M}\{J\} = \frac{1}{4}\left(N^2 - \sum_{j=1}^{k} n_j^2\right) = \frac{1}{4}(28^2 - (100 + 100 + 64)) = 130,$$

$$\mathbf{D}\{J\} = \frac{1}{72}\left(N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right) = \frac{1}{72}(28^2(2\cdot 28+3) -$$
$$-(100\cdot 23 + 100\cdot 23 + 64\cdot 19)) = 578{,}56.$$

Тогда реализация стандартизованной статистики J^*

$$J^* = T(z_N) = \frac{J - \mathbf{M}\{J\}}{\sqrt{\mathbf{D}\{J\}}} = 3{,}07.$$

Статистика J^* при справедливости гипотезы H_0 асимптотически нормальна. Критическая область критерия уровня значимости $\alpha = 0{,}05$ имеет вид $(u_{1-\alpha}; +\infty) = (u_{0{,}95}; +\infty) = (1{,}65; +\infty)$. Таким образом, реализация статистики J^* попадает в критическую область, и гипотеза H_0 отвергается на уровне значимости 0,05 в пользу альтернативы H_2.

Важно отметить, что полученная реализация статистики H', равная 10,13, совпадает с квантилью распределения $\mathcal{H}_2$ уровня $1-\alpha \approx \approx 0{,}99$, а реализация статистики J^*, равная 3,07, совпадает с квантилью распределения $\mathcal{N}(0;1)$ уровня $1-\alpha \approx 0{,}999$. ■

8.7. Задачи для самостоятельного решения

1. Проведено исследование по оценке роли чистой мотивации (знания цели работы) на выполнение скучных монотонных операций при вытачивании металлической заготовки. Восемнадцать рабочих одинаковой квалификации были случайным образом разделены на три группы. Группа A не имела информации о требуемой производительности труда, группа B получила лишь общие сведения, группа C имела точную информацию о задании. В табл. 8.7 приведено количество обработанных заготовок. Можно ли на уровне значимости 0,05 считать, что производительность труда растет с осведомленностью?

Таблица 8.7

A	40	35	38	43	44	41
B	38	40	47	44	40	42
C	48	40	45	43	46	44

2. Докажите, что статистика (8.15) $\widehat{\gamma}$ является несмещенной оценкой параметра γ из определения 8.1.

3. Найдите распределение статистики (8.16) и докажите, что статистика (8.16) является центральной статистикой для контраста параметров γ из определения 8.1.

4. Имеются данные о величине прожиточного минимума трудоспособного населения (на душу населения руб. в месяц), установленной в субъектах РФ за IV квартал 2008 г. (табл. 8.8).

Таблица 8.8

Центральный федеральный округ	Сибирский федеральный округ	Дальневосточный федеральный округ
Белгородская область 4143	Республика Бурятия 5039	Чукотский автономный округ 10 061
Брянская область 4292	Республика Тыва 5005	Камчатский край 10 146
Ивановская область 4623	Кемеровская область 4407	Приморский край 6525
Курская область 4424	Красноярский край 5625	Амурская область 6069
Московская область 5786	Омская область 4836	Сахалинская область 7972
Смоленская область 4838	Республика Алтай 5907	Республика Саха 8128
Рязанская область 4772		
г. Москва 7510		

Можно ли считать, что величина прожиточного минимума во всех представленных федеральных округах в среднем одинакова? Уровень значимости выберите равным 0,05.

5. Пусть величины прожиточного минимума трудоспособного населения X_{ij}, где $j = 1, 2, 3$ — номер федерального округа, а $i = 1, \ldots, n_j$ — номер региона в соответствующем федеральном округе, представленные в задаче 8.3, описываются моделью (8.12), в которой ε_{ij} имеют распределение $\mathcal{N}(0; \sigma^2)$. Постройте точечные оценки и доверительные интервалы надежности 0,95 для контрастов $\gamma_1 = \theta_1 - \theta_2$, $\gamma_2 = \theta_1 - \theta_3$ и $\gamma_3 = \theta_2 - \theta_3$ в модели (8.12).

Ответ: $\widehat{\gamma}_1 = -88$, $I_1 = (-1464{,}4, 1288{,}4)$; $\widehat{\gamma}_2 = -3101{,}7$, $I_2 = (-4478{,}1, -1725{,}2)$; $\widehat{\gamma}_3 = -3013{,}7$, $I_3 = (-4485{,}1, -1542{,}2)$.

6. Три группы случайно отобранных людей обучались навыкам скорочтения тремя разными методами. В конце обучения проводился зачет, на котором оценивалась скорость чтения. Обучающиеся показали следующие результаты (страниц за десять минут). Первая группа: 20, 23, 24, 24, 25, 26, 28, 30, 31, 32. Вторая группа: 38, 42, 42, 44, 47, 48, 49, 50, 51, 52. Третья группа: 29, 32, 33, 35, 35, 37, 38, 39, 40, 42. Можно ли считать на уровне значимости 0,05, что предлагаемые методы обучения имеют различную эффективность?

7. Время (в сек.) химической реакции при различном содержании катализатора распределилось следующим образом (табл. 8.9)

Таблица 8.9

Содержание катализатора, %	Номер эксперимента					
	1	2	3	4	5	6
5	8,2	6,8	8,0	7,5	7,0	7,2
10	5,0	6,1	7,0	6,3	5,5	
15	4,9	5,0	6,2	5,5	4,5	6,0

Можно ли считать на уровне значимости 0,01, что увеличение содержания катализатора в среднем уменьшает время химической реакции?

8. Докажите эквивалентность формул (8.3) и (8.4).

9. Проверка гипотезы о независимости случайных величин

Большое количество задач в экономике, социологии, биологии, технике связано с исследованием зависимости между двумя (или несколькими) показателями или признаками, которыми характеризуется объект. Например, связаны ли показатель ВВП и темп прироста населения; доходы и уровень образования, пол, возраст человека; коэффициент интеллекта и калорийность питания; уровень холестерина в крови и степень физической активности человека; урожайность пшеницы и количество осадков; участие в благотворительной деятельности и материальное положение. Понятно, что выявленная зависимость (или независимость) изучаемых признаков должна привести исследователя к определенным практическим выводам. В этом разделе будет представлено несколько статистических критериев проверки независимости СВ. Выбор того или иного критерия обусловлен шкалой, в которой производится измерение признаков. Так, если показатели измеряются в номинальной шкале, т.е. представляются наименованиями категорий, которые не могут быть упорядочены, то для проверки независимости применяют критерий хи-квадрат. Если установлено наличие статистической связи между номинальными признаками, то силу этой связи можно характеризовать коэффициентом взаимной сопряженности Пирсона, коэффициентом Крамера, мерами прогноза Гутмана. Если признаки измерены в порядковой шкале (т.е. к данным применимо только сравнение типа «хуже-лучше»), то для проверки независимости применяют ранговые критерии Спирмена и Кендалла, а коэффициенты ранговой корреляции Спирмена и согласованности Кендалла служат измерителями силы связи между двумя порядковыми переменными.

Для выявления независимости показателей, измеряемых в количественной шкале, можно использовать критерий хи-квадрат состоятельный против любых альтернатив о зависимости. К сожалению, применение этого универсального критерия может быть сопряжено с некоторыми техническими сложностями. Поэтому удобнее бывает проверить гипотезу о некоррелированности признаков с помощью критерия, основанного на выборочном коэффициенте корреляции. Если гипотеза о некоррелированности отвергнута, то коэффициент выборочной корреляции характеризует силу связи между признаками. В случае же принятия гипотезы о некоррелированности следует провести дополнительное исследование, обратившись, например, к критерию хи-квадрат. Здесь важно отметить, что для показателей, имеющих двумерное гауссовское распределение, некоррелированность показателей эквивалентна их независимости. Поэтому в гауссовском случае критерий, основанный на выборочном коэффициенте корреляции, позволяет полностью решить проблему исследования зависимости между наблюдаемыми показателями. Для проверки независимости количественных показателей можно также использовать ранговый критерий Спирмена или критерий Кендалла. Эти критерии обладают рядом достоинств, однако их существенный недостаток состоит в том, что они способны уловить наличие только монотонной связи между признаками. Для исследования зависимости между несколькими показателями можно использовать критерий, основанный на множественном коэффициенте корреляции, и критерий, основанный на коэффициенте конкордации Кендалла.

9.1. Теоретические положения

Пусть выборки $\mathbb{X}_n = [X_1, \ldots, X_n]^\top$ и $\mathbb{Y}_n = [Y_1, \ldots, Y_n]^\top$ порождены СВ X и Y соответственно. Предполагается, что $\mathbb{X}_n$ и $\mathbb{Y}_n$ получены в процессе совместного наблюдения за X и Y, а именно: если x_k — реализация СВ X_k, $k = 1, \ldots, n$ (т.е. реализация СВ X в k-м опыте), то y_k — реализация СВ Y в этом же опыте. Обозначим $\mathbb{W}_n = [W_1, \ldots, W_n]^\top$ — двумерную выборку с элементами $W_k = (X_k, Y_k)$, $k = 1, \ldots, n$. Обозначим через $F_X(x)$ и $F_Y(y)$ функции распределения СВ X и Y соответственно, а через $F_W(x, y)$ — функцию распределения случайного вектора $W = [X, Y]^\top$, компонентами которого являются изучаемые СВ X и Y.

Определение 9.1. Статистическая гипотеза вида

$$H_0: \ F_W(x, y) = F_X(x) F_Y(y), \ \ \forall x, y \in \mathbb{R}^1 \tag{9.1}$$

называется гипотезой о независимости СВ X и Y.

Рассмотрим критерии проверки гипотезы H_0 вида (9.1) при различных предположениях о законах распределения $F_W(x, y)$, $F_X(x)$, $F_Y(y)$.

9.2. Критерий, основанный на выборочном коэффициенте корреляции

Пусть справедливо предположение о том, что случайный вектор $W = [X, Y]^\top$ — гауссовский, причем $\mathbf{D}\{X\} > 0$ и $\mathbf{D}\{Y\} > 0$.

Тогда гипотеза H_0 вида (9.1) о независимости СВ X и Y эквивалентна гипотезе

$$H_0: \ r_{XY} = 0, \tag{9.2}$$

где $r_{XY} = \dfrac{k_{XY}}{\sqrt{\mathbf{D}\{X\}\,\mathbf{D}\{Y\}}}$ — коэффициент корреляции СВ X и Y, а k_{XY} — их ковариация.

Эквивалентность (9.1) и (9.2) следует из того, что компоненты гауссовского вектора X и Y независимы тогда и только тогда, когда X и Y некоррелированы (см. свойство 2 разд. 21.5).

Оценкой неизвестного коэффициента корреляции r_{XY} является *выборочный коэффициент корреляции*

$$\widehat{r}_{XY}(n) = \frac{\widehat{k}_{XY}(n)}{\overline{S}_X \overline{S}_Y}, \tag{9.3}$$

где $\overline{S}^2_X$, $\overline{S}^2_Y$ — выборочные дисперсии, построенные по выборкам $\mathbb{X}_n$ и $\mathbb{Y}_n$ соответственно, а $\widehat{k}_{XY}(n)$ — выборочная ковариация (см. определение 1.10), построенная по двумерной выборке $\mathbb{W}_n$.

Можно показать [33], что при выполнении сделанного предположения и справедливости гипотезы H_0 вида (9.2), статистика

$$T(\mathbb{W}_n) = \frac{\sqrt{n-2}\,\widehat{r}_{XY}(n)}{\sqrt{1-\widehat{r}^{\,2}_{XY}(n)}} \tag{9.4}$$

имеет распределение Стьюдента $\mathcal{T}_l$ с $l = n-2$ степенями свободы.

При $n \to \infty$ и справедливости H_0 статистика вида

$$\widetilde{T}(\mathbb{W}_n) = \sqrt{n}\,\widehat{r}_{XY}(n) \tag{9.5}$$

асимптотически нормальна.

Критические области уровня значимости α для критериев, основанных на статистиках (9.4) и (9.5), приведены в табл. 9.1, где $t_\gamma(l)$, u_γ — квантили уровня γ распределений $\mathcal{T}_l$ и $\mathcal{N}(0;1)$ соответственно.

Таблица 9.1

H_A	Критические области для $T(\mathbb{W}_n)$	Критические области для $\widetilde{T}(\mathbb{W}_n)$
$r_{XY} < 0$	$(-\infty; t_\alpha(n-2))$	$(-\infty; u_\alpha)$
$r_{XY} > 0$	$(t_{1-\alpha}(n-2); +\infty)$	$(u_{1-\alpha}; +\infty)$
$r_{XY} \neq 0$	$\left(-\infty; t_{\frac{\alpha}{2}}(n-2)\right) \cup$ $\cup \left(t_{1-\frac{\alpha}{2}}(n-2); +\infty\right)$	$\left(-\infty; u_{\frac{\alpha}{2}}\right) \cup \left(u_{1-\frac{\alpha}{2}}; +\infty\right)$

Важно отметить, что если вектор $W = (X,Y)^\top$ — гауссовский, то утверждение о том, что СВ X и Y зависимы справедливо тогда и только тогда, когда $r_{XY} \neq 0$. Таким образом, в рассматриваемом случае альтернативная гипотеза общего вида

$$H_A:\ \exists\, x, y \in \mathbb{R}^1 \text{ такие, что } F_W(x,y) \neq F_X(x) F_Y(y) \tag{9.6}$$

эквивалентна гипотезе

$$H_1:\ r_{XY} \neq 0.$$

Критерии, основанные на статистиках (9.4) и (9.5), в гауссовском случае являются состоятельными против альтернативы (9.6). Если же распределение вектора W отлично от гауссовского, то эти критерии состоятельны против альтернатив вида H_1: $r_{XY} \neq 0$, H_2: $r_{XY} < 0$, H_3: $r_{XY} > 0$, означающих коррелированность СВ X и Y.

Если гипотеза H_0 отвергнута, т.е. СВ X и Y зависимы, то коэффициент корреляции r_{XY} может служить характеристикой силы связи между X и Y. Чтобы построить доверительный интервал для r_{XY}, необходимо знать распределение $\widehat{r}_{XY}(n)$ при справедливости гипотезы H_1: $r_{XY} \neq 0$. Известно [21], что в этом случае

$$\mathbf{M}\{\widehat{r}_{XY}(n)\} = r_{XY}\left(1 - \frac{1 - r_{XY}^2}{2n}\right) + O(n^2),$$

$$\mathbf{D}\{\widehat{r}_{XY}(n)\} = \frac{(1 - r_{XY}^2)^2}{n} + O(n^2),$$

и при $n \to \infty$ статистика $\dfrac{\widehat{r}_{XY}(n) - r_{XY}\left(1 - \dfrac{1 - r_{XY}^2}{2n}\right)}{(1 - r_{XY}^2)/\sqrt{n}}$ асимптотически нормальна.

Тот факт, что дисперсия статистики $\widehat{r}_{XY}(n)$ явно зависит от неизвестного параметра r_{XY}, снижает точность асимптотического доверительного интервала. Более того, если абсолютное значение r_{XY} близко к единице, то одна из границ асимптотического доверительного интервала может оказаться меньше -1 или больше 1.

В связи с этим Фишером [36] было построено *z-преобразование* коэффициента выборочной корреляции

$$\widehat{z} = \operatorname{arcth} \widehat{r}_{XY}(n) = \frac{1}{2} \ln\left(\frac{1 + \widehat{r}_{XY}(n)}{1 - \widehat{r}_{XY}(n)}\right),$$

распределение которого сходится по распределению к нормальному быстрее, чем распределение самой статистики $\widehat{r}_{XY}(n)$, а дисперсия $\widehat{z}$ не зависит от r_{XY}.

Доказано, что

$$\mathbf{M}\{\widehat{z}\} = \frac{1}{2} \ln \frac{1 + r_{XY}}{1 - r_{XY}} + \frac{r_{XY}}{2(n-1)} + O(n^2), \quad \mathbf{D}\{\widehat{z}\} = \frac{1}{n-3} + O(n^2).$$

Тогда доверительный интервал параметра $z = \operatorname{arcth} r_{XY}$ уровня надежности $1-p$ имеет вид

$$I_1 = \left[\operatorname{arcth}\widehat{r}_{XY}(n) - \frac{\widehat{r}_{XY}(n)}{2(n-1)} - \frac{u_{1-\frac{p}{2}}}{\sqrt{n-3}};\right.$$

$$\left.\operatorname{arcth}\widehat{r}_{XY}(n) - \frac{\widehat{r}_{XY}(n)}{2(n-1)} + \frac{u_{1-\frac{p}{2}}}{\sqrt{n-3}}\right] = [z_1; z_2],$$

а искомый доверительный интервал для коэффициента корреляции r_{XY} уровня надежности $1-p$ вид

$$I_2 = [\operatorname{th} z_1;\ \operatorname{th} z_2],$$

где $u_{1-\frac{p}{2}}$ — квантиль распределения $\mathcal{N}(0;1)$ уровня $1-\frac{p}{2}$, $\operatorname{th} x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ — тангенс гиперболический величины x.

Отметим, что слагаемым $\frac{r_{XY}}{2(n-1)}$ в формуле для $\mathbf{M}\{\widehat{z}\}$ можно, вообще говоря, пренебрегать, так как оно имеет больший порядок малости по сравнению с $\sqrt{\mathbf{D}\{\widehat{z}\}} = \frac{1}{\sqrt{n-3}}$.

9.3. Критерий Кендалла

При нарушении предположения о гауссовости вектора W критерий, основанный на выборочном коэффициенте корреляции, может оказаться ненадежным, так как является состоятельным только для альтернатив вида H_1, H_2 или H_3, или даже неприменимым (если, например, СВ X и Y не имеют конечных дисперсий). Кроме того, возможны ситуации, при которых наблюдаемые величины не имеют количественного выражения, и экспериментальные данные могут быть представлены лишь в виде ранжировок. В этом случае можно использовать ранговый критерий Спирмена или ранговый критерий Кендалла.

Пусть справедливо предположение о том, что функции распределения $F_X(x)$, $F_Y(y)$ и $F_W(x, y)$ непрерывны.

Определение 9.2. Назовем *параметром согласованности* СВ X и Y, являющихся компонентами вектора W, величину

$$\tau_{XY} = 1 - 2\mathbf{P}\{(X_2 - X_1)(Y_2 - Y_1) < 0\},$$

где (X_1, Y_1) и (X_2, Y_2) — независимые двумерные СВ, имеющие распределение $F_W(x, y)$.

Параметр τ_{XY}, согласно определению 9.2, может принимать значения от -1 до 1, причем значения ± 1 достигаются в том случае, когда

$Y = \varphi(X)$, где $\varphi(\cdot)$ — строго монотонная функция. Действительно, если $\varphi(\cdot)$ — строго возрастающая функция, то

$$\mathbf{P}\{(X_2 - X_1)(Y_2 - Y_1) < 0\} = \mathbf{P}\{(X_2 - X_1)(\varphi(X_2) - \varphi(X_1)) < 0\} = 0,$$

т.е. $\tau_{XY} = 1$.

Если $\varphi(\cdot)$ — монотонно убывающая функция, то

$$\mathbf{P}\{(X_2 - X_1)(\varphi(X_2) - \varphi(X_1)) < 0\} = 1,$$

т.е. $\tau_{XY} = -1$.

Понятно, что в случае независимости СВ X и Y (т.е. верна гипотеза H_0 вида (9.1)) параметр τ_{XY} равен нулю. Однако существуют ситуации (см. пример 9.3), когда СВ X и Y зависимы, а $\tau_{XY} = 0$.

Для того чтобы построить оценку параметра согласованности τ_{XY}, введем следующие обозначения и определения.

Пусть (X_i, Y_i) и (X_j, Y_j) — элементы выборки $\mathbb{W}_n$, $1 \leqslant i, j \leqslant n$.

Определение 9.3. Пары (X_i, Y_i) и (X_j, Y_j) называются *согласованными*, если $\operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\} = 1$, и *несогласованными*, если $\operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\} = -1$.

Обозначим через K случайное число несогласованных пар среди всех $C_n^2 = \dfrac{n(n-1)}{2}$ пар выборки $\mathbb{W}_n$, а через Q — число согласованных пар выборки $\mathbb{W}_n$.

Статистика

$$\widehat{\tau}_{XY}(n) = 1 - \frac{4K}{n(n-1)} \tag{9.7}$$

называется *коэффициентом согласованности Кендалла.*

Нетрудно показать, что возможны и другие формы записи коэффициента согласованности Кендалла:

$$\begin{aligned} \widehat{\tau}_{XY}(n) = \frac{2(Q-K)}{n(n-1)} &= \frac{2 \sum\limits_{1 \leqslant i < j \leqslant n} \operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\}}{n(n-1)} = \\ &= \frac{2 \sum\limits_{1 \leqslant i < j \leqslant n} \operatorname{sign}\{(R_i - R_j)(S_i - S_j)\}}{n(n-1)}, \end{aligned} \tag{9.8}$$

где R_i, R_j — ранги элементов X_i, X_j в выборке $\mathbb{X}_n$, а S_i, S_j — ранги Y_i, Y_j в выборке $\mathbb{Y}_n$.

Можно показать, что

$$\mathbf{M}\{\widehat{\tau}_{XY}(n)\} = \tau_{XY}, \tag{9.9}$$

т.е. коэффициент $\widehat{\tau}_{XY}(n)$ является несмещенной оценкой параметра τ_{XY}.

Если в выборках $\mathbb{X}_n$ и $\mathbb{Y}_n$ имеются связки, то при вычислении коэффициента $\widehat{\tau}_{XY}(n)$ следует внести поправку

$$\widehat{\tau}_{XY}(n) = \frac{\sum\limits_{1 \leqslant i < j \leqslant n} \operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\}}{\sqrt{\frac{1}{2}n(n-1) - u_1}\sqrt{\frac{1}{2}n(n-1) - u_2}}, \tag{9.10}$$

где $u_1 = \frac{1}{2}\sum\limits_{k=1}^{q} u_{1k}(u_{1k} - 1)$, $u_2 = \frac{1}{2}\sum\limits_{k=1}^{g} u_{2k}(u_{2k} - 1)$, q — количество связок в выборке $\mathbb{X}_n$, u_{1k} — размер k-й связки выборки $\mathbb{X}_n$, g — количество связок в выборке $\mathbb{Y}_n$, u_{2k} — размер k-й связки выборки $\mathbb{Y}_n$.

Кендалл доказал [18], что при справедливости гипотезы H_0 вида (9.1)

$$\mathbf{M}\{\widehat{\tau}_{XY}(n)\} = 0, \quad \mathbf{D}\{\widehat{\tau}_{XY}(n)\} = \frac{4n+10}{9n(n-1)},$$

а нормированный коэффициент согласованности

$$T_\tau(\mathbb{W}_n) = \frac{\widehat{\tau}_{XY}(n)}{\sqrt{\mathbf{D}\{\widehat{\tau}_{XY}(n)\}}} \approx \frac{3\sqrt{n}\,\widehat{\tau}_{XY}(n)}{2} \tag{9.11}$$

асимптотически нормален.

Квантили распределения статистики $\widehat{\tau}_{XY}(n)$ при справедливости H_0 вида (9.1) для $4 \leqslant n \leqslant 40$ табулированы в [39].

Критические области уровня значимости α критерия Кендалла, основанного на статистике (9.11), соответствующие различным альтернативам H_A, приведены в табл. 9.2.

Таблица 9.2

H_A	Критические области для $T_\tau(\mathbb{W}_n)$
$\tau_{XY} < 0$	$(-\infty; u_\alpha)$
$\tau_{XY} > 0$	$(u_{1-\alpha}; +\infty)$
$\tau_{XY} \neq 0$	$\left(-\infty; u_{\frac{\alpha}{2}}\right) \cup \left(u_{1-\frac{\alpha}{2}}; +\infty\right)$

Критерий Кендалла является состоятельным только для альтернатив H_A: $\tau_{XY} < 0$, H_A: $\tau_{XY} > 0$ и H_A: $\tau_{XY} \neq 0$. Важно отметить, что известно асимптотическое распределение коэффициента согласованности Кендалла $\widehat{\tau}_{XY}(n)$ и в том случае, когда гипотеза H_0 вида (9.1) неверна. Это обстоятельство позволяет построить асимптотический доверительный интервал для параметра согласованности τ_{XY}.

Известна также АОЭ $e = e(\widehat{\tau}_{XY}(n), \widehat{r}_{XY}(n))$ критерия Кендалла, основанного на статистике (9.11), по отношению к критерию, основанному на выборочном коэффициенте корреляции (9.5). Так, например, если X и Y — гауссовские величины, то $e = 0{,}912$, если равномерные, то $e = 1$, если имеют распределение Лапласа, то $e = 1{,}266$.

9.4. Критерий Спирмена

Пусть справедливо предположение о том, что функции распределения $F_X(x)$, $F_Y(y)$ и $F_W(x, y)$ непрерывны.

Ранговая статистика

$$\widehat{\rho}_{XY}(n) = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}}, \tag{9.12}$$

где R_i — ранг элемента X_i в выборке $\mathbb{X}_n$; S_i — ранг Y_i в выборке $\mathbb{Y}_n$, $i = 1, \ldots, n$, а $\overline{R}$ и $\overline{S}$ — соответствующие средние арифметические рангов, называется *коэффициентом ранговой корреляции Спирмена.*

Основанием для введения термина «ранговый коэффициент корреляции» послужило следующее соображение: если в определении (9.3) выборочного коэффициента корреляции $\widehat{r}_{XY}(n)$ заменить элементы выборок X_i, Y_i, $i = 1, \ldots, n$ их рангами R_i и S_i соответственно, то полученная таким образом статистика совпадает с $\widehat{\rho}_{XY}(n)$.

Статистику (9.12) можно преобразовать в более удобную форму

$$\widehat{\rho}_{XY}(n) = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(R_i - S_i)^2. \tag{9.13}$$

При наличии связок в выборках нужно внести следующую поправку в формулу (9.13):

$$\widehat{\rho}_{XY}(n) = 1 - \frac{\sum_{i=1}^{n}(R_i - S_i)^2}{\frac{1}{6}(n^3 - n) - (u_1 + u_2)}, \tag{9.14}$$

где $u_1 = \dfrac{1}{12}\sum_{k=1}^{q}(u_{1k}^3 - u_{1k})$; $u_2 = \dfrac{1}{12}\sum_{k=1}^{g}(u_{2k}^3 - u_{2k})$; q — количество связок в выборке $\mathbb{X}_n$; u_{1k} — размер k-й связки выборки $\mathbb{X}_n$; g — количество связок в выборке $\mathbb{Y}_n$; u_{2k} — размер k-й связки выборки $\mathbb{Y}_n$.

Квантили распределения статистики Спирмена $\widehat{\rho}_{XY}(n)$ при справедливости H_0 вида (9.1) для $4 \leqslant n \leqslant 100$ табулированы в [39]. Доказано [18], что при справедливости H_0 статистика

$$T_\rho(\mathbb{W}_n) = \sqrt{n-1}\,\widehat{\rho}_{XY}(n) \tag{9.15}$$

асимптотически нормальна.

Рассмотрим взаимосвязь между статистиками $\widehat{\rho}_{XY}(n)$ и $\widehat{\tau}_{XY}(n)$.

Теорема 9.1. *Пусть выборка $\mathbb{X}_n$ упорядочена в порядке возрастания, а $T_1, \ldots, T_n$ — ранги соответствующих элементов $Y_1, \ldots, Y_n$.*

Обозначим $I(t) = \begin{cases} 1, & t > 0, \\ 0, & t \leqslant 0. \end{cases}$ *Тогда*

$$\widehat{\rho}_{XY}(n) = 1 - \frac{12}{n^3 - n} \sum_{1 \leqslant i < j \leqslant n} (j - i) I(T_i - T_j),$$

$$\widehat{\tau}_{XY}(n) = 1 - \frac{4}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} I(T_i - T_j).$$

Доказательство теоремы 9.1 приведено в [38] (см. теорему 4.4.1).

Из теоремы 9.1 следует, что за исключением крайних ситуаций, когда $\widehat{\rho}_{XY}(n) = \widehat{\tau}_{XY}(n) = 1$ или $\widehat{\rho}_{XY}(n) = \widehat{\tau}_{XY}(n) = -1$ коэффициенты $\widehat{\rho}_{XY}(n)$ и $\widehat{\tau}_{XY}(n)$, вообще говоря, не равны. Однако при справедливости H_0 вида (9.1) эти статистики сильно коррелированы, а именно, как показано в [18],

$$r(\widehat{\rho}_{XY}(n), \widehat{\tau}_{XY}(n)) = \frac{2(n+1)}{\sqrt{2n(2n+5)}} \to 1 \text{ при } n \to \infty.$$

Критерий Спирмена, основанный на статистике (9.13), состоятелен для тех же альтернатив, что и критерий Кендалла. Соответствующие критические области для критерия со статистикой $T_\rho(\mathbb{W}_n)$ приведены в табл. 9.2.

Отметим, что распределение коэффициента ранговой корреляции Спирмена при нарушении гипотезы H_0 вида (9.1) изучено недостаточно.

9.5. Критерий хи-квадрат

Критерий предназначен для проверки гипотезы H_0 вида (9.1) против альтернативы H_A общего вида (9.6). Предположим сначала, что случайный вектор $W = (X, Y)$ является дискретным, и его первая компонента X принимает конечное множество значений $\{a_1, \ldots, a_m\}$, а вторая компонента Y — конечное множество значений $\{b_1, \ldots, b_k\}$.

Обозначим через n_{ij}, $i = 1, \ldots, m$; $j = 1, \ldots, k$ случайное число пар (X_l, Y_l) элементов двумерной выборки $\mathbb{W}_n = [(X_1, Y_1), \ldots, (X_n, Y_n)]$, реализации (x_l, y_l) которых равны $\left(a_i, b_j\right)$. Пусть также $n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$, $i = 1, \ldots, m$, $n_{\cdot j} = \sum_{i=1}^{m} n_{ij}$, $j = 1, \ldots, k$.

Понятно, что $n_{i\cdot}$ — количество элементов выборки $\mathbb{X}_n$, принявших значение a_i $i = 1,\ldots,m$, а $n_{\cdot j}$ — количество элементов выборки $\mathbb{Y}_n$, принявших значение b_j, $j = 1,\ldots,k$. Заметим также, что $\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k} n_{ij} = n$ по построению.

Обозначим через $p^*_{ij} = \dfrac{n_{ij}}{n}$, $i = 1,\ldots,m$, $j = 1,\ldots,k$ частоту появления соответствующих значений $\left(a_i, b_j\right)$ в выборке $\mathbb{W}_n$. Аналогично, $p^*_{i\cdot} = \dfrac{n_{i\cdot}}{n}$ — частота появления значения a_i, $i = 1,\ldots,m$, а $p^*_{\cdot j} = \dfrac{n_{\cdot j}}{n}$ — частота появления значения b_j, $j = 1,\ldots,k$.

Если компоненты вектора $W = (X, Y)^\top$ имеют другую структуру, например, СВ X и Y являются непрерывными, то следует провести предварительную группировку данных. Для этого область V_X всех возможных значений СВ X разбивается на $m > 1$ непересекающихся интервалов $\{\Delta_{X,i}, i = 1,\ldots,m\}$ так, что $\bigcup\limits_{i=1}^{m} \Delta_{X,i} = V_X$. Аналогично, разобьем область V_Y всех возможных значений СВ Y на $k > 1$ непересекающихся интервалов $\{\Delta_{Y,j}, j = 1,\ldots,k\}$ так, что $\bigcup\limits_{j=1}^{k} \Delta_{Y,j} = V_Y$.

При такой структуре n_{ij} — это случайное число пар (X_l, Y_l) элементов двумерной выборки $[(X_1, Y_1),\ldots,(X_n, Y_n)]$, реализации (x_l, y_l), $l = 1,\ldots,n$ которых попали в прямоугольник $\Delta_{ij} = \Delta_{X,i} \times \times \Delta_{Y,j}$, $i = 1,\ldots,m$, $j = 1,\ldots,k$. Тогда $n_{i\cdot}$ — количество элементов выборки $\mathbb{X}_n$ реализации которых попали в $\Delta_{X,i}$, $i = 1,\ldots,m$, а $n_{\cdot j}$ — количество элементов выборки $\mathbb{Y}_n$ реализации которых попали в $\Delta_{Y,j}$, $j = 1,\ldots,k$.

Статистика критерия хи-квадрат имеет вид

$$\widehat{\chi}^2_n = T_G(\mathbb{W}_n) = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(p^*_{ij} - p^*_{i\cdot}p^*_{\cdot j})^2}{p^*_{i\cdot}p^*_{\cdot j}} = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(n_{ij} - \dfrac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{n_{i\cdot}n_{\cdot j}}. \tag{9.16}$$

Статистику (9.16) принято называть *статистикой хи-квадрат, или статистикой Пирсона.*

Согласно теореме Фишера–Пирсона [33] асимптотическое распределение статистики (9.16) при справедливости гипотезы H_0 вида (9.1) есть распределение хи-квадрат $\mathcal{H}_r$ с $r = (m-1)(k-1)$ степенями свободы, т.е.

$$\widehat{\chi}^2_n \sim \mathcal{H}_r \text{ при } n \to \infty.$$

Большие расхождения между частотами p^*_{ij} и соответствующими им произведениями частот $p^*_{i\cdot}p^*_{\cdot j}$ говорят о нарушении гипотезы H_0

вида (9.1). Таким образом, в пользу альтернативной гипотезы (9.6) свидетельствуют большие значения статистики (9.16).

Следовательно, критическая область уровня значимости α для данного критерия имеет вид $(k_{1-\alpha}(r), +\infty)$, где $k_{1-\alpha}(r)$ — квантиль уровня $1-\alpha$ распределения $\mathcal{H}_r$.

Для удобства вычислений можно использовать другую форму записи статистики (9.16) вида

$$\widehat{\chi}_n^2 = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij})^2}{n_{i\cdot}n_{\cdot j}} - 1\right). \tag{9.17}$$

Для случая $m = k = 2$, часто встречающегося на практике, формула (9.16) принимает простой вид:

$$\widehat{\chi}_n^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}. \tag{9.18}$$

Отметим, что критерий хи-квадрат является состоятельным для альтернатив общего вида (9.6).

9.6. Проверка гипотезы о независимости двух номинальных признаков

При проведении социологических и психологических исследований часто приходится иметь дело с экспериментальными данными, которые не только не имеют количественного выражения, но и не могут быть упорядочены. Например, пол, профессия, принадлежность к политической партии, отношение к какой-либо религиозной конфессии и т.д. Такие категоризованные данные принято называть признаками, измеренными в номинальной шкале. Рассмотрим задачу выявления статистической зависимости (независимости) между признаками, измеренными в номинальной шкале, и укажем меры, описывающие силу связи между ними.

Пусть признак A измеряется в номинальной шкале и имеет категории (градации) $A_1, \ldots, A_m$, а признак B, измеренный в номинальной шкале, имеет категории $B_1, \ldots, B_k$. Например, признак A — цвет глаз человека, а признак B — пол человека. Тогда признак A имеет категории: A_1 — карий, A_2 — зеленый, A_3 — серый, A_4 — голубой, а признак B категории: B_1 — мужской, B_2 — женский.

Определим следующие случайные события:

$A_i = \{$признак A у случайно выбранного объекта имеет i-ю категорию$\}$, $i = 1, \ldots, m$;

$B_j = \{$признак B у случайно выбранного объекта имеет j-ю категорию$\}$, $j = 1, \ldots, k$.

Обозначим $p_{ij} = \mathbf{P}(A_i \cdot B_j)$, $p_{i\cdot} = \mathbf{P}(A_i)$, $p_{\cdot j} = \mathbf{P}(B_j)$, $i = 1, \ldots, m$, $j = 1, \ldots, k$.

Определение 9.4. Признаки A и B, измеренные в номинальной шкале, называются независимыми, если

$$p_{ij} = p_{\cdot i} p_{\cdot j}, \ \forall i = 1, \ldots, m, \ j = 1, \ldots, k. \tag{9.19}$$

Пусть случайным образом выбрано n объектов, и у каждого из этих объектов измерен признак A и признак B. Результаты измерений удобно представить в виде таблицы 9.3 размера $m \times k$. В этой таблице n_{ij} обозначает количество объектов, у которых признак A имеет категорию A_i и признак B категорию B_j; $n_{i\cdot} = \sum\limits_{j=1}^{k} n_{ij}$, $i = 1, \ldots, m$ — количество объектов, у которых признак A имеет категорию A_i; $n_{\cdot j} = \sum\limits_{i=1}^{m} n_{ij}$, $j = 1, \ldots, k$ — количество объектов, у которых признак B имеет категорию B_j. Таблицу 9.3 называют *таблицей сопряженности признаков A и B*.

Таблица 9.3

	B_1	$\cdots$	B_k	
A_1	n_{11}	$\cdots$	n_{1k}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_m	n_{m1}	$\cdots$	n_{mk}	$n_{m\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot k}$	n

Имея таблицу сопряженности признаков, можно оценить неизвестные вероятности p_{ij}, $p_{\cdot i}$ и $p_{\cdot j}$, $i = 1, \ldots, m$, $j = 1, \ldots, k$. Обозначим через $p^*_{ij} = \dfrac{n_{ij}}{n}$ частоту случайного события $A_i \cdot B_j$, $i = 1, \ldots, m$, $j = 1, \ldots, k$, через $p^*_{i\cdot} = \dfrac{n_{i\cdot}}{n}$ — частоту события A_i, а через $p^*_{\cdot j} = \dfrac{n_{\cdot j}}{n}$ частоту события B_j. Введенные частоты p^*_{ij}, $p^*_{i\cdot}$, $p^*_{\cdot j}$ являются несмещенными и сильно состоятельными оценками соответствующих вероятностей p_{ij}, $p_{\cdot i}$ и $p_{\cdot j}$.

Для проверки гипотезы H_0 (9.19) о независимости признаков A и B применяется критерий Пирсона хи-квадрат со статистикой вида

$$\widehat{\chi}^2_n = n \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(p^*_{ij} - p^*_{i} p^*_{\cdot j})^2}{p^*_{i} p^*_{\cdot j}} = n \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(n_{ij} - \dfrac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{n_{i\cdot} n_{\cdot j}}. \tag{9.20}$$

Статистика (9.20) при справедливости гипотезы H_0 и $n \to \infty$ имеет распределение хи-квадрат $\mathcal{H}_r$ с $r = (m-1)(k-1)$ степенями свободы, а критическая область уровня значимости α вид $(k_{1-\alpha}(r); +\infty)$, где $k_{1-\alpha}(r)$ — квантиль уровня $1-\alpha$ распределения $\mathcal{H}_r$.

Рассмотренный критерий хи-квадрат является состоятельным для альтернативы общего вида

$$H_A : \exists\, i, j \text{ такие, что } p_{ij} \neq p_{i\cdot}\, p_{\cdot j}\,. \tag{9.21}$$

Теперь проведем формализацию рассмотренной задачи в терминах случайных величин и выборок. Свяжем с признаками A и B случайные величины X_A и X_B. А именно, рассмотрим полиномиальную СВ X_A с параметрами $(1, p_{1\cdot}, \ldots, p_{m\cdot})$. Первый параметр, равный единице, обозначает количество проведенных испытаний. Предполагается, что в испытании может произойти одно из m описанных выше событий $A_1, \ldots, A_m$ с вероятностями $p_{1\cdot}, \ldots, p_{m\cdot}$ соответственно. Рассмотрим также полиномиальную СВ X_B с параметрами $(1, p_{\cdot 1}, \ldots, p_{\cdot k})$, где первый параметр обозначает количество проведенных испытаний, и в испытании может произойти одно из событий $B_1, \ldots, B_k$ с вероятностями $p_{\cdot 1}, \ldots, p_{\cdot k}$ соответственно. Рассмотрим теперь совместное распределение СВ X_A и X_B, обозначив через p_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k$, вероятность одновременного появления событий A_i и B_j в проведенном испытании. Реализациями пары случайных величин (X_A, X_B) являются матрицы размера $m \times k$, в которых один элемент равен единице, а остальные $m \cdot k - 1$ элементов равны нулю. Если элемент равный единице стоит в i-й строке и j-м столбце матрицы, то это означает, что в опыте реализовалось событие $A_i \cdot B_j$. Нетрудно видеть, что независимость признаков A и B, определенная в (9.19), эквивалентна независимости случайных величин X_A и X_B.

Пусть теперь имеется выборка объема n, порожденная парой случайных величин (X_A, X_B). Каждый элемент такой выборки является матрицей, состоящей из $m \cdot k - 1$ нулей и одной единицы. Такую выборку можно также представить в виде табл. 9.3 размера $m \times k$, где n_{ij} — случайное число матриц, у которых единицы стоят в i-й строке и j-м столбце, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant k$.

В рамках данной модели для проверки гипотезы о независимости случайных величин X_A и X_B применяется описанный выше критерий Пирсона хи-квадрат со статистикой (9.20).

Меры связи, основанные на статистике $\widehat{\chi}^2_n$. Понятно, что большие значения статистики $\widehat{\chi}^2_n$ говорят о наличии зависимости между признаками A и B. Однако, непосредственно величина $\widehat{\chi}^2_n$ не позволяет судить о степени этой зависимости, так как величина $\widehat{\chi}^2_n \to \infty$ при неограниченном возрастании n, если признаки A и B зависимы.

В качестве меры связи признаков A и B К. Пирсон предложил *коэффициент взаимной сопряженности* (*или коэффициент Пирсона*)

$$P = \sqrt{\frac{\widehat{\chi}_n^2}{\widehat{\chi}_n^2 + n}}. \tag{9.22}$$

Основанием для его введения послужил следующий факт. Если для двумерной гауссовской выборки $[(X_1, Y_1), \ldots, (X_n, Y_n)]$ провести описанное выше разбиение на прямоугольники $\Delta_{ij} = \Delta_{X,i} \times \Delta_{Y,i}$, $i = 1, \ldots, m$, $j = 1, \ldots, k$, то при возрастании m и k коэффициент

$$P^2 = \frac{\widehat{\chi}_n^2}{\widehat{\chi}_n^2 + n} \to r_{XY}^2,$$

где $\widehat{\chi}_n^2$ — статистика вида (9.16), а r_{XY}^2 — коэффициент корреляции СВ X и Y, порождающих выборки $\mathbb{X}_n$ и $\mathbb{Y}_n$.

Однако, в отличие от r_{XY}, максимальное значение P равно $\sqrt{\dfrac{l-1}{l}} < 1$, где $l = \min(m, k)$. Чтобы устранить этот недостаток, Крамер ввел другую меру

$$C = \sqrt{\frac{\widehat{\chi}_n^2}{n \cdot \min\{(m-1), (k-1)\}}}, \tag{9.23}$$

которая называется *коэффициентом Крамера.*

Значение коэффициента Крамера $C \in [0; 1]$ и верхний предел $C = 1$ достигается тогда и только тогда, когда каждая строка (при $m \geqslant k$) или каждый столбец (при $m \leqslant k$) табл. 9.3 содержит лишь один отличный от нуля элемент.

Значения коэффициентов P и C, близкие к 1, говорят о сильной связи между признаками A и B. Если гипотеза о независимости признаков A и B отвергнута, то принято считать, что значения коэффициентов P и C в интервале $[0; 0{,}3)$ говорят о слабой силе связи признаков A и B, значения в интервале $[0{,}3; 0{,}7)$ — об умеренной силе связи и значения в интервале $[0{,}7; 1]$ — о значительной силе связи.

9.7. Коэффициенты связи, основанные на прогнозе

Пусть известно совместное распределение пары случайных величин (X_A, X_B), описанное в п. 9.6. Назовем модальным (наиболее вероятным) исходом полиномиальной СВ X_B с параметрами $(1, p_{\cdot 1}, \ldots, p_{\cdot k})$ такое событие B_j, для которого $p_{\cdot j} = \max\limits_{1 \leqslant l \leqslant k} p_{\cdot l}$.

В качестве прогноза номинальной переменной (признака) B с категориями $B_1, \ldots, B_k$ естественно выбрать такую категорию B_j, которая представляет собой модальный исход соответствующей СВ X_B. Вероятность ошибки такого прогноза (назовем его первым прогнозом) будет

$$p_1 = 1 - \max_{1\leqslant l\leqslant k} p_{\cdot l}.$$

При определении первого прогноза не было учтено значение СВ X_A. Если же СВ X_A и X_B зависимы, то естественно предположить, что прогноз модальной категории признака B может быть улучшен (т.е. вероятность ошибки прогноза будет уменьшена), если при прогнозировании будет учтено совместное распределение случайных величин X_A и X_B.

Пусть известно, что в результате проведенного испытания реализовалось событие A_i. Назовем вторым прогнозом признака B такую категорию B_j, для которой

$$p_{ij} = \max_{1\leqslant l\leqslant k} p_{il}.$$

Вероятность ошибки второго прогноза составит

$$p_2 = 1 - \sum_{i=1}^{m} \max_{1\leqslant l\leqslant k} p_{il}.$$

Гутманом [4] была предложена *мера прогноза* λ_B, равная относительному уменьшению вероятности ошибки предсказания модальной категории признака B при переходе от первого прогноза ко второму.

Мерой λ_B называется величина

$$\lambda_B = \frac{p_1 - p_2}{p_1} = \frac{\sum\limits_{i=1}^{m} \max\limits_{1\leqslant l\leqslant k} p_{il} - \max\limits_{1\leqslant l\leqslant k} p_{\cdot l}}{1 - \max\limits_{1\leqslant l\leqslant k} p_{\cdot l}}. \tag{9.24}$$

Аналогично, мерой прогноза признака A называется

$$\lambda_A = \frac{\sum\limits_{j=1}^{k} \max\limits_{1\leqslant l\leqslant m} p_{lj} - \max\limits_{1\leqslant l\leqslant m} p_{l\cdot}}{1 - \max\limits_{1\leqslant l\leqslant m} p_{l\cdot}}. \tag{9.25}$$

Мера λ_B (λ_A) характеризует улучшение качества прогноза модальной категории признака B (признака A), которое обусловлено учетом совместного распределения СВ X_A и X_B. Меры λ_B и λ_A принимают значения от 0 до 1 и имеют следующую интерпретацию. Величина $\lambda_B \cdot 100\%$ ($\lambda_A \cdot 100\%$) показывает, на сколько процентов улучшится прогноз модальной категории признака B (признака A), если при прогнозировании будет учтено совместное распределение соответствующих СВ X_A и X_B.

Меры λ_B и λ_A асимметричны, так как при прогнозировании один из признаков рассматривается как причина, а другой как следствие. Если непонятно, какой из признаков является причиной, а какой следствием, то в качестве меры прогноза рассматривают симметричную меру $\lambda = \dfrac{\lambda_A + \lambda_B}{2}$.

Пусть имеется таблица сопряженности признаков A и B (см. табл. 9.3), в которой представлена выборка, порожденная парой случайных величин (X_A, X_B). Оценим по этой выборке меру λ_B.

Оценкой вероятности ошибки первого прогноза является соответствующая частота

$$\widehat{p}_1 = 1 - \frac{1}{n}\max_{1\leqslant l\leqslant k} n_{\cdot l},$$

а оценкой вероятности ошибки второго прогноза

$$\widehat{p}_2 = 1 - \frac{1}{n}\sum_{i=1}^{m}\max_{1\leqslant l\leqslant k} n_{il}.$$

Тогда оценкой меры λ_B будет коэффициент

$$\widehat{\lambda}_B = \frac{\widehat{p}_1 - \widehat{p}_2}{\widehat{p}_1} = \frac{\sum\limits_{i=1}^{m}\max\limits_{1\leqslant l\leqslant k} n_{il} - \max\limits_{1\leqslant l\leqslant k} n_{\cdot l}}{n - \max\limits_{1\leqslant l\leqslant k} n_{\cdot l}}.$$

Аналогично, оценкой меры λ_A будет коэффициент

$$\widehat{\lambda}_A = \frac{\sum\limits_{j=1}^{k}\max\limits_{1\leqslant l\leqslant m} n_{lj} - \max\limits_{1\leqslant l\leqslant m} n_{l\cdot}}{n - \max\limits_{1\leqslant l\leqslant m} n_{l\cdot}},$$

а оценкой меры λ — коэффициент $\widehat{\lambda} = \dfrac{\widehat{\lambda}_A + \widehat{\lambda}_B}{2}$.

Для мер Гутмана λ_B и λ_A можно построить не только точечные оценки $\widehat{\lambda}_B$, $\widehat{\lambda}_A$, но и асимптотические доверительные интервалы (см. пример 9.6). В [4] доказано, что статистика

$$(\widehat{\lambda}_B - \lambda_B)\sqrt{\frac{\left[n - \max\limits_{1\leqslant j\leqslant k} n_{\cdot j}\right]^3}{\left[n - \sum\limits_{i=1}^{m}\max\limits_{1\leqslant j\leqslant k} n_{ij}\right]\left[\sum\limits_{i=1}^{m}\max\limits_{1\leqslant j\leqslant k} n_{ij} + \max\limits_{1\leqslant j\leqslant k} n_{\cdot j} - 2\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k} n_{ij}\delta_{ij}\delta_j\right]}}, \tag{9.26}$$

где

$$\delta_{ij} = \begin{cases} 1, & \text{если индекс } j \text{ такой, что}\, n_{ij} = \max\limits_{1\leqslant l\leqslant k} n_{il}, \\ 0, & \text{иначе,} \end{cases} \tag{9.27}$$

$$\delta_{j} = \begin{cases} 1, & \text{если индекс } j \text{ такой, что}\, n_{\cdot j} = \max\limits_{1\leqslant l\leqslant k} n_{\cdot l}, \\ 0, & \text{иначе} \end{cases} \tag{9.28}$$

и статистика

$$(\widehat{\lambda}_A - \lambda_A)\sqrt{\frac{\left[n - \max\limits_{1\leqslant i\leqslant m} n_{i\cdot}\right]^3}{\left[n - \sum\limits_{j=1}^{k} \max\limits_{1\leqslant i\leqslant m} n_{ij}\right]\left[\sum\limits_{j=1}^{k} \max\limits_{1\leqslant i\leqslant m} n_{ij} + \max\limits_{1\leqslant i\leqslant m} n_{i\cdot} - 2\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{m} n_{ij}\delta_{ij}\delta_{i}\right]}}, \tag{9.29}$$

где

$$\delta_{ij} = \begin{cases} 1, & \text{если индекс } i \text{ такой, что } n_{ij} = \max\limits_{1\leqslant l\leqslant m} n_{lj}, \\ 0, & \text{иначе,} \end{cases} \tag{9.30}$$

$$\delta_{i} = \begin{cases} 1, & \text{если индекс } i \text{ такой, что } n_{i\cdot} = \max\limits_{1\leqslant l\leqslant m} n_{l\cdot}, \\ 0, & \text{иначе} \end{cases} \tag{9.31}$$

асимптотически нормальны.

Основываясь на этом утверждении, нетрудно видеть, что статистики (9.26) и (9.29) являются центральными статистиками параметров λ_B и λ_A соответственно.

9.8. Исследование зависимости между несколькими СВ

Как было показано в разделе 9.2, коэффициент корреляции r_{XY} может быть использован в качестве меры, описывающей силу связи между двумя СВ X и Y. Однако встречаются ситуации, когда коррелированность двух СВ является лишь отражением того факта, что обе они коррелированы с некоторой третьей СВ или совокупностью СВ. В этом случае говорят о наличии «ложной» корреляции. Для того чтобы выяснить, является ли наблюдаемая коррелированность «ложной», требуется устранить влияние третьих величин. С этой целью рассматривают условное распределение двух СВ при фиксированных значениях третьих величин, и определяют частный коэффициент корреляции.

Частные коэффициенты корреляции. Пусть $\xi = [\xi_1, \ldots \xi_l]^\top$ — невырожденный гауссовский вектор с корреляционной матрицей

$K_\xi = \{r_{ij}\}$, $i, j = 1, \ldots, l$, где r_{ij} — коэффициент корреляции между ξ_i и ξ_j.

Определение 9.5. *Частным коэффициентом корреляции СВ* ξ_1 *и* ξ_2 при фиксированных значениях $\xi_3, \ldots, \xi_l$ называется

$$r_{12;3,\ldots,l} = \frac{-K_{12}}{\sqrt{K_{11}K_{22}}}, \tag{9.32}$$

где K_{ij} — алгебраическое дополнение элемента r_{ij} матрицы K_ξ.

Аналогично определяются частные коэффициенты корреляции между любыми двумя компонентами вектора ξ при фиксированных значениях других компонент.

Нетрудно показать (см. пример 9.7), что при $l = 3$ частным коэффициентом корреляции СВ ξ_1 и ξ_2 при фиксированной ξ_3 будет

$$r_{12;3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}.$$

По различию между $r_{12;3,\ldots,l}$ и $r_{\xi_1\xi_2}$ можно судить о том, зависимы ли ξ_1 и ξ_2 между собой или зависимость между ними есть следствие зависимости каждой из них от величин $\xi_3, \ldots, \xi_l$. Так, если корреляция между ξ_1 и ξ_2 уменьшается, когда зафиксированы $\xi_3, \ldots, \xi_l$, то это означает, что зависимость между ξ_1 и ξ_2 частично обусловлена воздействием величин $\xi_3, \ldots \xi_l$. Если частная корреляция равна нулю или мала, то зависимость между ξ_1 и ξ_2 целиком обусловлена воздействием СВ $\xi_3, \ldots, \xi_l$. Наоборот, если частная корреляция больше парной $r_{\xi_1\xi_2}$, то величины $\xi_3, \ldots \xi_l$ ослабляют связь между ξ_1 и ξ_2.

Пусть имеется l выборок $X_1 = [X_{11}, \ldots, X_{n1}]^\top, \ldots, X_l = [X_{1l}, \ldots, X_{nl}]^\top$, порожденных СВ $\xi_1, \ldots, \xi_l$ соответственно. Предполагается, что выборки $X_1, \ldots, X_l$ получены в процессе совместного наблюдения за $\xi_1, \ldots \xi_l$, а именно: если x_{k1} — реализация СВ ξ_1 в k-м опыте, то $x_{k2}, \ldots, x_{kl}$ — реализации соответственно $\xi_2, \ldots, \xi_l$ в k-м опыте, $k = 1, \ldots, n$.

Обозначим через $\widehat{r}_{ij}$ — выборочный коэффициент корреляции СВ ξ_i и ξ_j, построенный по выборкам X_i и X_j, а через $\widehat{K}_\xi$ — симметричную матрицу размера $l \times l$, элементами которой являются $\widehat{r}_{ij}$, $i, j = 1, \ldots, l$, причем $\widehat{r}_{ii} = 1$ для любого $i = 1, \ldots, l$.

Оценками неизвестных частных коэффициентов корреляции служат выборочные частные коэффициенты корреляции, которые получаются путем замены в формуле (9.32) коэффициентов корреляции r_{ij} на соответствующие оценки $\widehat{r}_{ij}$.

В [21] показано, что распределение выборочного частного коэффициента корреляции $\widehat{r}_{12;3,\ldots,l}$, вычисленного по выборке объема n, такое же, как у выборочного коэффициента корреляции $\widehat{r}_{12}$, основанного на $n - d$ наблюдениях, где $d = l - 2$ — количество фиксированных переменных. Таким образом, имеется возможность для построения доверительных интервалов частного коэффициента корреляции и проверки гипотезы о равенстве его нулю.

Множественный коэффициент корреляции.

Определение 9.6. *Множественным коэффициентом корреляции* между СВ ξ_1 и совокупностью СВ $\xi_2, \dots \xi_l$ называется величина

$$R_{1(2\dots l)} = \sqrt{1 - \frac{\det K_{\xi}}{K_{11}}}, \tag{9.33}$$

где K_{11} — алгебраическое дополнение элемента $(1, 1)$ матрицы K_{ξ}.

Можно показать, что $R^2_{1(2\dots l)}$ выражается через частные коэффициенты корреляции следующим образом

$$R^2_{1(2\dots l)} = 1 - \left(1 - r^2_{12}\right)\left(1 - r^2_{13;2}\right) \cdot \ldots \cdot \left(1 - r^2_{1l;2,3,\dots,l-1}\right). \tag{9.34}$$

Используя соотношение (9.34), покажем, что множественный коэффициент корреляции характеризует зависимость между компонентами вектора $\xi = (\xi_1, \dots, \xi_l)^\top$. Обозначим через $I^{(j)}$ — подмножество элементов из множества $\{2, \dots, l\}$, которое не содержит элемент j, а через $r_{1j;I^{(j)}}$ — частный коэффициент корреляции между ξ_1 и ξ_j при фиксированных $\xi_{i_1}, \dots, \xi_{i_m}$, где индексы $i_1, \dots, i_m \in I^{(j)}$.

Если $R_{1(2\dots l)} = 0$, то, как следует из (9.34), для любого $j = 2, \dots, l$ и любого множества индексов $I^{(j)}$ частные коэффициеты $r_{1j;I^{(j)}} = 0$. Последнее означает, что СВ ξ_1 некоррелирована со всеми остальными компонентами вектора ξ. Если же $R_{1(2\dots l)} = 1$, то по крайней мере один из частных коэффициентов корреляции $r_{1j;I^{(j)}}$ должен быть равен единице. Это означает, что СВ ξ_1 является линейной функцией случайных величин $\xi_2, \dots, \xi_l$.

Отметим также еще некоторые важные свойства множественного коэффициента корреляции:

1) $0 \leqslant R_{1(2\dots l)} \leqslant 1$;

2) $R_{1(2\dots l)} \geqslant \left|r_{1j;I^{(j)}}\right|$;

3) $R^2_{1(2)} = r^2_{12}$;

4) $R^2_{1(2)} \leqslant R^2_{1(2,3)} \leqslant \dots \leqslant R^2_{1(2\dots l)}$.

Последнее свойство означает, что коэффициент множественной корреляции нельзя уменьшить путем расширения множества СВ, относительно которых измеряется зависимость СВ ξ_1.

Рассмотрим оценку $\widehat{R}_{1(2\dots l)}$ неизвестного коэффициента $R_{1(2\dots l)}$, которая получается заменой матрицы K_{ξ} в формуле (9.34) матрицей $\widehat{K}_{\xi}$.

Фишер доказал [21], что при справедливости гипотезы

$$H_0: \ R_{1(2\ldots l)} = 0 \tag{9.35}$$

статистика

$$T_n(X_1, \ldots, X_l) = \widehat{F} = \frac{\frac{1}{l-1}\widehat{R}^2_{1(2\ldots l)}}{\frac{1}{n-l}\left(1 - \widehat{R}^2_{1(2\ldots l)}\right)} \tag{9.36}$$

имеет F-распределение $F(l-1; n-l)$.

Критическая область уровня значимости α критерия со статистикой (9.36) для проверки гипотезы (9.35) против альтернативы H_1: $R_{1(2\ldots l)} > 0$ имеет вид $(f_{1-\alpha}(l-1; n-l); +\infty)$, где $f_{1-\alpha}(l-1; n-l)$ — квантиль распределения $F(l-1; n-l)$ уровня $1-\alpha$.

Коэффициент конкордации Кендалла. Пусть имеется m ранжировок $\{R_{11}, \ldots, R_{n1}\}, \ldots, \{R_{1m}, \ldots, R_{nm}\}$, построенных по выборкам $X_1 = [X_{11}, \ldots, X_{n1}]^\top, \ldots, X_m = [X_{1m}, \ldots, X_{nm}]^\top$, порожденных непрерывными СВ $\xi_1, \ldots, \xi_m$ соответственно. Предполагается, что выборки получены в результате совместного наблюдения за СВ $\xi_1, \ldots, \xi_m$.

Требуется проверить гипотезу о независимости СВ $\xi_1, \ldots, \xi_m$ вида

$$H_0: \ F_\xi(x_1, \ldots, x_m) = F_{\xi_1}(x_1) \cdot \ldots \cdot F_{\xi_m}(x_m) \ \forall x_1, \ldots, x_m \in \mathbb{R}^1, \tag{9.37}$$

где $F_\xi(x_1, \ldots, x_m)$ — функция распределения случайного вектора $\xi = (\xi_1, \ldots, \xi_m)^\top$, а $F_{\xi_i}(x_i)$ — функции распределения СВ ξ_i, $i = 1, \ldots, m$.

Коэффициентом конкордации Кендалла называется статистика

$$\begin{gathered} T_n(X_1, \ldots, X_m) = \widehat{W}_n(m) = \\ = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^{n} \left[\sum_{j=1}^{m} R_{ij} - \frac{m(n+1)}{2} \right]^2. \end{gathered} \tag{9.38}$$

Для $3 \leqslant n \leqslant 7$ и $3 \leqslant m \leqslant 20$ есть таблицы квантилей распределения статистики $\widehat{W}_n(m)$ при справедливости гипотезы H_0 вида (9.37).

При справедливости H_0 вида (9.37) и достаточно большом m статистика

$$m(n-1)\widehat{W}_n(m) \tag{9.39}$$

имеет распределение хи-квадрат H_r с $r = n-1$ степенями свободы. Критическая область уровня значимости α критерия со статистикой (9.39) имеет вид: $(k_{1-\alpha}(n-1); +\infty)$.

Отметим свойства коэффициента конкордации:

1) $0 \leqslant \widehat{W}_n(m) \leqslant 1$;

2) $\widehat{W}_n(m) = 1$ тогда и только тогда, когда все m ранжировок совпадают;

3) пусть $\overline{\rho}(m)$ — среднее арифметическое значений ранговых коэффициентов корреляции Спирмена, вычисленных по всем C_m^2 парам ранжировок выборок $X_1, \ldots, X_m$, тогда $\widehat{W}_n(m) = \dfrac{m\overline{\rho}(m) - 1}{m - 1}$.

Критерий, основанный на коэффициенте конкордации Кендалла, используется в задачах исследования согласованности экспертной группы. Пусть имеется n объектов, характеризующихся некоторым качественным показателем, и группа из m экспертов, способных оценить данный показатель. Каждый из m экспертов производит ранжирование всех объектов по рассматриваемому показателю. Обозначим R_{ij} — ранг присвоенный i-му объекту j-м экспертом, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Тогда коэффициент конкордации $\widehat{W}_n(m)$ может служить оценкой степени согласованности суждений экспертов. Действительно, согласно свойству 2), коэффициент $\widehat{W}_n(m)$ принимает свое максимальное значение в том случае, когда ранжировки всех экспертов совпадают. Если же различия между ранжировками экспертов велики, то суммы рангов, присвоенные каждому из n объектов $\sum\limits_{j=1}^{m} R_{ij}$, $i = 1, \ldots, n$ будут близки к среднему значению суммы рангов всех экспертов, равному $\dfrac{1}{n}\left(m\sum\limits_{i=1}^{n} i\right) = \dfrac{m(n+1)}{2}$, а коэффициент $\widehat{W}_n(m)$ близок к нулю.

Таким образом, если реализация статистики $\widehat{W}_n(m)$ близка к нулю, то говорят, что суждения экспертов не характеризуется общностью предпочтений, т.е. экспертная группа рассогласована. Если же реализация $\widehat{W}_n(m)$ близка к единице, то это свидетельствует в пользу того, что экспертная группа обладает единой системой предпочтений, т.е. является согласованной.

9.9. Примеры

Пример 9.1. Имеются данные [15] о ВВП в паритетах покупательной способности (показатель X) и коэффициенте младенческой смертности в промиллях (показатель Y) по 15 странам за 1995 г. (табл. 9.4). Применяя критерии Спирмена и Кендалла, выясните, являются ли показатели X и Y зависимыми. Уровень значимости выберите равным 0,05.

Таблица 9.4

i	Страна	x_i	y_i	i	Страна	x_i	y_i
1	Мозамбик	3,0	113	9	Бразилия	20	44
2	Чад	2,6	117	10	Греция	43,4	8
3	Бангладеш	5,1	79	11	Республика Корея	42,4	10
4	Индия	5,2	68	12	Италия	73,7	7
5	Египет	14,2	56	13	Канада	78,3	6
6	Белоруссия	15,6	13	14	США	100	8
7	Польша	20	14	15	Швейцария	95,9	6
8	Мексика	23,7	33				

Решение. Пусть выборка $[(X_1, Y_1), \ldots, (X_n, Y_n)]^\top$, $n = 15$, порождена двумерным случайным вектором $W = (X, Y)^\top$, имеющим некоторое непрерывное распределение $F_W(x, y)$. Проверим гипотезу H_0 о независимости СВ X и Y вида (9.1). В качестве альтернативной гипотезы можно выбрать гипотезу H_A: $\tau_{XY} \neq 0$, которая означает, что между СВ X и Y имеется монотонная связь. Если же предполагается, что между СВ X и Y имеется отрицательная монотонная связь, т.е. с ростом одной из величин другая величина убывает, то в качестве альтернативы следует выбрать гипотезу H_A: $\tau_{XY} < 0$.

Применим критерий Спирмена. Для вычисления статистики критерия составим таблицу рангов

Таблица 9.5

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
r_i	2	1	3	4	5	6	7,5	9	7,5	11	10	12	13	15	14
s_i	14	15	13	12	11	7	8	9	10	4,5	6	3	1,5	4,5	1,5

в которой r_i — реализация ранга элемента x_i в выборке $[X_1, \ldots, X_{15}]^\top$, а s_i — реализация ранга элемента y_i в выборке $[Y_1, \ldots, Y_{15}]^\top$.

Реализация статистики Спирмена (9.13)

$$\widehat{\rho}_{XY}(n) = 1 - \frac{6[(2-14)^2 + (1-15)^2 + \ldots + (14-1{,}5)^2]}{15^3 - 15} =$$

$$= 1 - \frac{6[144 + 196 + \ldots + 156{,}25]}{3360} = 1 - \frac{6 \cdot 1085{,}5}{3360} = -0{,}938.$$

Так как в выборках имеются связки, то при вычислении $\widehat{\rho}_{XY}(n)$ следует внести поправку (9.14). В реализации выборки соответствующей СВ X есть одна связка размера два, т.е. $u_{11} = 2$, а в выборке, соответствующей СВ Y, — две связки размера два, т.е. $u_{21} = 2$ и $u_{22} = 2$. Тогда $u_1 = \dfrac{1}{12}(2^3 - 2) = 0{,}5$, $u_2 = \dfrac{1}{12}(2^3 - 2)2 = 1$, а

$$\widehat{\rho}_{XY}(n) = 1 - \frac{\sum\limits_{i=1}^{n}(r_i - s_i)^2}{\frac{1}{6}(n^3 - n) - (u_1 + u_2)} = 1 - \frac{1085{,}5}{\frac{1}{6}3360 - 1{,}5} = -0{,}944.$$

Если в качестве альтернативной гипотезы выбрана H_A: $\tau_{XY} \neq$ $\neq 0$, то критическая область будет иметь вид $\left[-1; z_{\frac{\alpha}{2},n}\right) \cup \left(z_{1-\frac{\alpha}{2},n}; 1\right]$, где $z_{\frac{\alpha}{2},n}$, $z_{1-\frac{\alpha}{2},n}$ — квантили уровня $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$ распределения коэффициента ранговой корреляции Спирмена при справедливости гипотезы H_0 вида (9.1) для выборки объема n. По таблицам [39] находим $z_{0,975,15} = 0{,}521$ и, учитывая, что $z_{1-\frac{\alpha}{2},n} = -z_{\frac{\alpha}{2},n}$, имеем $z_{0,025,15} = -0{,}521$.

Тогда реализация статистики $\widehat{\rho}_{XY}(n) = -0{,}944$ попадает в критическую область $[-1; -0{,}521) \cup (0{,}521; 1]$. Таким образом, гипотеза о независимости СВ X и Y отвергается на уровне значимости 0,05 в пользу альтернативы H_A: $\tau_{XY} \neq 0$ о том, что СВ X и Y зависимы.

Если в качестве альтернативной гипотезы выбрана гипотеза H_A: $\tau_{XY} < 0$, то критическая область будет иметь вид: $[-1; z_{\alpha,n}) =$ $= [-1; z_{0,05,15}) = [-1; -0{,}443)$. Реализация статистики $\widehat{\rho}_{XY}(n)$ также попадает в указанную критическую область. Следовательно, на уровне значимости 0,05 гипотеза H_0 отвергается в пользу H_A: $\tau_{XY} < 0$ о том, что между СВ X и Y имеется отрицательная монотонная связь.

Применим теперь критерий Кендалла. Поскольку в наблюдениях имеются связки, то для вычисления коэффициента $\widehat{\tau}_{XY}(n)$ надо воспользоваться формулой (9.10).

$$\widehat{\tau}_{XY}(n) = \frac{\sum\limits_{1\leqslant i<j\leqslant n} \operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\}}{\sqrt{\frac{1}{2}n(n-1) - u_1}\sqrt{\frac{1}{2}n(n-1) - u_2}}.$$

Для удобства вычислений реализацию двумерной выборки $(x_1, y_1), \ldots, (x_{15}, y_{15})$ запишем таким образом, чтобы $x_1 \leqslant x_2 \leqslant$ $\leqslant \ldots \leqslant x_{15}$. Получим табл. 9.6.

Таблица 9.6

i	1	2	3	4	5	6	7	8
x_i	2,6	3,0	5,1	5,2	14,2	15,6	20	20
y_i	117	113	79	68	56	13	14	44

i	9	10	11	12	13	15	14	
x_i	23,7	42,4	43,4	73,7	78,3	95,9	100	
y_i	33	10	8	7	6	6	8	

В силу того, что реализация выборки иксов упорядочена по возрастанию, имеем

$$\sum_{1\leqslant i<j\leqslant n} \operatorname{sign}((x_i - x_j)(y_i - y_j)) = \sum_{i=1}^{14}\sum_{j=i+1}^{15} \operatorname{sign}((x_i - x_j)(y_i - y_j)) =$$
$$= \sum_{i=1}^{14}\sum_{j=i+1}^{15} \operatorname{sign}((-1)(y_i - y_j)) = \sum_{i=1}^{14}\sum_{j=i+1}^{15} \operatorname{sign}(y_j - y_i).$$

Если $i = 1$, а $j = 2$, то $\operatorname{sign}(y_2 - y_1) = \operatorname{sign}(113 - 117) = -1$. Проведя аналогичные вычисления для $i = 1$ и $j = 3, \ldots, 15$, получаем, что $\sum\limits_{j=2}^{15} \operatorname{sign}(y_j - y_1) = -14$, а

$$\sum_{i=1}^{14} \sum_{j=i+1}^{15} \operatorname{sign}(y_j - y_i) = -14 - 13 - 12 - 11 - 10 - 3 - 5 - 7 - \\ -6 - 5 - 3 - 1 + 1 + 1 = -88.$$

В выборке иксов есть одна связка размера два, т.е. $u_{11} = 2$, а в выборке игреков две связки размера два, т.е. $u_{21} = 2$, $u_{22} = 2$. Тогда $u_1 = \frac{1}{2} 2 \cdot 1 = 1$, $u_2 = \left(\frac{1}{2} 2 \cdot 1\right) 2 = 2$.

Тогда реализация коэффициента согласованности

$$\widehat{\tau}_{XY}(n) = \frac{-88}{\sqrt{\dfrac{15 \cdot 14}{2} - 1}\sqrt{\dfrac{15 \cdot 14}{2} - 2}} = -0{,}85.$$

Если проверяется гипотеза H_0 вида (9.1) о независимости СВ X и Y против альтернативы H_A: $\tau_{XY} < 0$, то критическая область имеет вид $[-1; z_{\alpha,n})$, где $z_{\alpha,n}$ — квантиль уровня α распределения коэффициента согласованности Кендалла при справедливости H_0 для выборки объема n. По таблицам [39] находим ближайшую к $\alpha = 0{,}05$ квантиль $z_{0{,}046{,}15} = -0{,}33$. Таким образом, гипотеза о независимости СВ X и Y отвергается на уровне значимости $\alpha = 0{,}046$ в пользу альтернативы H_A: $\tau_{XY} < 0$. ■

Пример 9.2. Изучается зависимость между показателем X (ВВП в паритетах покупательной способности) и показателем Y (коэффициент младенческой смертности в промиллях). По имеющимся данным этих показателей для 30 стран была вычислена реализация коэффициента выборочной корреляции $\widehat{r}_{XY}(30) = -0{,}734$. Предполагая, что наблюдаемая выборка $(X_1, Y_1), \ldots, (X_{30}, Y_{30})$, порожденная случайным вектором $W = (X, Y)^\top$, соответствует двумерному гауссовскому распределению, сделайте вывод о зависимости (независимости) СВ X и Y. Постройте асимптотический доверительный интервал уровня надежности 0,95 для коэффициента корреляции r_{XY} СВ X и Y. Постройте также доверительный интервал уровня надежности 0,95 для r_{XY}, используя преобразование Фишера.

Решение. Если двумерная выборка $\{(X_i, Y_i),\ i = 1, \ldots, n\}$, где X_i — ВВП, а Y_i — коэффициент младенческой смертности в i-й стране, порождена двумерным гауссовским вектором вектором $W = (X, Y)^\top$, то гипотеза о независимости СВ X и Y вида (9.1) эквивалентна гипотезе вида (9.2) некоррелированности этих СВ. В данной задаче естественной альтернативой гипотезе H_0 вида (9.2) является гипотеза H_A: $r_{XY} < 0$ об отрицательной коррелированности показателей X и Y (см. пример 9.1). Для проверки H_0 используем критерий, основанный на выборочном коэффициенте корреляции.

Статистика (9.4) этого критерия

$$T_r(\mathbb{W}_n) = \frac{\sqrt{n-2}\,\widehat{r}_{XY}(n)}{\sqrt{1-\widehat{r}^2_{XY}(n)}}$$

имеет при справедливости H_0 распределение Стьюдента T_{n-2}.

Реализация этой статистики

$$T_r(\mathbb{W}_n) = \frac{-0{,}734\sqrt{28}}{\sqrt{1-0{,}734^2}} = -5{,}72.$$

Критическая область критерия уровня значимости $\alpha = 0{,}05$ при альтернативе $H_A : r_{XY} < 0$ имеет вид: $(-\infty; t_{0,05}(n-2)]$. По табл. 22.4 находим квантиль уровня 0,05 распределения Стьюдента с $n-2 = 28$ степенями свободы $t_{0,05}(28) = -t_{0,95}(28) = -1{,}701$. Так как реализация статистики попадает в критическую область $(-\infty; -1{,}701]$, то на уровне значимости 0,05 гипотеза о независимости ВВП и коэффициента младенческой смертности отвергается в пользу того, что эти показатели отрицательно коррелированы.

Будем считать, что объем выборки $n = 30$ достаточно велик, и построим асимптотический доверительный интервал для коэффициента корреляции r_{XY} уровня надежности 0,95. Так как СВ $\widehat{r}_{XY}(n)$ имеет асимптотическое гауссовское распределение с параметрами

$$\mathbf{M}\{\widehat{r}_{XY}(n)\} = r_{XY} - \frac{r_{XY}(1-r^2_{XY})}{2n}, \quad \mathbf{D}\{\widehat{r}_{XY}(n)\} = \frac{(1-r^2_{XY})^2}{n},$$

то при больших n

$$\mathbf{P}\left\{-u_{0,975} \leqslant \frac{\widehat{r}_{XY}(n) - \mathbf{M}\{\widehat{r}_{XY}(n)\}}{\sqrt{\mathbf{D}\{\widehat{r}_{XY}(n)\}}} \leqslant u_{0,975}\right\} = 0{,}95. \qquad (9.40)$$

Поскольку $\mathbf{D}\{\widehat{r}_{XY}(n)\}$ и смещение оценки $\widehat{r}_{XY}$, равное $\dfrac{r_{XY}(1-r^2_{XY})}{2n}$, зависят от неизвестного значения r_{XY}, то в формуле (9.40) значение r_{XY} в выражениях $\mathbf{D}\{\widehat{r}_{XY}(n)\}$ и $\dfrac{r_{XY}(1-r^2_{XY})}{2n}$ заменяется выборочным коэффициентом корреляции $\widehat{r}_{XY}(n)$. Из-за этого обстоятельства построенный асимптотический доверительный интервал будет иметь надежность, отличную от 0,95.

Проведем преобразование в (9.40) так, чтобы получить интервал для величины r_{XY}:

$$\mathbf{P}\left\{-u_{0,975} \leqslant \frac{\widehat{r}_{XY}(n) - \left(r_{XY} - \dfrac{\widehat{r}_{XY}(n)(1-\widehat{r}^2_{XY}(n))}{2n}\right)}{\frac{1}{\sqrt{n}}\,(1-\widehat{r}^2_{XY}(n))} \leqslant u_{0,975}\right\} = 0{,}95;$$

$$\mathbf{P}\left\{\frac{-u_{0,975}(1-\widehat{r}^2_{XY}(n))}{\sqrt{n}} - \frac{\widehat{r}_{XY}(n)(1-\widehat{r}^2_{XY}(n))}{2n} - \widehat{r}_{XY}(n) \leqslant -r_{XY} \leqslant \right.$$
$$\left. \leqslant \frac{u_{0,975}(1-\widehat{r}^2_{XY}(n))}{\sqrt{n}} - \frac{\widehat{r}_{XY}(n)(1-\widehat{r}^2_{XY}(n))}{2n} - \widehat{r}_{XY}(n)\right\} = 0{,}95;$$

$$\mathbf{P}\left\{\widehat{r}_{XY}(n) + \frac{\widehat{r}_{XY}(n)(1-\widehat{r}^2_{XY}(n))}{2n} - \frac{u_{0,975}(1-\widehat{r}^2_{XY}(n))}{\sqrt{n}} \leqslant r_{XY} \leqslant \right.$$
$$\left. \leqslant \widehat{r}_{XY}(n) + \frac{\widehat{r}_{XY}(n)(1-\widehat{r}^2_{XY}(n))}{2n} + \frac{u_{0,975}(1-\widehat{r}^2_{XY}(n))}{\sqrt{n}}\right\} = 0{,}95.$$

Подставляя имеющуюся реализацию $\widehat{r}_{XY}(n)$ в последнюю формулу, получим

$$\mathbf{P}\left\{-0{,}734 + \frac{-0{,}734(1-0{,}734^2)}{2\cdot 30} - \frac{1{,}96(1-0{,}734^2)}{\sqrt{30}} \leqslant r_{XY} \leqslant \right.$$
$$\left. \leqslant -0{,}734 + \frac{-0{,}734(1-0{,}734^2)}{2\cdot 30} + \frac{1{,}96(1-0{,}734^2)}{\sqrt{30}}\right\} = 0{,}95;$$

$$\mathbf{P}\left\{-0{,}734 - 0{,}006 - 0{,}165 \leqslant r_{XY} \leqslant -0{,}734 - 0{,}006 + 0{,}165\right\} = 0{,}95.$$

Итак, $\mathbf{P}\left\{-0{,}905 \leqslant r_{XY} \leqslant -0{,}575\right\} = 0{,}95$.

Построим теперь асимптотический доверительный интервал надежности 0,95 для r_{XY}, используя z-преобразование Фишера. Асимптотический доверительный интервал надежности 0,95 для величины $z = \dfrac{1}{2}\ln\dfrac{1+r_{XY}}{1-r_{XY}}$ имеет вид

$$\mathbf{P}\left\{\widehat{z} - \frac{\widehat{r}_{XY}(n)}{2(n-1)} - \frac{u_{0,975}}{\sqrt{n-3}} \leqslant z \leqslant \widehat{z} - \frac{\widehat{r}_{XY}(n)}{2(n-1)} + \frac{u_{0,975}}{\sqrt{n-3}}\right\} = 0{,}95, \quad (9.41)$$

где $\widehat{z} = \dfrac{1}{2}\ln\dfrac{1+\widehat{r}_{XY}(n)}{1-\widehat{r}_{XY}(n)}$. Подставляя в формулу (9.41) реализацию статистики $\widehat{z} = -0{,}937$ и $\widehat{r}_{XY}(n) = -0{,}734$, получим

$$\mathbf{P}\left\{-0{,}937 + 0{,}013 - 0{,}377 \leqslant z \leqslant -0{,}937 + 0{,}013 + 0{,}377\right\} =$$
$$= \mathbf{P}\left\{-1{,}301 \leqslant z \leqslant -0{,}544\right\} = 0{,}95.$$

Поскольку $\operatorname{th} z = \dfrac{e^z - e^{-z}}{e^z + e^{-z}} = r_{XY}$, то

$$\mathbf{P}\left\{\operatorname{th}(-1{,}301) \leqslant r_{XY} \leqslant \operatorname{th}(-0{,}544)\right\} =$$
$$= \mathbf{P}\left\{-0{,}860 \leqslant r_{XY} \leqslant -0{,}496\right\} = 0{,}95.$$

Таким образом, асимптотический доверительный интервал уровня надежности 0,95 для r_{XY} есть $[-0{,}860; -0{,}496]$. ■

Пример 9.3. Приведите пример зависимых СВ X и Y, параметр согласованности которых τ_{XY} равен нулю.

Решение. Пусть СВ X имеет четную плотность распределения $f_X(x)$, т.е. $f_X(-x) = f_X(x)$ для $\forall x \in \mathbb{R}^1$. Пусть СВ Y связана со СВ X соотношением $Y = X^2$. Понятно, что СВ X и Y являются зависимыми по построению. Покажем теперь, что $\tau_{XY} = 0$.

Пусть СВ X_1 и X_2 независимы и имеют одинаковую плотность распределения $f_X(x)$, тогда СВ $Y_1 = X_1^2$ и $Y_2 = X_2^2$ также независимы и имеют одинаковую плотность распределения $f_Y(y)$. Вычислим

$$\tau_{XY} = 1 - 2\mathbf{P}((X_1 - X_2)(Y_1 - Y_2) < 0) =$$
$$= 1 - 2\mathbf{P}((X_1 - X_2)(X_1^2 - X_2^2) < 0) =$$
$$= 1 - 2\mathbf{P}((X_1 - X_2)^2(X_1 + X_2) < 0) = 1 - 2\mathbf{P}((X_1 + X_2) < 0) =$$
$$= 1 - 2\int\limits_{-\infty}^{+\infty} dx_1 \int\limits_{-\infty}^{-x_1} f_X(x_1) f_X(x_2)\, dx_2 = 1 - 2\int\limits_{-\infty}^{+\infty} f_X(x_1) F_X(-x_1)\, dx_1 =$$
$$= \left\langle \begin{array}{l} z = -x_1 \\ dz = -dx_1 \end{array} \right\rangle = 1 + 2\int\limits_{+\infty}^{-\infty} f_X(-z) F_X(z)\, dz =$$
$$= 1 - 2\int\limits_{-\infty}^{+\infty} f_X(-z) F_X(z)\, dz = 1 - 2\frac{F_X^2(z)}{2}\Bigg|_{-\infty}^{+\infty} = 1 - 1 = 0. \ ■$$

Пример 9.4. В табл. 9.7 представлены сведения о возрасте X (в годах) и среднемесячной заработной плате Y (в тыс. руб.) 30 сотрудников некоторой организации.

Таблица 9.7

X	19	20	22	24	28	30	31	32	34	37
Y	8,5	9,2	11,2	10,4	16,2	17,4	14,3	24,9	22,8	20,8
X	39	40	41	43	45	46	47	48	50	51
Y	34,1	30,4	28,8	33,8	35,3	34,4	32,3	29,4	31,8	30,3
X	53	54	55	58	60	62	65	68	70	72
Y	28,7	31,9	25,5	19,9	22,3	20,6	18,3	14,7	14,1	15,0

Применяя критерий Спирмена и критерий хи-квадрат, проверьте гипотезу о том, что возраст сотрудника и величина его заработной платы независимы. Уровень значимости считайте равным 0,05. Прокомментируйте полученные результаты.

Решение. Пусть данные о X и Y представляются двумерной выборкой $\mathbb{W}_n$ объема $n = 30$ с элементами $(X_1, Y_1), \ldots, (X_n, Y_n)$, соответствующей неизвестному непрерывному распределению $F_W(x, y)$, где $W = (X, Y)^T$.

Проверим гипотезу H_0 вида (9.1) о независимости СВ X и Y, используя критерий Спирмена. Для того чтобы вычислить коэффициент ранговой корреляции (9.13), составим таблицу рангов (табл. 9.8), в которой r_i — реализации рангов X_i в выборке $[X_1, \ldots, X_n]^\top$, а s_i — реализации рангов Y_i в выборке $[Y_1, \ldots, Y_n]^\top$, $i = 1, \ldots, n$.

Таблица 9.8

r_i	1	2	3	4	5	6	7	8	9	10
s_i	1	2	4	3	9	10	6	17	16	14
r_i	11	12	13	14	15	16	17	18	19	20
s_i	28	23	20	27	30	29	26	21	24	22
r_i	21	22	23	24	25	26	27	28	29	30
s_i	19	25	18	12	15	13	11	7	5	8

Вычислив $\sum\limits_{i=1}^{30}(r_i - s_i)^2 = 0 + 0 + 1 + 1 + 4^2 + \ldots + 24^2 + 22^2 = 3530$, найдем реализацию статистики

$$\widehat{\rho}_{XY}(n) = 1 - \frac{6 \cdot 3530}{30^3 - 30} = 0{,}215.$$

Критическая область критерия Спирмена уровня значимости $\alpha = 0{,}05$ при альтернативе H_A: $\tau_{XY} \neq 0$ имеет вид $[-1; -0{,}362) \cup$ $\cup (0{,}362; 1]$, где значения $-0{,}362$ и $0{,}362$ — квантили уровня 0,025 и 0,975, соответственно, распределения коэффициента ранговой корреляции при справедливости гипотезы H_0. Так как реализация статистики $\widehat{\rho}_{XY}(n)$ не попала в критическую область, то на уровне значимости 0,05 принимается гипотеза о независимости СВ X и Y.

Применим критерий хи-квадрат. Разобьем множество значений СВ X на $m = 3$ непересекающихся интервала:

$$\Delta_{X,1} = [0; 35),\ \Delta_{X,2} = [35; 55),\ \Delta_{X,3} = [55; 100],$$

а множество значений СВ Y на $k = 2$ интервала

$$\Delta_{Y,1} = [0; 21\,000),\ \Delta_{Y,2} = [21\,000; +\infty).$$

В прямоугольник $\Delta_{X,1} \times \Delta_{Y,1}$ попало 7 реализаций выборки $\mathbb{W}_{30}$, т.е. $n_{11} = 7$. Далее, $n_{12} = 2$, $n_{21} = 1$, $n_{22} = 12$, $n_{31} = 6$, $n_{32} = 2$.

Соответственно, $n_{1\cdot} = \sum\limits_{j=1}^{2} n_{1j} = 9$, $n_{2\cdot} = 13$, $n_{3\cdot} = 8$, $n_{\cdot 1} = \sum\limits_{i=1}^{3} n_{i1} = 14$, $n_{\cdot 2} = 16$.

Вычислим реализацию стаитистики хи-квадрат вида (9.17)

$$\widehat{\chi}_n^2 = 30\left[\frac{7^2}{14\cdot 9} + \frac{2^2}{9\cdot 19} + \frac{1^2}{13\cdot 14} + \frac{12^2}{16\cdot 13} + \frac{6^2}{8\cdot 14} + \frac{2^2}{16\cdot 8} - 1\right] = 14{,}03.$$

При справедливости гипотезы H_0 статистика (9.17) имеет распределение $\mathcal{H}_r$ с $r = (m-1)(k-1) = 2$. Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид: $(k_{0,95}(2); +\infty)$, где $k_{0,95}(2) = 5{,}99$ — квантиль уровня 0,95 распределения $\mathcal{H}_2$. Таким образом, реализация статистики критерия попала в критическую область. Следовательно, на уровне значимости 0,05 гипотеза о независимости СВ X и Y отвергается.

При применении критериев хи-квадрат и Спирмена для проверки гипотезы о независимости СВ X и Y были получены разные выводы о справедливости этой гипотезы. Этот факт можно объяснить следующим образом. Критерий хи-квадрат является состоятельным для альтернативы общего вида (9.6), т.е. этот критерий способен «улавливать» зависимость любого типа. Критерий Спирмена является состоятельным только для альтернатив вида $\tau_{XY} \neq \neq 0$, $\tau_{XY} < 0$, $\tau_{XY} > 0$, т.е. способен обнаружить лишь монотонную зависимость между СВ. В данном примере монотонной зависимости между возрастом и величиной заработной платы человека нет. Однако представленные данные позволяют заметить следующие тенденции: самые высокие заработки имеют сотрудники среднего возраста, а относительно невысокие — сотрудники молодого и пожилого возраста. ■

Пример 9.5. Имеются следующие данные о специализации и поле 900 английских студентов (табл. 9.9).

Таблица 9.9

Специализация	Пол		Всего
	М	Ж	
Искусствоведение	165	185	350
Естественные науки	168	92	260
Социально-экономические науки	115	105	220
Музыка	32	38	70
Всего	480	420	900

Выясните, имеется ли зависимость между выбранной специальностью и полом студента. Оцените силу связи.

Решение. Каждый объект (респондент) в данной задаче характеризуется двумя признаками. Пусть признак A — выбранная специализация, а признак B — пол студента. Тогда A имеет градации: A_1 — искусствоведение, A_2 — естественные науки, A_3 — социально-экономические науки, A_4 — музыка; а B — градации B_1 — мужской и B_2 — женский.

Рассмотрим полиномиальную СВ X_A с параметрами $(1; p_{1\cdot}, \ldots, p_{4\cdot})$, где $p_{i\cdot}$ — вероятность появления в опыте события A_i, $i = 1, \ldots, 4$ и полиномиальную СВ X_B с параметрами $(1; p_{\cdot 1}, p_{\cdot 2})$, где $p_{\cdot 1}$, $p_{\cdot 2}$ — вероятности появления событий B_1 и B_2. Обозначим p_{ij}, $i = 1, \ldots, 4$, $j = 1, 2$ — вероятность одновременного появления в опыте событий A_i и B_j. Тогда проверка гипотезы о независимости признаков A и B эквивалентна проверке гипотезы H_0 о независимости СВ X_A и X_B. А именно,

$$H_0: \ p_{ij} = p_{i\cdot} p_{\cdot j} \ \text{ для } \forall i = 1, \ldots, 4, j = 1, 2.$$

Альтернативная гипотеза имеет вид (9.21).

Для проверки гипотезы H_0 используем критерий хи-квадрат, статистика которого имеет вид (9.20)

$$\widehat{\chi}_n^2 = n \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{n_{i\cdot} n_{\cdot j}}.$$

Согласно представленной таблице сопряженности (табл. 9.9), реализации частот принимают следующие значения

$$\frac{n_{1\cdot} n_{\cdot 1}}{n} = \frac{350 \cdot 480}{900} = 186{,}7, \quad \frac{n_{1\cdot} n_{\cdot 2}}{n} = \frac{350 \cdot 420}{900} = 163{,}3,$$
$$\frac{n_{2\cdot} n_{\cdot 1}}{n} = 138{,}7, \quad \frac{n_{2\cdot} n_{\cdot 2}}{n} = 121{,}3, \quad \frac{n_{3\cdot} n_{\cdot 1}}{n} = 117{,}3,$$
$$\frac{n_{3\cdot} n_{\cdot 2}}{n} = 102{,}7, \quad \frac{n_{4\cdot} n_{\cdot 1}}{n} = 37{,}3, \quad \frac{n_{4\cdot} n_{\cdot 2}}{n} = 32{,}7,$$

а реализация статистики

$$\widehat{\chi}_n^2 = 2{,}52 + 6{,}19 + 0{,}05 + 0{,}75 + 2{,}88 + 7{,}08 + 0{,}05 + 0{,}86 = 20{,}38.$$

При справедливости гипотезы H_0 статистика хи-квадрат имеет распределение $\mathcal{H}_r$ с $r = (k-1)(m-1) = 3$ степенями свободы. Выберем уровень значимости $\alpha = 0{,}05$, тогда критическая область имеет вид: $\left(k_{0,95}(3); +\infty\right) = (7{,}81; +\infty)$. Реализация статистики попадает в критическую область. Следовательно, гипотеза о независимости признаков A (выбранная специализация) и B (пол студента) отвергается.

Оценим силу связи между признаками A и B с помощью коэффициентов Пирсона и Крамера:

$$P = \sqrt{\frac{\widehat{\chi}_n^2}{\widehat{\chi}_n^2 + n}} = 0{,}149, \quad C = \sqrt{\frac{\widehat{\chi}_n^2}{n \cdot \min\{(k-1), (m-1)\}}} = 0{,}15.$$

Значения этих коэффициентов близки к нулю, что говорит о достаточно слабой силе выявленной связи. ■

Пример 9.6. В 2009 г. центром исследования гражданского общества и некоммерческого сектора НИУ ВШЭ была сформирована репрезентативная выборка из 2000 респондентов. Среди ста вопросов анкеты были, в частности, такие:

1) какое из шести перечисленных описаний точнее всего соответствует материальному положению вашей семьи;

2) удовлетворены ли вы своим здоровьем.

На первый вопрос предлагались ответы:

— денег не хватает даже на питание (категория A_1);

— на питание денег хватает, но не хватает на покупку одежды и обуви (категория A_2);

— на покупку одежды и обуви денег хватает, но не хватает на покупку бытовой техники (категория A_3);

— денег вполне хватает на покупку крупной бытовой техники, но не можем купить новый автомобиль (категория A_4);

— денег хватает на все, кроме таких дорогих приобретений, как квартира, дом (категория A_5);

— материальных затруднений не испытываем, при необходимости могли бы приобрести квартиру, дом (категория A_6).

Ответы на второй вопрос: удовлетворен (категория B_1) и не удовлетворен (категория B_2). Результаты опроса представлены в таблице сопряженности (табл. 9.10) признаков A (материальное положение семьи) и B (удовлетворенность состоянием своего здоровья).

Таблица 9.10

	B_1	B_2	
A_1	83	154	237
A_2	278	354	632
A_3	470	299	769
A_4	204	76	280
A_5	46	20	66
A_6	13	3	16
	1094	906	2000

Оцените меры прогноза Гутмана λ_B и λ_A, постройте асимптотические доверительные интервалы уровня надежности 0,95 этих мер.

Решение. Оценкой меры прогноза Гутмана λ_B является

$$\widehat{\lambda}_B = \frac{\sum\limits_{i=1}^{m} \max\limits_{1\leqslant j\leqslant k} n_{ij} - \max\limits_{1\leqslant j\leqslant k} n_{\cdot j}}{n - \max\limits_{1\leqslant j\leqslant k} n_{\cdot j}}, \quad k=2,\ m=6.$$

Согласно табл. 9.10, максимальное значение сумм реализаций по столбцам имеет первый столбец, т.е. $\max\limits_{1\leqslant j\leqslant 2} n_{\cdot j} = n_{\cdot 1} = 1094$, а $\sum\limits_{i=1}^{6} \max\limits_{1\leqslant j\leqslant 2} n_{ij} = 154+354+470+204+46+13 = 1241$. Тогда реализация оценки

$$\widehat{\lambda}_B = \frac{1241-1094}{2000-1094} = 0{,}168.$$

Аналогично, оценка меры прогноза Гутмана λ_A есть

$$\widehat{\lambda}_A = \frac{\sum\limits_{j=1}^{k} \max\limits_{1\leqslant i\leqslant m} n_{ij} - \max\limits_{1\leqslant i\leqslant m} n_{i\cdot}}{n - \max\limits_{1\leqslant i\leqslant m} n_{i\cdot}}, \quad k=2,\ m=6.$$

По табл. 9.10 находим $\max\limits_{1\leqslant i\leqslant 6} n_{i\cdot} = n_{3\cdot} = 769$, а $\sum\limits_{j=1}^{2} \max\limits_{1\leqslant i\leqslant 6} n_{ij} = 470 + {}$ $+354 = 824$. Реализация оценки

$$\widehat{\lambda}_A = \frac{824-769}{2000-769} = 0{,}045.$$

Оценка для симметричной меры прогноза λ будет

$$\widehat{\lambda} = \frac{\widehat{\lambda}_A + \widehat{\lambda}_B}{2} = 0{,}107.$$

Построенные оценки позволяют сказать, что прогноз модальной (наиболее вероятной) категории признака B (удовлетворенность состоянием своего здоровья) улучшится на 16,8%, если при прогнозировании будет учтено значение признака A (материальное положение семьи), а прогноз модальной категории признака A улучшится на 4,5%, если при прогнозировании будет учтено значение признака B.

Построим теперь асимптотический доверительный интервал надежности 0,95 для $\widehat{\lambda}_B$, пользуясь тем, что статистика (9.26) имеет асимптотически нормальное распределение. Тогда

$$\mathbf{P}\left\{\widehat{\lambda}_B - u_{0{,}975}\sqrt{\frac{\left[n - \sum\limits_{i=1}^{m} \max\limits_{1\leqslant j\leqslant k} n_{ij}\right] Q_B}{(n - \max\limits_{1\leqslant j\leqslant k} n_{\cdot j})^3}} \leqslant \lambda_B \leqslant \right.$$

$$\left. \leqslant \widehat{\lambda}_B + u_{0{,}975}\sqrt{\frac{\left[n - \sum\limits_{i=1}^{m} \max\limits_{1\leqslant j\leqslant k} n_{ij}\right] Q_B}{(n - \max\limits_{1\leqslant j\leqslant k} n_{\cdot j})^3}}\right\} = 0{,}95,$$

где $Q_B = \left[\sum\limits_{i=1}^{m} \max\limits_{1\leqslant j\leqslant k} n_{ij} + \max\limits_{1\leqslant j\leqslant k} n_{\cdot j} - 2\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k} n_{ij}\delta_{ij}\delta_j\right]$, а δ_{ij} и δ_j определены в (9.27) и (9.28). Найдем $s = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k} n_{ij}\delta_{ij}\delta_j$. Так как $n_{\cdot 1}$ является максимумом среди $n_{\cdot 1}$ и $n_{\cdot 2}$, то $\delta_1 = 1$, $\delta_2 = 0$ и $s = \sum\limits_{i=1}^{m} n_{i1}\delta_{i1}$. Максимальными значениями n_{ij} в строках будут значения n_{12}, n_{22}, n_{31}, n_{41}, n_{51} и n_{61}. Среди этих значений $n_{31} = 470$, $n_{14} = 204$, $n_{51} = 46$ и $n_{61} = 13$ находятся в первом столбце, т.е. $\delta_{31} = \delta_{41} = \delta_{51} = \delta_{61} = 1$, а $\delta_{11} = \delta_{21} = 0$. Таким образом, $s = 470 + 204 + 46 + 13 = 733$. Теперь

$$\mathbf{P}\left\{0{,}168 - 1{,}96\sqrt{\frac{(2000-1241)(1241+1094-2\cdot 733)}{(2000-1094)^3}} \leqslant \lambda_B \leqslant \right.$$
$$\left. \leqslant 0{,}168 + 1{,}96\sqrt{\frac{(2000-1241)(1241+1094-2\cdot 733)}{(2000-1094)^3}}\right\} = 0{,}95.$$

$$\mathbf{P}\left\{0{,}168 - 1{,}96\cdot 0{,}03 \leqslant \lambda_B \leqslant 0{,}168 + 1{,}96\cdot 0{,}03\right\} =$$
$$= \mathbf{P}\{0{,}109 \leqslant \lambda_B \leqslant 0{,}227\} = 0{,}95.$$

Аналогично, для коэффициента λ_A получим, что

$$\mathbf{P}\left\{\widehat{\lambda}_A - u_{0{,}975}\sqrt{\frac{\left[n - \sum\limits_{j=1}^{k}\max\limits_{1\leqslant i\leqslant m} n_{ij}\right] Q_A}{\left[n - \max\limits_{1\leqslant i\leqslant m} n_{i\cdot}\right]^3}} \leqslant \lambda_A \leqslant \right.$$
$$\left. \leqslant \widehat{\lambda}_A + u_{0{,}975}\sqrt{\frac{\left[n - \sum\limits_{j=1}^{k}\max\limits_{1\leqslant i\leqslant m} n_{ij}\right] Q_A}{\left[n - \max\limits_{1\leqslant i\leqslant m} n_{i\cdot}\right]^3}}\right\} = 0{,}95,$$

где $Q_A = \left[\sum\limits_{j=1}^{k} \max\limits_{1\leqslant i\leqslant m} n_{ij} + \max\limits_{1\leqslant i\leqslant m} n_{i\cdot} - 2\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{m} n_{ij}\delta_{ij}\delta_i\right]$, а δ_{ij} и δ_i определены в (9.30) и (9.31).

Вычислив $\sum\limits_{j=1}^{2}\sum\limits_{i=1}^{6} n_{ij}\delta_{ij}\delta_i = \sum\limits_{j=1}^{2} n_{3j}\delta_{3j} = n_{31} = = 470$, получим

$$\mathbf{P}\left\{0{,}045 - u_{0{,}975}\sqrt{\frac{(2000-824)(824+769-2\cdot 470)}{(2000-769)^3}} \leqslant \lambda_A \leqslant \right.$$
$$\left. \leqslant 0{,}045 - u_{0{,}975}\sqrt{\frac{(2000-824)(824+769-2\cdot 470)}{(2000-769)^3}}\right\} = 0{,}95;$$

$$\mathbf{P}\{0{,}045 - 1{,}96\cdot 0{,}02 \leqslant \lambda_A \leqslant 0{,}045 + 1{,}96\cdot 0{,}02\} =$$
$$= \mathbf{P}\{0{,}005 \leqslant \lambda_A \leqslant 0{,}085\} = 0{,}95. \ \blacksquare$$

Пример 9.7. В некоторой области Англии исследовалось влияние погоды на урожай (данные взяты из монографии [21]). Рассматривалось три показателя: урожай сена в центнерах на акр (X_1), весеннее количество осадков в дюймах (X_2) и накопленная за весну температура выше 42°F (X_3). По данным двадцатилетних наблюдений были вычислены реализации выборочных коэффициентов корреляции: $\widehat{r}_{X_1X_2} = 0{,}8$, $\widehat{r}_{X_1X_3} = -0{,}4$, $\widehat{r}_{X_2X_3} = = -0{,}56$. Оцените частные коэффициенты корреляции $r_{12;3}$, $r_{13;2}$ и $r_{23;1}$. Прокомментируйте полученный результат.

Решение. Пусть K_X — корреляционная матрица вектора $X = (X_1, X_2, X_3)^\top$. Пользуясь определением 9.5, найдем частные коэффициенты корреляции $r_{12;3}$, $r_{12;2}$ и $r_{23;1}$:

$$r_{12;3} = \frac{-K_{12}}{\sqrt{K_{11}K_{22}}} = \frac{-(r_{12}-r_{13}r_{23})(-1)^{1+2}}{\sqrt{(1-r_{23}^2)(1-r_{13}^2)}} = \frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{23}^2)(1-r_{13}^2)}},$$
$$r_{13;2} = \frac{-K_{13}}{\sqrt{K_{11}K_{33}}} = \frac{-(r_{12}r_{23}-r_{13})(-1)^{1+3}}{\sqrt{(1-r_{23}^2)(1-r_{12}^2)}} = \frac{r_{13}-r_{12}r_{23}}{\sqrt{(1-r_{23}^2)(1-r_{12}^2)}},$$
$$r_{23;1} = \frac{-K_{23}}{\sqrt{K_{22}K_{33}}} = \frac{-(r_{23}-r_{12}r_{13})(-1)^{2+3}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}} = \frac{r_{23}-r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}.$$

Заменяя в полученных формулах неизвестные коэффициенты корреляции r_{ij}, $1 \leqslant i < j \leqslant 3$ их выборочными оценками $\widehat{r}_{ij}$, получаем оценки частных коэффициентов корреляции

$$\widehat{r}_{12;3} = \frac{\widehat{r}_{12}-\widehat{r}_{13}\widehat{r}_{23}}{\sqrt{(1-\widehat{r}_{13}^2)(1-\widehat{r}_{23}^2)}} = \frac{0{,}8-(-0{,}4)(-0{,}56)}{\sqrt{(1-0{,}4^2)(1-0{,}56^2)}} = 0{,}759,$$

$$\widehat{r}_{13;2} = \frac{\widehat{r}_{13}-\widehat{r}_{12}\widehat{r}_{23}}{\sqrt{(1-\widehat{r}_{12}^2)(1-\widehat{r}_{23}^2)}} = \frac{-0{,}4-0{,}8(-0{,}56)}{\sqrt{(1-0{,}8^2)(1-0{,}56^2)}} = 0{,}097,$$

$$\widehat{r}_{23;1} = \frac{\widehat{r}_{23} - \widehat{r}_{12}\widehat{r}_{13}}{\sqrt{(1-\widehat{r}_{12}^2)(1-\widehat{r}_{13}^2)}} = \frac{-0{,}56 - 0{,}8(-0{,}4)}{\sqrt{(1-0{,}8^2)(1-0{,}4^2)}} = -0{,}436.$$

Оценки парных коэффициентов корреляции показывают, что урожайность X_1 и количество осадков X_2 имеют положительную корреляцию, а урожайность X_1 и накопленные температуры X_3 — отрицательную. Последнее можно интепретировать как неблагоприятное влияние высоких температур на урожай. Однако оценка частного коэффициента корреляции $\widehat{r}_{13;2}$ между урожайностью и накопленными температурами при фиксированном количестве осадков оказывается положительной. Это означает, что существует положительная связь между урожаем и температурой, когда влияние осадков устранено. Отрицательное значение $\widehat{r}_{13}$ является следствием воздействия фактора осадков. ■

П р и м е р 9.8. Автосалон предоставил сведения о продажной цене ξ_1, ширине ξ_2, длине ξ_3 и массе ξ_4 автомобиля. За последний месяц было продано 34 автомобиля. На основании этих данных вычислены выборочные коэффициенты корреляции: $\widehat{r}_{12}(n) = 0{,}33$, $\widehat{r}_{13}(n) = 0{,}16$, $\widehat{r}_{14}(n) = 0{,}53$, $\widehat{r}_{23}(n) = 0{,}71$, $\widehat{r}_{24}(n) = 0{,}72$, $\widehat{r}_{34}(n) = 0{,}63$. Оцените множественный коэффициент корреляции $R_{1(2,3,4)}$ между продажной ценой автомобиля и совокупностью его трех технических характеристик, описывающих длину, высоту и массу. Проверить гипотезу о том, что $R_{1(2,3,4)} = 0$, предполагая, что данные имеют гауссовское распределение. Прокомментируйте полученный результат.

Р е ш е н и е. Пусть выборки $X_1 = [X_{11}, \ldots, X_{n1}]^\top$, $X_2 = [X_{12}, \ldots, X_{n2}]^\top$, $X_3 = [X_{13}, \ldots, X_{n3}]^\top$, $X_4 = [X_{14}, \ldots, X_{n4}]^\top$ объема $n = 34$ порождены СВ ξ_1, ξ_2, ξ_3 и ξ_4 соответственно, которые являются компонентами гауссовского вектора $\xi = (\xi_1, \ldots, \xi_4)^\top$.

Оценкой множественного коэффициента корреляции между СВ ξ_1 и совокупностью ξ_2, ξ_3, ξ_4 является

$$\widehat{R}_{1(2\ldots4)} = \sqrt{1 - \frac{\det \widehat{K}_\xi}{\widehat{K}_{11}}},$$

где $\widehat{K}_\xi$ — матрица, составленная из выборочных коэффициентов корреляции $\widehat{r}_{ij}(n)$, $i, j = 1, \ldots, 4$; $\widehat{K}_{11}$ — алгебраическое дополнение элемента $(1,1)$ матрицы $\widehat{K}_\xi$.

По условию реализация матрицы $\widehat{K}_\xi$ имеет вид

$$\widehat{K} = \begin{bmatrix} 1 & 0{,}33 & 0{,}16 & 0{,}53 \\ 0{,}33 & 1 & 0{,}71 & 0{,}72 \\ 0{,}16 & 0{,}71 & 1 & 0{,}63 \\ 0{,}53 & 0{,}72 & 0{,}63 & 1 \end{bmatrix}.$$

Вычисляя $\widehat{K}_{11} = \begin{vmatrix} 1 & 0{,}71 & 0{,}72 \\ 0{,}71 & 1 & 0{,}63 \\ 0{,}72 & 0{,}63 & 1 \end{vmatrix} = 0{,}22$, и $\det \widehat{K}_{\xi} = 0{,}15$, получим что $\widehat{R}^2_{1(2,3,4)} = 1 - \dfrac{0{,}15}{0{,}22} = 0{,}32$, а $\widehat{R}_{1(2,3,4)} = 0{,}57$.

Проверим гипотезу H_0: $R_{1(2,3,4)} = 0$ против альтернативы H_A: $R_{1(2\ldots4)} > 0$, используя критерий, основанный на статистике (9.36).

Реализация статистики (9.36)

$$\widehat{F} = \frac{\frac{1}{3}0{,}32}{\frac{1}{30}(1-0{,}32)} = 4{,}706$$

При справедливости H_0 статистика (9.36) будет иметь F-распределение с $l-1 = 3$ и $n-l = 30$ степенями свободы. Критическая область уровня значимости $\alpha = 0{,}05$ критерия со статистикой (9.36) имеет вид $(8{,}617; +\infty)$, где 8,617 есть квантиль уровня 0,95 распределения $F(3;30)$. Реализация статистики не попадает в критическую область, следовательно, на уровне значимости 0,05 гипотеза H_0: $R_{1(2,3,4)} = 0$ принимается. То есть можно считать, что статистическая связь между продажной ценой автомобиля и совокупностью таких его технических характеристик, как длина, ширина и масса, отсутствует. ■

Пример 9.9. Три квалифицированных эксперта (A, B и C) проранжировали в порядке предпочтения семь представленных бизнес-проектов. Результаты представлены в табл. 9.11.

Таблица 9.11

	1	2	3	4	5	6	7
A	1	4	2	5	3	7	6
B	2	1	3	4	5	6	7
C	2	1	4	5	3	7	6

Можно ли считать, что данная экспертная группа обладает общей системой предпочтений?

Решение. Пусть СВ ξ_j, $j = 1,2,3$ — оценка качества представленных бизнес-проектов, согласно оценке j-го эксперта, а $\{R_{11}, \ldots, R_{n1}\}$, $\{R_{12}, \ldots, R_{n2}\}$, $\{R_{13}, \ldots, R_{n3}\}$, $n = 7$ — ранжировки выборок $X_1 = [X_{11}, \ldots, X_{n1}]^T$, $X_2 = [X_{12}, \ldots, X_{n2}]^T$, $X_3 = [X_{13}, \ldots, X_{n3}]^T$, порожденных СВ ξ_1, ξ_2 и ξ_3.

Из справедливости гипотезы вида (9.37) о независимости СВ ξ_1, ξ_2 и ξ_3 будет следовать, в частности, тот факт, что при каждой ранжировке выборки X_1 вероятность появления любого набора рангов выборки X_2 или выборки X_3 будет равна $\frac{1}{n!}$. То есть мнения экспертов рассогласованы.

Проверим гипотезу вида (9.37), используя критерий основанный на коэффициенте конкордации Кендалла (9.38) $\widehat{W}_n(m) = \dfrac{12}{m^2(n^3-n)} \sum\limits_{i=1}^{n} \left[\sum\limits_{j=1}^{m} R_{ij} - \dfrac{m(n+1)}{2} \right]^2$.

Для вычисления реализации статистики $\widehat{W}_n(m)$ определим сначала реализации $r_{i\cdot}$ статистик $\sum\limits_{j=1}^{m} R_{ij}$, $m = 3$, $i = 1,\ldots,7$. Имеем $r_{1\cdot} = \sum\limits_{j=1}^{3} r_{ij} = 1+2+2 = 5$, $r_{2\cdot} = 4+1+1 = 6$, $r_{3\cdot} = 2+3+4 = 9$, $r_{4\cdot} = 14$, $r_{5\cdot} = 11$, $r_{6\cdot} = 20$, $r_{7\cdot} = 19$. Тогда реализация коэффициента конкордации

$$\widehat{W}_n(3) = \frac{12}{3^2(7^3-7)} \left((5-12)^2 + \ldots + (19-12)^2\right) = \frac{12}{9(7^3-7)} \cdot 212 = 0{,}84.$$

Реализация коэффициента конкордации близка к единице. Однако для проверки гипотезы (9.37) следует указать критическую область. В [24] представлены квантили уровня 0,95 и 0,99 статистики $S = \sum\limits_{i=1}^{n} \left(\sum\limits_{j=1}^{m} R_{ij} - \dfrac{m(n+1)}{2} \right)^2$ при справедливости гипотезы H_0 вида (9.37). Так для $n = 7$ и $m = 3$ квантиль уровня 0,95 статистики S равна 157,3.

Реализация статистики S, равная 212, попадает в критическую область. Следовательно, гипотеза о независимости СВ ξ_1, ξ_2, ξ_3 отвергается на уровне значимости 0,05. Таким образом, можно считать на уровне значимости 0,05, что группа экспертов согласована, т.е. обладает единой системой предпочтений. ■

9.10. Задачи для самостоятельного решения

1. У каждого из 6800 респондентов измерялось два признака. Признак A (цвет глаз) имеет три градации: карий, серый, зеленый. Признак B (цвет волос) имеет градации: брюнет, шатен, блондин. Данные представлены в табл. 9.12.

Таблица 9.12

	Брюнет	Шатен	Блондин
Карий	1768	806	236
Серый	946	1387	800
Зеленый	115	438	304

Имеется ли статистическая зависимость между признаками A и B на уровне значимости 0,05? Оцените меры прогноза Гутмана λ_A и λ_B.

Ответ: признаки A и B зависимы, $\widehat{\lambda}_A = 0{,}19$, $\widehat{\lambda}_B = 0{,}22$.

2. Спортсмены, ранги которых построении по росту были равны $1, 2, \ldots, 13$, показали в прыжке в длину следующие результаты (в м): 6,94; 7,12; 7,01; 6,98; 7,24; 7,42; 7,13; 6,95; 7,34; 7,82; 7,23; 7,05; 7,15. Имеется ли зависимость между ростом спортсмена и длиной его прыжка? Уровень значимости считайте равным 0,05.

Ответ: гипотеза о независимости между ростом спортсмена и длиной его прыжка не отвергается критерием Спирмена и критерием Кендалла.

3. Докажите равенства (9.8).

4. Докажите, что коэффициент ранговой корреляции $\widehat{\rho}_{XY}(n)$, определенный формулой (9.12), можно преобразовать к виду (9.13).

5. Докажите справедливость формулы (9.34).

6. Докажите, что $\mathbf{M}\{\widehat{\tau}_{XY}\} = \tau_{XY}$.

7. В детской клинике проведено обследование 40 новорожденных определенной категории. Выборочный коэффициент корреляции $\widehat{r}_{XY}(n)$ между ростом (X) и весом (Y) этих новорожденных составил 0,76. На уровне значимости 0,05 проверьте гипотезу о независимости показателей X и Y. Используя z-преобразование Фишера, постройте асимптотический доверительный интервал для коэффициента корреляции r_{XY} уровня надежности 0,95.

Ответ: гипотеза о независимости случайных величин X и Y отвергается, $I = (0{,}58; 0{,}86)$.

8. Имеется 12 одинаковых по размеру дисков, окраска которых отличается тоном — от светло-голубого до темно-синего. Для того чтобы оценить способность модельера одежды различать цветовые оттенки, ему предлагают расположить диски в порядке увеличения степени интенсивности цвета. Модельер установил следующий порядок дисков: 1, 4, 7, 2, 3, 5, 8, 12, 10, 6, 11, 9. Объективная оценка интенсивности цвета, полученная с помощью колориметрического испытания: $1, 2, \ldots, 12$. Опираясь на эти данные, охарактеризуйте способность модельера различать оттенки цвета.

Ответ: гипотеза о независимости объективного показателя интенсивности цвета и оценки интенсивности, данной модельером, отвергается критерием Спирмена и критерием Кендалла. Значения $\widehat{\rho}_{XY}(n) = 0{,}748$ и $\widehat{\tau}_{XY}(n) = 0{,}58$ свидетельствуют о достаточно хороших способностях модельера правильно различать оттенки.

9. В табл. 9.13 приведены 2663 случая, классифицированных по двум признакам: A — наличие прививки против холеры, B — отсутствие заболевания. Проверьте гипотезу о независимости признаков A и B на уровне значимости 0,05.

Таблица 9.13

	B	$\overline{B}$
A	1625	5
$\overline{A}$	1022	11

Ответ: гипотеза отвергается.

10. Судейская коллегия из шести человек оценивает выступления пяти фигуристов, вышедших в финал соревнований. Результаты представлены в табл. 9.14.

Таблица 9.14

Судья	Спортсмен				
	A	B	C	D	E
1	1	2	3	4	5
2	1	3	4	2	5
3	4	3	2	1	5
4	1	2	3	4	5
5	2	1	3	4	5
6	5	4	3	2	1

Вычислите коэффициент конкордации Кендалла. Можно ли считать на уровне значимости 0,05, что данная судейская коллегия обладает единой системой предпочтений?

Указание. Квантиль уровня 0,95 статистики $\widehat{W}_5(6)$ при справедливости гипотезы H_0 вида (9.37) равна 0,378.

Ответ: $\widehat{W}_5(6) = 0{,}24$. Данная судейская коллегия не обладает единой системой предпочтений.

11. По имеющимся данным об экономических показателях X_1, X_2 и X_3 для 20 регионов России вычислены выборочные коэффициенты корреляции $\widehat{r}_{X_1X_2} = 0{,}746$, $\widehat{r}_{X_1X_3} = 0{,}507$, $\widehat{r}_{X_2X_3} = 0{,}432$. Предполагается, что наблюдения имеют гауссовское распределение. На уровне значимости 0,05 проверьте гипотезу о независимости показателей X_1 и X_3, используя информацию только о коэффициенте $\widehat{r}_{X_1X_3}$. Вычислите частный коэффициент корреляции $\widehat{r}_{13;2}$ между X_1 и X_3 при фиксированном показателе X_2. Проверьте гипотезу H_0: $r_{13;2} = 0$.

Ответ: гипотеза о независимости X_1 и X_3 отвергается; $\widehat{r}_{13;2} = 0{,}3$; гипотеза H_0 принимается.

12. Центром исследования гражданского общества и некоммерческого сектора НИУ ВШЭ была сформирована репрезентативная выборка из 2000 респондентов. Респондентов попросили ответить, удовлетворены ли они своими доходами. Из 1097 участвовавших в опросе женщин 18 затруднились ответить, 839 ответили отрицательно и 240 — положительно. Из 903 мужчин 13 затруднились ответить, 652 ответили отрицательно и 238 — положительно. Можно ли утверждать, опираясь на эти данные, что удовлетворенность своими доходами и пол зависимы? Уровень значимости считайте равным 0,05.

Г Л А В А III

МЕТОДЫ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ

10. Линейная модель множественной регрессии

10.1. Линейная модель регрессии

Пусть X — скалярная переменная, описывающая некоторый экономический или какой-либо другой показатель. Мы предполагаем, что X зависит от p-мерного вектора h некоторых факторов в том смысле, что изменения вектора h вызывают вполне определенные изменения X. Последнее предположение математически можно представить в виде функциональной зависимости

$$X = f(h), \quad X \in \mathbb{R}^1,\ h \in \mathbb{R}^p,$$

где $f(h)$ — некоторая числовая функция, зависящая от p переменных.

Одной из важнейших задач является решение проблемы определения функции $f(\cdot)$, описывающей связь между *объясняемой переменной* X и *объясняющими переменными* h по результатам наблюдений за X и h.

Предположим, что у нас имеется возможность получить совместные наблюдения за X и h в n опытах. Обозначим результаты наблюдений через $\{X_k, h_k\}$, где $k = 1, \ldots, n$. С учетом введенной выше связи между X и h, связь между результатами отдельных наблюдений можно представить в виде

$$X_k = f(h_k) + \varepsilon_k, \quad k = 1, \ldots, n,$$

где ε_k — случайная ошибка k-го наблюдения.

Естественно, без каких-то предварительных договоренностей о возможном виде функции $f(\cdot)$ задача ее восстановления по результатам наблюдений, число которых конечно, представляется практически неразрешимой. Поэтому мы сделаем следующее предположение о виде исследуемой зависимости: функция $f(h)$ линейна по переменным h. Последнее означает, что $f(h)$ можно представить в виде линейной комбинации компонент вектора h с некоторыми неизвестными неслучайными коэффициентами $\theta_1, \ldots, \theta_p$:

$$f(h) = h_1\theta_1 + \cdots + h_p\theta_p = h^\top\theta, \quad \theta = \{\theta_1, \ldots, \theta_p\}^\top \in \mathbb{R}^p.$$

Теперь введенную выше модель наблюдения за h и X можно представить в виде

$$X_k = h_k^\top\theta + \varepsilon_k, \quad k = 1, \ldots, n. \tag{10.1}$$

Модель (10.1) называется моделью *множественной линейной регрессии.*

Относительно случайных ошибок $\{\varepsilon_k\}$ сделаем следующие предположения:

1) при каждом $k = 1, \ldots, n$ случайная ошибка ε_k имеет гауссовское распределение с нулевым математическим ожиданием и дисперсией σ^2, т.е. $\varepsilon_k \sim \mathcal{N}\left(0; \sigma^2\right)$;

2) случайные ошибки $\{\varepsilon_k,\ k = 1, \ldots, n\}$ независимы в совокупности.

Таким образом, мы предполагаем, что ошибки наблюдений образуют последовательность независимых гауссовских случайных величин с параметрами $\left(0; \sigma^2\right)$.

Относительно модели (10.1) в рамках принятых выше предположений сделаем следующие замечания.

а) восстановление зависимости $X = f(h)$ сводится к оцениванию вектора параметров θ по наблюдениям $\{X_k, h_k\}$, $k = 1, \ldots, n$;

б) предположение о линейной зависимости X от h во многих практически важных случаях выполняется с достаточно высокой точностью. Если же зависимость X от h является существенно нелинейной, то для оценивания $f(h)$ применяются методы нелинейного регрессионного анализа, основы которого изложены в разд. 15;

в) модель (10.1), в которой ошибки наблюдения имеют одинаковые дисперсии, равные σ^2, называется *гомоскедастичной линейной регрессионной моделью.* Гомоскедастичность модели означает, что точности всех наблюдений одинаковы. Предположение о том, что ошибки имеют нулевые математические ожидания, означает отсутствие *систематической погрешности* наблюдений. *Гетероскедастичные модели*, в которых дисперсии ошибок наблюдения различны, т.е. $\mathbf{D}\{\varepsilon_k\} = \sigma_k^2$, $k = 1, \ldots, n$, будут рассмотрены далее в разд. 11, 12;

г) из предположения о независимости ошибок наблюдений следует, что они некоррелированы: $\mathbf{cov}(\varepsilon_k, \varepsilon_m) = 0$, если $k \neq m$. Это условие также можно ослабить, если воспользоваться некоторой дополнительной математической моделью, описывающей корреляционные зависимости в рядах наблюдений. Соответствующие обобщения будут рассмотрены далее в разд. 11.2 и 12.

Представим теперь модель линейной регрессии (10.1) в векторно-матричном виде. Для этого введем обозначения: $Z_n = \{X_1, \ldots, X_n\}^\top$,

$E_n = \{\varepsilon_1, \ldots, \varepsilon_n\}^\top$, H_n — матрица размера $(n \times p)$, k-й строкой которой является $h_k^\top$.

Вектор Z_n результатов наблюдений за объясняемой переменной является *неоднородной выборкой*, а H_n называется обычно *регрессионной матрицей* (*матрицей плана*). Заметим, что случайный вектор E_n ошибок наблюдений имеет гауссовское n-мерное распределение с параметрами $(0; \sigma^2 I_n)$: $E_n \sim \mathcal{N}(0; \sigma^2 I_n)$, где I_n — единичная матрица размера $(n \times n)$. Используя введенные обозначения, представим модель (10.1) в окончательном виде:

$$Z_n = H_n \theta + E_n. \tag{10.2}$$

Определение 10.1. Модель (10.2), удовлетворяющая всем сделанным выше предположениям, называется *моделью линейной множественной гауссовской регрессии* порядка p.

В заключение заметим, что порядок p равен числу неизвестных параметров $\{\theta_1, \ldots, \theta_p\}$ в модели (10.2), подлежащих оцениванию по выборке Z_n. Для случая $p = 2$ обычно рассматривается линейная регрессия следующего вида:

$$X_k = \theta_1 + h_k \theta_2 + \varepsilon_k, \quad k = 1, \ldots, n,$$

где $\{h_k\}$ — наблюдения за одной объясняющей переменной. Такая модель обычно называется *простой линейной регрессией*. Если же для описания линейной регрессии используется термин «множественная», то это означает, что порядок модели $p \geqslant 3$ (т.е. объясняющих переменных не менее двух).

10.2. Метод наименьших квадратов

Рассмотрим задачу оценивания по выборке Z_n вектора $Y = L\theta$, где L — заданная неслучайная матрица размера $(q \times p)$. Для построения оценки $\widehat{Y}_n$ вектора Y воспользуемся методом максимального правдоподобия. По предположению вектор наблюдений Z_n в модели (10.2) имеет гауссовское распределение:

$$Z_n \sim \mathcal{N}(H_n \theta; \sigma^2 I_n). \tag{10.3}$$

Теорема 10.1. *Пусть матрица $W_n = H_n^\top H_n$ — невырожденная. Тогда оценка метода максимального правдоподобия (МП-оценка) $\widehat{Y}_n$ вектора Y по выборке Z_n имеет вид*

$$\widehat{Y}_n = L \widehat{\theta}_n, \tag{10.4}$$

где $\widehat{\theta}_n$ — оценка вектора θ вида

$$\widehat{\theta}_n = W_n^{-1} H_n^\top Z_n. \tag{10.5}$$

Из доказательства теоремы 10.1 (см. пример 10.1) следует, что $\widehat{\theta}_n$ есть точка минимума квадратичной функции потерь:

$$\widehat{\theta}_n = \arg\min_{\theta} |Z_n - H_n\theta|^2. \tag{10.6}$$

Определение 10.2. Оценка $\widehat{\theta}_n$, определяемая из условия (10.6) и имеющая вид (10.5), называется *оценкой метода наименьших квадратов* (МНК-оценкой) вектора θ в модели линейной регрессии (10.2).

Статистические свойства ошибки $\Delta\widehat{Y}_n = \widehat{Y}_n - Y$ оценки $\widehat{Y}_n$ вектора Y приведены в следующей теореме.

Теорема 10.2. *Оценка $\widehat{Y}_n$ и ее ошибка $\Delta\widehat{Y}_n$ обладают следующими свойствами:*

1) $\mathbf{M}\left\{\widehat{Y}_n\right\} = Y$, $\mathbf{M}\left\{\Delta\widehat{Y}_n\right\} = 0$ — *несмещенность*;

2) $\widehat{Y}_n \sim \mathcal{N}\left(Y; \widehat{K}_{Y_n}\right)$, $\Delta\widehat{Y}_n \sim \mathcal{N}\left(0; \widehat{K}_{Y_n}\right)$, *где ковариационная матрица ошибки $\Delta\widehat{Y}_n$ оценки $\widehat{Y}_n$ равна $\widehat{K}_{Y_n} = \mathbf{cov}\{\Delta\widehat{Y}_n, \Delta\widehat{Y}_n\} = \sigma^2 L W_n^{-1} L^\top$*;

3) *оценка $\widehat{Y}_n$ эффективна по Рао—Крамеру (см. определение 4.5).*

Так как при $L = I$ оценка $\widehat{Y}_n$ превращается в $\widehat{\theta}_n$, то справедливо следующее утверждение.

Следствие 10.1. *Оценка $\widehat{\theta}_n$ вида* (10.5) *имеет распределение $\mathcal{N}\left(\theta; \widehat{K}_n\right)$, где $\widehat{K}_n = \sigma^2 W_n^{-1}$ — ковариационная матрица ее ошибки $\Delta\widehat{\theta}_n = \widehat{\theta}_n - \theta$. МНК-оценка $\widehat{\theta}_n$ — несмещенная и эффективная по Рао—Крамеру.*

Следствие 10.2. *СВ $\xi = \dfrac{1}{\sigma^2}|Z_n - H_n\widehat{\theta}_n|^2$ имеет хи-квадрат распределение с $r = n - p$ степенями свободы: $\xi \sim \mathcal{H}_{n-p}$.*

Если отказаться от предположения о том, что вектор ошибок E_n — гауссовский, то $\widehat{Y}_n$ теряет свойство эффективности, но остается наилучшей (по с.к.-критерию, см. определение 2.7) среди всех линейных несмещенных оценок.

Определение 10.3. Оценка $\widetilde{Y}_n$ вектора Y по наблюдениям Z_n называется *линейной несмещенной оценкой*, если она имеет вид $\widetilde{Y}_n = A_n Z_n$, где A_n — неслучайная матрица размера $(q \times n)$, причем $\mathbf{M}\left\{\widetilde{Y}_n\right\} = Y$.

Теорема 10.3 (Гаусс—Марков). *Пусть $\widehat{K}_{Y_n}$ — ковариационная матрица ошибки оценки $\widehat{Y}_n$ вида* (10.4), (10.5), *а $\widetilde{K}_{Y_n}$ — ковариационная матрица произвольной линейной несмещенной оценки $\widetilde{Y}_n$ вектора Y. Тогда $\widehat{K}_{Y_n} \leqslant \widetilde{K}_{Y_n}$ (т.е. матрица $\widetilde{K}_{Y_n} - \widehat{K}_{Y_n}$ является неотрицательно определенной).*

Таким образом, оценка $\widehat{Y}_n$ является *наилучшей линейной несмещенной оценкой* (*НЛН-оценкой*) вектора Y в модели линейной регрессии (10.2).

10.3. Коэффициент детерминации

Выше мы предположили, что переменная X зависит от h значимым образом. Эту гипотезу, как и всякое базовое предположение, следует подвергнуть проверке, используя имеющиеся экспериментальные данные.

Пусть вектор регрессоров h_k в k-м опыте имеет вид

$$h_k = \{1, h_{k,1}, \ldots, h_{k,p-1}\}^\top.$$

Если $p \geqslant 2$, то предполагается зависимость показателя X от объясняющих переменных h. Если же $p = 1$, то модель наблюдения (10.1) принимает вид

$$X_k = \theta_1 + \varepsilon_k, \quad k = 1, \ldots, n,$$

т.е. зависимости X от h фактически нет. Таким образом, нам достаточно сравнить между собой модели порядков $p = 1$ и $p \geqslant 2$.

Введем следующие обозначения:

$\overline{X}_n = \dfrac{1}{n} \sum\limits_{k=1}^{n} X_k$ — выборочное среднее;

$\widehat{X}_k = h_k^\top \widehat{\theta}_n$ — оценка для значения переменной X в k-м опыте ($\widehat{\theta}_n$ — МНК-оценка вектора θ для $p \geqslant 2$);

$\widehat{\varepsilon}_k = X_k - \widehat{X}_k$ — остаток k-го наблюдения (т.е. оценка ошибки ε_k k-го наблюдения).

Нетрудно проверить [13], что между величинами $\overline{X}_n$, $\{\widehat{X}_k\}$ и $\{\widehat{\varepsilon}_k\}$ имеется зависимость следующего вида:

$$\sum_{k=1}^{n} \left(X_k - \overline{X}_n\right)^2 = \sum_{k=1}^{n} \left(\widehat{X}_k - \overline{X}_n\right)^2 + \sum_{k=1}^{n} \left(\widehat{\varepsilon}_k\right)^2. \tag{10.7}$$

Таким образом, разброс объясняемой переменной около выборочного среднего $\overline{X}_n$ равен сумме разброса, «объясняемого регрессией», и разброса, который объяснить не удалось (остаточного разброса).

Определение 10.4. Величина

$$R^2 = \frac{\sum\limits_{k=1}^{n} \left(\widehat{X}_k - \overline{X}_n\right)^2}{\sum\limits_{k=1}^{n} \left(X_k - \overline{X}_n\right)^2} \tag{10.8}$$

называется *коэффициентом детерминации* регрессии порядка $p \geqslant 2$.

Из (10.7) и (10.8) следует, что

$$R^2 = 1 - \frac{\sum\limits_{k=1}^{n} (\widehat{\varepsilon}_k)^2}{\sum\limits_{k=1}^{n} \left(X_k - \overline{X}_n\right)^2}. \tag{10.9}$$

Так как $\sum\limits_{k=1}^{n} (\widehat{\varepsilon}_k)^2 \leqslant \sum\limits_{k=1}^{n} \left(X_k - \overline{X}_n\right)^2$ в силу того, что модель порядка $p \geqslant 2$ «не хуже» модели порядка $p = 1$, то из (10.9) следует, что для всех $p \geqslant 2$

$$0 \leqslant R^2 \leqslant 1.$$

Из приведенных соотношений следует, что коэффициент детерминации R^2 показывает, в какой степени модель порядка $p \geqslant 2$ «лучше объясняет» величину X, чем тривиальная модель $p = 1$ (т.е. X не зависит от объясняющих переменных): чем ближе R^2 к 1, тем большее превосходство имеет регрессия порядка $p \geqslant 2$ над регрессией порядка $p = 1$. Наоборот, если R^2 близок к нулю, то это означает, что модель линейной регрессии плохо описывает объясняемую переменную X. Заметим также, что близость R^2 к 1 еще не означает, что модель регрессии порядка $p \geqslant 2$ правильно описывает зависимость X от h, т.е. по степени близости R^2 к 1 еще нельзя судить об адекватности принятой модели (10.2).

Чтобы коэффициент детерминации не возрастал с увеличением числа регрессоров p, также рассматривают *несмещенный коэффициент детерминации*:

$$\widetilde{R}^2 = 1 - \frac{n-1}{n-p} \frac{\sum\limits_{k=1}^{n} (\widehat{\varepsilon}_k)^2}{\sum\limits_{k=1}^{n} \left(X_k - \overline{X}_n\right)^2}.$$

10.4. Критерий Фишера

Рассмотрим линейную регрессионную модель $Z_n = H_n\theta + E$ в которой $X \in \mathbb{R}^{n\times p}$, $E \sim \mathcal{N}(0; \sigma I_n)$ и матрица H_n имеет ранг p.

Критерий Фишера применяется для проверки гипотезы

$$H_0:\ \ A\theta = c, \tag{10.10}$$

где $A \in \mathbb{R}^{q\times p}$, $c \in \mathbb{R}^q$ — известные матрица и вектор.

Альтернативной гипотезой является H_A: $A\theta \neq c$.

Статистика критерия Фишера имеет вид:

$$F(Z_n) = \frac{(A\widehat{\theta}_n - c)^\top \left[A({H_n}^\top H_n)^{-1}A^\top\right]^{-1}(A\widehat{\theta}_n - c)}{\frac{q}{n-p}(Z_n - H_n\widehat{\theta}_n)^\top(Z_n - H_n\widehat{\theta}_n)}. \tag{10.11}$$

Теорема 10.4. *Если гипотеза H_0 верна, то статистика $F(Z_n)$ имеет распределение Фишера $F(q; n-p)$ с q и $n-p$ степенями свободы.*

Из теоремы 10.4 следует, что критическая область критерия Фишера имеет вид

$$(f_{1-\alpha}(q; n-p); \infty), \tag{10.12}$$

где α — уровень значимости критерия, $f_{1-\alpha}(q; n-p)$ — квантиль уровня $1-\alpha$ распределения $F(q; n-p)$.

Подробнее с критерием Фишера можно познакомиться в монографии [34].

10.5. Примеры

Пример 10.1. Докажите теорему 10.1.

Решение. Найдем МП-оценку вектора θ. По условию закон распределения наблюдения X_k имеет плотность

$$p_k(x; \theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x - h_k^\top\theta)^2}{2\sigma^2}\right\}.$$

Поэтому функция правдоподобия выборки Z_n принимает вид

$$L_n(\theta; Z_n) = \prod_{k=1}^{n} p_k(X_k; \theta) = (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(X_k - h_k^\top\theta)^2\right\}.$$

Последнее выражение, используя матричные обозначения в (10.2), можно представить следующим образом:

$$L_n(\theta; Z_n) = (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}|Z_n - H_n\theta|^2\right\}. \tag{10.13}$$

Из (10.13) следует, что задача максимизации $L_n(\theta; Z_n)$ по θ эквивалентна минимизации по θ функции $\mathcal{L}_n(\theta) = |Z_n - H_n\theta|^2$. Таким образом, МП-оценку $\widehat{\theta}_n$ можно определить из условия

$$\widehat{\theta}_n = \arg\min_{\theta}\mathcal{L}_n(\theta).$$

Найдем явный вид оценки $\widehat{\theta}_n$.

$$\mathcal{L}_n(\theta) = |Z_n - H_n\theta|^2 = Z_n^\top Z_n - 2\theta^\top H_n^\top Z_n + \theta^\top H_n^\top H_n\theta.$$

Вычислим $\dfrac{\partial \mathcal{L}_n(\theta)}{\partial \theta}$, воспользовавшись следующими правилами матричного дифференцирования: $\dfrac{\partial(\theta^\top A)}{\partial\theta} = A$; $\dfrac{\partial(\theta^\top A\theta)}{\partial\theta} = (A + A^\top)\theta$. Если $A = A^\top$, то $\dfrac{\partial(\theta^\top A\theta)}{\partial\theta} = 2A\theta$.

Итак, $\dfrac{\partial \mathcal{L}_n(\theta)}{\partial \theta} = -2H_n^\top Z_n + 2H_n^\top H_n\theta$.

Необходимое условие экстремума функции $\mathcal{L}_n(\theta)$ дает нам следующие соотношения:

$$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} = 0 \implies H_n^\top H_n\theta = H_n^\top Z_n.$$

Так как матрица $W_n = H_n^\top H_n > 0$ по условию, полученная система уравнений имеет единственное решение:

$$\widehat{\theta}_n = W_n^{-1} H_n^\top Z_n.$$

Заметим, что функция $\mathcal{L}_n(\theta)$ строго выпукла по θ, поэтому $\widehat{\theta}_n$ — единственная точка минимума $\mathcal{L}_n(\theta)$, и, следовательно, единственная МП-оценка вектора θ.

Так как $Y = L\theta$, то МП-оценка для Y имеет вид $\widehat{Y} = L\widehat{\theta}_n$, что непосредственно следует из принципа инвариантности для МП-оценок, согласно которому МП-оценка по Z_n для любой параметрической функции $\varphi(\theta)$ равна $\varphi(\widehat{\theta}_n)$, где $\widehat{\theta}_n$ — оценка для θ по Z_n. ■

Пример 10.2. Пусть в модели наблюдения (10.2) вектор ошибок E_n имеет ковариационную матрицу $K_{E_n} = \sigma^2 I_n$, где $\sigma > 0$. Покажите, что МНК-оценка $\widehat{\theta}_n$ является наилучшей линейной несмещенной оценкой вектора θ.

Решение. Пусть $\widetilde{\theta}_n = A_n Z_n$ — произвольная линейная несмещенная оценка θ. Тогда

$$\mathbf{M}\left\{\widetilde{\theta}_n\right\} = A_n \mathbf{M}\{Z_n\} = A_n\left(H_n\theta + \mathbf{M}\{E_n\}\right) = A_n H_n\theta = \theta.$$

Отсюда $(A_n H_n - I_n)\,\theta = 0$. Последнее в силу произвольности θ означает, что

$$A_n H_n = I.$$

Заметим, что $\widehat{\theta}_n = A_n^0 Z_n$, где $A_n^0 = W_n^{-1} H_n^\top$, причем

$$\mathbf{M}\left\{\widehat{\theta}_n\right\} = W_n^{-1} H_n^\top H_n\theta + W_n^{-1} H_n \mathbf{M}\{E_n\} = W_n^{-1} W_n\theta = \theta.$$

Таким образом, $\widehat{\theta}_n$ — линейная несмещенная оценка. Отсюда следует, что $A_n^0 H_n = I_n$. Обозначим $T_n = A_n - A_n^0$.

$$T_n H_n = (A_n - A_n^0) H_n = A_n H_n - A_n^0 H_n = I_n - I_n = 0.$$

Отсюда $T_n\left(A_n^0\right)^\top = T_n H_n W_n^{-1} = 0$.

Найдем ковариационную матрицу $\widetilde{K}_n$ ошибки $\Delta\widetilde{\theta}_n = \widetilde{\theta}_n - \theta$ оценки $\widetilde{\theta}_n$. Заметим, что $\widetilde{\theta}_n = \theta + A_n E_n$. Отсюда

$$\widetilde{K}_n = \mathbf{M}\Big\{\Delta\widetilde{\theta}_n\left(\Delta\widetilde{\theta}_n\right)^\top\Big\} = \mathbf{M}\Big\{A_n E_n (A_n E_n)^\top\Big\} = \\ = A_n \mathbf{M}\big\{E_n E_n^\top\big\} A_n^\top = \sigma^2 A_n A_n^\top.$$

Учитывая, что $A_n = A_n^0 + T_n$ и $T_n\left(A_n^0\right)^\top = 0$, получаем

$$\widetilde{K}_n = \sigma^2\left(A_n^0 + T_n\right)\left(A_n^0 + T_n\right)^\top = \sigma^2 A_n^0\left(A_n^0\right)^\top + T_n\left(A_n^0\right)^\top + \\ + \left(T_n\left(A_n^0\right)^\top\right)^\top + \sigma^2 T_n (T_n)^\top = \sigma^2 A_n^0\left(A_n^0\right)^\top + \sigma^2 T_n (T_n)^\top \geqslant \widehat{K}_n,$$

где $\widehat{K}_n = \sigma^2 A_n^0\left(A_n^0\right)^\top = \sigma^2 W_n^{-1}$ — ковариационная матрица ошибки $\Delta\widehat{\theta}_n = \widehat{\theta}_n - \theta$ МНК-оценки $\widehat{\theta}_n$, а $\sigma^2 T_n (T_n)^\top \geqslant 0$ при любой величине матрицы T_n. Итак, наименьшей ковариационной матрицей ошибки среди всех линейных несмещенных оценок обладает МНК-оценка $\widehat{\theta}_n$.

Так как $K_{\widetilde{Y}_n} = L\widetilde{K}_n L^\top$, а $K_{\widehat{Y}_n} = L\widehat{K}_n L^\top$, то из $\widehat{K}_n \leqslant \widetilde{K}_n$ немедленно следует, что $K_{\widehat{Y}_n} \leqslant K_{\widetilde{Y}_n}$. ■

Пример 10.3. Пусть задана линейная регрессионная модель первого порядка

$$X_k = \theta h_k + \varepsilon_k, \quad k = 1, \ldots, n,$$

где ε_k — независимые случайные величины с математическим ожиданием $\mathbf{M}\{\varepsilon_k\} = 0$ и известной дисперсией $\mathbf{D}\{\varepsilon_k\} = \sigma_\varepsilon^2$.

Постройте МНК-оценку параметра θ, докажите ее несмещенность и найдите дисперсию ошибки оценки $\widehat{\theta}_n$.

Решение. Найдем оценку $\widehat{\theta}_n$ как решение задачи оптимизации

$$J(\theta) = \sum_{k=1}^{n} (X_k - \theta h_k)^2 \to \min_{\theta}.$$

Из необходимых условий экстремума следует, что $\widehat{\theta}_n$ является решением уравнения

$$\frac{d\,J(\theta)}{d\,\theta} = -2\sum_{k=1}^{n} h_k (X_k - \theta h_k) = 0,$$

т.е. $\widehat{\theta}_n = \dfrac{\sum\limits_{k=1}^{n} h_k X_k}{\sum\limits_{k=1}^{n} h_k^2}$.

Покажем, что найденная оценка является несмещенной:

$$\mathbf{M}\left\{\widehat{\theta}_n\right\} = \mathbf{M}\left\{\frac{\sum\limits_{k=1}^{n} h_k X_k}{\sum\limits_{k=1}^{n} h_k^2}\right\} = \frac{\sum\limits_{k=1}^{n} h_k \mathbf{M}\{\theta h_k + \varepsilon_k\}}{\sum\limits_{k=1}^{n} h_k^2} =$$

$$= \frac{\theta \sum\limits_{k=1}^{n} h_k^2 + \sum\limits_{k=1}^{n} \mathbf{M}\{\varepsilon_k\} \sum\limits_{k=1}^{n} h_k^2}{\sum\limits_{k=1}^{n} h_k^2} = \theta,$$

так как $\mathbf{M}\{\varepsilon_k\} = 0$ для любого $k = 1, \ldots, n$.

Теперь найдем дисперсию ошибки оценки $\Delta\widehat{\theta}_n = \theta - \widehat{\theta}_n$

$$\mathbf{D}\left\{\Delta\widehat{\theta}_n\right\} = \mathbf{D}\left\{\theta - \widehat{\theta}_n\right\} = \mathbf{D}\left\{\theta - \frac{\sum\limits_{k=1}^{n} h_k X_k}{\sum\limits_{k=1}^{n} h_k^2}\right\} = \mathbf{D}\left\{\frac{\sum\limits_{k=1}^{n} h_k X_k}{\sum\limits_{k=1}^{n} h_k^2}\right\} =$$

$$= \frac{\mathbf{D}\left\{\sum\limits_{k=1}^{n} h_k X_k\right\}}{\left(\sum\limits_{k=1}^{n} h_k^2\right)^2} = \frac{\sum\limits_{k=1}^{n} h_k^2 \mathbf{D}\{X_k\}}{\left(\sum\limits_{k=1}^{n} h_k^2\right)^2} = \frac{\sum\limits_{k=1}^{n} h_k^2 \mathbf{D}\{\theta h_k + \varepsilon_k\}}{\left(\sum\limits_{k=1}^{n} h_k^2\right)^2} =$$

$$= \frac{\sum\limits_{k=1}^{n} h_k^2 \sigma_\varepsilon^2}{\left(\sum\limits_{k=1}^{n} h_k^2\right)^2} = \frac{\sigma_\varepsilon^2}{\sum\limits_{k=1}^{n} h_k^2}. \quad \blacksquare$$

П р и м е р 10.4. Линейная функция $\varphi(t;\theta) = \theta_1 + \theta_2 t$ измеряется в дискретные моменты $\{t_k\}$ по схеме

$$X_k = \varphi(t_k;\theta) + \varepsilon_k, \quad k = 1, \ldots, 5.$$

Случайные ошибки измерений $\{\varepsilon_k\}$ — независимые центрированные гауссовские СВ с дисперсией $\sigma^2 = 0{,}01$. Результаты наблюдений $\{x_k\}$ приведены в табл. 10.1.

Таблица 10.1

k	1	2	3	4	5
t_k	0	1	2	3	4
x_k	−1,10	1,15	3,20	4,85	7,10

Найдите реализацию МНК-оценки $\widehat{\varphi}(t;\theta)$ наблюдаемой функции.

Р е ш е н и е. По условию $\varphi(t;\theta) = h^\top(t)\theta$, где $h^\top(t) = \{1;t\}$, $\theta = \{\theta_1, \theta_2\}^\top$. Поэтому $h^\top(t_k) = \{1;t_k\}$. Реализация МНК-оценки имеет вид

$$\widehat{\theta}_n = \left(H_n^\top H_n\right)^{-1} H_n^\top z_n,$$

где

$$z_n = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -1{,}10 \\ 1{,}15 \\ 3{,}20 \\ 4{,}85 \\ 7{,}10 \end{bmatrix}, \quad H_n = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}.$$

Отсюда

$$W_n = H_n^\top H_n = \begin{bmatrix} n & \sum\limits_{k=1}^{n} t_k \\ \sum\limits_{k=1}^{n} t_k & \sum\limits_{k=1}^{n} t_k^2 \end{bmatrix} = \begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix}.$$

$$H_n^\top z_n = \begin{bmatrix} \sum\limits_{k=1}^{n} x_k \\ \sum\limits_{k=1}^{n} t_k x_k \end{bmatrix} = \begin{bmatrix} 15{,}2 \\ 50{,}5 \end{bmatrix}.$$

Реализация МНК-оценки вектора $\theta = [\theta_1; \theta_2]^\top$ имеет вид

$$\widehat{\theta}_n = W_n^{-1} H_n^\top z_n = \begin{bmatrix} 0{,}6 & -0{,}2 \\ -0{,}2 & 0{,}1 \end{bmatrix} \cdot \begin{bmatrix} 15{,}2 \\ 50{,}5 \end{bmatrix} = \begin{bmatrix} -0{,}98 \\ 2{,}01 \end{bmatrix}.$$

Согласно принципу инвариантности (см. теорему 3.1)

$$\widehat{\varphi}(t;\theta) = \varphi(t;\widehat{\theta}_n) = h^\top(t)\widehat{\theta}_n = -0{,}98 + 2{,}01t. \ \blacksquare$$

П р и м е р 10.5. В условиях примера 10.4 найдите закон распределения оценки $\widehat{\varphi}(t;\theta)$ полезного сигнала и постройте доверительный интервал надежности $q = 0{,}95$ для $\varphi(t;\theta)$ в момент $t = 5$.

Р е ш е н и е. Так как $\widehat{\varphi}(t;\theta) = h^\top(t)\widehat{\theta}_n$, то

$$\mathbf{M}\{\widehat{\varphi}(t;\theta)\} = h^\top(t)\mathbf{M}\left\{\widehat{\theta}_n\right\} = h^\top(t)\theta,$$

$$\mathbf{D}\{\widehat{\varphi}(t;\theta)\} = h^\top(t)\widehat{K}_n h(t) = \sigma_n^2 h^\top(t) W_n^{-1} h(t),$$

где $\widehat{K}_n = \sigma_n^2 W_n^{-1}$ — ковариационная матрица ошибки оценки $\widehat{\theta}_n$. Итак,

$$\widehat{\varphi}(t;\theta) \sim \mathcal{N}\left(h^\top(t)\theta;\ \sigma_n^2 h^\top(t) W_n^{-1} h(t)\right).$$

При этом ошибка оценки $\Delta\widehat{\varphi}(t;\theta) = \widehat{\varphi}(t;\theta) - \varphi(t;\theta)$ имеет распределение $\mathcal{N}\left(0;\ \sigma_n^2 h^\top(t) W_n^{-1} h(t)\right)$.

В условиях примера 10.4 для $t = 5$ находим

$$\sigma_n^2 h^\top(t) W_n^{-1} h(t) = 0{,}01 \begin{bmatrix} 1 \\ 5 \end{bmatrix}^\top \cdot \begin{bmatrix} 0{,}6 & -0{,}2 \\ -0{,}2 & 0{,}1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 5 \end{bmatrix} = 0{,}011.$$

Итак, $\widehat{\varphi}(5;\theta) - \varphi(5;\theta) \sim \mathcal{N}(0; 0{,}011)$. Отсюда

$$\mathbf{P}\left(|\widehat{\varphi}(5;\theta) - \varphi(5;\theta)| \leqslant u_\alpha \sqrt{0{,}011}\right) = 0{,}95,$$

где $u_\alpha = 1{,}96$ — квантиль уровня $\alpha = 0{,}975$ распределения $\mathcal{N}(0;1)$. Таким образом, искомый доверительный интервал имеет вид $\left[\widehat{\varphi}(5;\theta) - 1{,}96\sqrt{0{,}011};\ \widehat{\varphi}(5;\theta) + 1{,}96\sqrt{0{,}011}\right]$. Найдем реализацию этого интервала, используя результаты примера 10.4. Так как $\widehat{\varphi}(5;\theta) = -0{,}98 + 2{,}01 \cdot 5 = 9{,}07$, окончательно получаем интервал [8,86; 9,28], который с надежностью 0,95 накрывает точное значение $\varphi(5;\theta)$ полезного сигнала в точке $t = 5$. ■

Пример 10.6. Основываясь на данных из табл. 10.2, охватывающих период с 1954 по 1965 г., найдите реализацию оценки параметров потребительской функции США:

$$X_k = \theta_1 + h_k\theta_2 + \varepsilon_k, \quad k = 1, \ldots, n,$$

где X_k — индивидуальное потребление (млрд долл.) в k-м году, h_k — личные доходы (млрд долл.) в k-м году, ε_k, $k = 1, \ldots, n$ — независимые случайные ошибки (данные взяты из: U. S. Department of Commerce, Office of Business Economics, Survey of Current Business, July, 1966). Вычислите реализации оценки ковариационной матрицы ошибки вектора $\theta = (\theta_1, \theta_2)^\top$ и коэффициента детерминации.

Таблица 10.2

Год	X_k	h_k	Год	X_k	h_k
1954	236	257	1960	325	350
1955	254	275	1961	335	364
1956	267	293	1962	355	385
1957	281	309	1963	375	405
1958	290	319	1964	401	437
1959	311	337	1965	431	469

Решение. По формуле $\widehat{\theta}_n = \left(H_n^\top H_n\right)^{-1} H_n^\top z_n$ найдем реализацию оценки вектора неизвестных параметров $\theta = \{\theta_1, \theta_2\}^\top$. В нашем примере

$$H_n^\top = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 257 & 275 & 293 & 309 & 319 & 337 & 350 & 364 & 385 & 405 & 437 & 469 \end{bmatrix},$$

$$z_n = [236 \;\; 254 \;\; 267 \;\; 281 \;\; 290 \;\; 311 \;\; 325 \;\; 335 \;\; 355 \;\; 375 \;\; 401 \;\; 431]^\top .$$

Тогда

$$H_n^\top H_n = \begin{bmatrix} n & \sum\limits_{k=1}^n h_k \\ \sum\limits_{k=1}^n h_k & \sum\limits_{k=1}^n h_k^2 \end{bmatrix} = \begin{bmatrix} 12 & 4\,200 \\ 4\,200 & 1\,394\,495 \end{bmatrix};$$

$$H_n^\top z_n = \begin{bmatrix} \sum\limits_{k=1}^n X_k \\ \sum\limits_{k=1}^n X_k h_k \end{bmatrix} = \begin{bmatrix} 3\,861 \\ 1\,282\,345 \end{bmatrix}.$$

Далее можно найти или обратную матрицу $\left(H_n^\top H_n\right)^{-1}$ или решить систему уравнений:

$$\begin{cases} 12\,\theta_1 + 4\,200\,\theta_2 = 3\,861 \\ 4\,200\,\theta_1 + 1\,394\,495\,\theta_2 = 1\,282\,345. \end{cases} \tag{10.14}$$

Решая систему уравнений (10.14), находим реализацию оценки $\widehat{\theta} = \{-2{,}927 \;\; ; 0{,}928\}^\top$, т.е. искомая зависимость имеет вид $X = -2{,}927 + 0{,}928h$.

Теперь найдем реализацию оценки ковариационной матрицы ошибки

$$\widehat{K}_n = \widehat{\sigma}_n^2 (H_n^\top H_n)^{-1},$$

где

$$\widehat{\sigma}_n^2 = \frac{1}{n-2} \sum_{k=1}^n \left(X_k - \widehat{X}_k\right)^2 = \frac{1}{n-1} \sum_{k=1}^n \left(X_k - (\widehat{\theta}_1 + h_k \widehat{\theta}_2)\right)^2 =$$
$$= \frac{1}{n-1} \sum_{k=1}^n \widehat{\varepsilon}_k^2 = \frac{1}{10} \Big[(236 - (-2{,}297 + 0{,}928 \cdot 257))^2 + \ldots +$$
$$+ (431 - (-2{,}927 + 0{,}928 \cdot 469))^2\Big] = 4{,}48.$$

Величина $\widehat{\sigma}_n^2$ является несмещенной оценкой неизвестной дисперсии σ_n^2 (см. теорему 1.3 из [13]).

Тогда

$$\widehat{K}_n = 4{,}48 \begin{bmatrix} 2{,}72 & -7 \cdot 10^{-3} \\ -7 \cdot 10^{-3} & 2{,}1 \cdot 10^{-5} \end{bmatrix} = \begin{bmatrix} 12{,}17 & -0{,}03 \\ -0{,}03 & 9{,}6 \cdot 10^{-5} \end{bmatrix}.$$

Вычислим реализацию коэффициента детерминации:

$$R^2 = 1 - \frac{\sum\limits_{k=1}^{n}(X_k - \widehat{X}_k)^2}{\sum\limits_{k=1}^{n}(X_k - \overline{X}_n)^2},$$

где

$$\overline{X}_n = \frac{1}{12}\sum_{k=1}^{12} X_k = 321{,}75; \quad \sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = 40\,070;$$

$$\sum_{k=1}^{n}(X_k - \widehat{X}_k)^2 = 44{,}79.$$

Тогда $R^2 = 1 - \dfrac{44{,}79}{40\,070} = 0{,}999$. Столь близкое к единице значение коэффициента детерминации позволяет сделать вывод о наличии связи между X и h, отличной от тревиальной. ■

Пример 10.7. По данным из примера 10.6, считая, что ошибки в наблюдениях имеют гауссовское распределение, проверьте с помощью критерия Фишера гипотезы: H_1: $3\theta_2 - \theta_1 = 0$, H_2: $\theta_1 = 0$, $\theta_2 = 0$, H_3: $\theta_2 = 1$ на уровне значимости $\alpha = 0{,}05$.

Решение. Чтобы свести гипотезы H_1, H_2 и H_3 к виду (10.10), положим:

1) $q_1 = 1$, $c_1 = 0$, $A_1 = \begin{bmatrix} -1 & 3 \end{bmatrix}$;

2) $q_2 = 2$, $c_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$;

3) $q_3 = 1$, $c_3 = 1$, $A_3 = \begin{bmatrix} 0 & 1 \end{bmatrix}$.

Используя результаты примера 10.6, вычислим реализации статистики критерия Фишера (10.11) для гипотез H_1, H_2, H_3 соответственно:

$$F_1 = 2{,}64, \quad F_2 = 1{,}431 \cdot 10^5, \quad F_3 = 54{,}35.$$

Для уровня значимости 0,05 критическая область критерия Фишера (10.12) имеет вид

1) $I_1 = (f_{0,95}(1;10);\infty) = (4{,}96;\infty)$;

2) $I_2 = (f_{0,95}(2;10);\infty) = (4{,}10;\infty)$;

3) $I_3 = (f_{0,95}(1;10);\infty) = (4{,}96;\infty)$.

Таким образом, в первом случае реализация статистики не попадает в критическую область и гипотеза H_1 принимается на уровне значимости 0,05, а во втором и третьем случаях реализация статистики попадает в критическую область и гипотезы H_2 и H_3 отвергаются на уровне значимости 0,05. ■

Пример 10.8. В табл. 10.3 приведены данные о производительности труда и уровне занятости для 11 стран за некоторый год (данные взяты из: *Kaldor N.* Causes of the Slow Rate of Economic

Growth of the United Kingdom. Cambridge Univercity Press, 1966). Производительность и уровень занятости измеряются как приросты в процентах за год. Постройте зависимость $X = \theta_1 + \theta_2 h + \theta_3 h^2$, где X — производительность труда, h — уровень занятости. Вычислите коэффициент детерминации R^2 и постройте прогноз производительности для Японии, если уровень занятости в этом году равен 5,8.

Таблица 10.3

Страна	Занятость	Производительность
Австрия	2,0	4,2
Бельгия	1,5	3,9
Канада	2,3	1,3
Дания	2,5	3,2
Франция	1,9	3,8
Италия	4,4	4,2
Нидерланды	1,9	4,1
Норвегия	0,5	4,4
ФРГ	2,7	4,5
Великобритания	0,6	2,8
США	0,8	2,6

Решение. Вычислим реализацию МНК-оценки вектора параметров $\theta = \{\theta_1, \theta_2, \theta_3\}^\top$ по формуле

$$\widehat{\theta}_n = \left(H_n^\top H_n\right)^{-1} H_n^\top Z_n, \tag{10.15}$$

где реализация выборки Z_n находится в третьем столбце таблицы, а матрица $H_n^\top$ имеет вид

$$H_n^\top = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2{,}0 & 1{,}5 & 2{,}3 & 2{,}5 & 1{,}9 & 4{,}4 & 1{,}9 & 0{,}5 & 2{,}7 & 0{,}6 & 0{,}8 \\ 4{,}0 & 2{,}25 & 5{,}29 & 6{,}25 & 3{,}61 & 19{,}36 & 3{,}61 & 0{,}25 & 7{,}29 & 0{,}36 & 0{,}64 \end{bmatrix}.$$

Вычислим обратную матрицу $\left(H_n^\top H_n\right)^{-1}$ и вектор $H_n^\top z_n$

$$\left(H_n^\top H_n\right)^{-1} = \begin{bmatrix} 1{,}003 & -0{,}862 & 0{,}154 \\ -0{,}862 & 0{,}894 & -0{,}177 \\ 0{,}154 & -0{,}177 & 0{,}039 \end{bmatrix}, \quad H_n^\top z_n = \begin{bmatrix} 39 \\ 76{,}84 \\ 198{,}86 \end{bmatrix}.$$

Тогда по формуле (10.15) получим $\widehat{\theta}_n = \begin{bmatrix} 3{,}53 \\ -0{,}18 \\ 0{,}07 \end{bmatrix}$.

Теперь вычислим реализацию коэффициента детерминации (см. пример 10.6) $R^2 = 1 - \dfrac{\sum\limits_{k=1}^{n} (X_k - \widehat{X}_k)^2}{\sum\limits_{k=1}^{n} (X_k - \overline{X}_n)^2} = 0{,}04967$. Поскольку реализация коэффициента очень близка к нулю, то это означает, что построенная нами зависимость ненамного лучше тривиальной модели $X(h) = \overline{X}_n = 3{,}545$. Это возможно в том случае, когда связь между уровнем занятости и производительностью практически отсутствует.

Для того чтобы подтвердить предположение об отсутствии связи (квадратичного вида), между уровнем занятости и производительностью проверим гипотезу H_0: $\theta_2 = 0$, $\theta_3 = 0$ с помощью критерия Фишера, считая, что ошибки в наблюдениях являются гауссовскими.

Чтобы свести гипотезу H_0 к виду (10.10), положим $q = 1$, $c = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ и $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Вычислим реализацию статистики критерия Фишера (10.11):

$$(A\widehat{\theta}_n - c) = \begin{bmatrix} -0{,}181 \\ 0{,}075 \end{bmatrix}, \quad A(H_n^\top H_n)^{-1}A^\top = \begin{bmatrix} 0{,}894 & -0{,}177 \\ -0{,}177 & 0{,}039 \end{bmatrix},$$

$$(Z_n - H_n\widehat{\theta}_n)^\top (Z_n - H_n\widehat{\theta}_n) = 9{,}13,$$

тогда

$$F = \frac{0{,}477}{\frac{2}{11-3}9{,}13} = 0{,}209.$$

Выберем уровень значимости равным 0,05 и построим критическую область критерия Фишера (10.12) $(f_{0,95}(2;8); \infty) = (4{,}46; \infty)$. Поскольку $0{,}209 \notin (4{,}46; \infty)$, т.е. реализация статистики критерия не попала в критическую область, то гипотеза H_0 принимается на уровне значимости 0,05 и можно говорить о том, что рассматриваемой зависимости между производительностью и занятостью нет.

Прогноз для Японии можно построить, подставив уровень занятости для этой страны в найденное уравнение регрессии: $X_{\text{яп}} = 3{,}53 - 0{,}18 \cdot 5{,}8 + 0{,}07 \cdot (5{,}8)^2 = 5{,}004$. Истинное значение производительности для Японии за этот год равно 7,8 и существенно отличается от сделанного нами прогноза, что еще раз говорит о низком качестве рассматриваемой зависимости $X(h)$. ■

10.6. Задачи для самостоятельного решения

1. Модель линейной регрессии имеет следующий частный вид:

$$X_k = \theta_1 + h_k\theta_2 + \varepsilon_k, \quad k = 1, \ldots, n,$$

где $\{\varepsilon_k\}$ — независимые центрированные СВ. Найдите явное аналитическое выражение для МНК-оценки $\widehat{\theta}_n$ вектора $\theta = \{\theta_1, \theta_2\}^\top$.

2.* Докажите соотношение (10.7):

$$\sum_{k=1}^{n}\left(X_k-\overline{X}_n\right)^2=\sum_{k=1}^{n}\left(\widehat{X}_n-\overline{X}_n\right)^2+\sum_{k=1}^{n}\widehat{\varepsilon}_k^2,$$

где $\overline{X}_n=\sum\limits_{k=1}^{n}X_k$, $\widehat{X}_k=h_k^\top\widehat{\theta}_n$, $\widehat{\varepsilon}_k=X_k-\widehat{X}_k$.

3. Пусть $X=H\theta+E$, где $\mathbf{M}\{E\}=0$, $\mathbf{cov}(E,E)=\sigma^2 I$, $H\in\mathbb{R}^{n\times p}$. Докажите, что если H и θ разбиты на блоки в форме

$$X\theta=[\,H_1\ H_2\,]\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix},$$

то МНК-оценка для θ_2 имеет вид:

$$\widehat{\theta}_2=\left[H_2^\top H_2-H_2^\top H_1(H_1^\top H_1)^{-1}H_1^\top H_2\right]^{-1}\left[H_2^\top X-H_2^\top H_1(H_1^\top H_1)^{-1}H_1^\top X\right].$$

4. По двум наблюдениям $X_1=2$, $X_2=4$ при $h_1=1$, $h_2=2$ постройте реализацию МНК-оценки неизвестных параметров в модели простой линейной регрессии. Вычислите коэффициент детерминации.

5. В условиях примера 10.4 проверьте с помощью критерия Фишера гипотезу H_0: $\theta_1=-1$, $\theta_2=2$ на уровне значимости $\alpha=0{,}01$, считая, что ошибки в наблюдениях имеют распределение $\mathcal{N}(0;\sigma^2)$.

6. Модель наблюдений имеет вид

$$X_k=k\theta+\varepsilon_k,\quad k=1,\ldots,n,$$

где $\{\varepsilon_k\}$ — независимые гауссовские СВ, $\varepsilon_k\sim\mathcal{N}(0;\sigma^2)$, $\sigma>0$. Используя МНК-оценку $\widehat{\theta}_n$ параметра θ, постройте для него доверительный интервал надежности $q=0{,}95$.

Ответ: $\left[\widehat{\theta}_n-1{,}96\,\sigma\sqrt{\psi(n)};\widehat{\theta}_n+1{,}96\,\sigma\sqrt{\psi(n)}\right]$, где $\widehat{\theta}_n=\psi(n)\sum\limits_{k=1}^{n}kX_k$, $\psi(n)=6(2n^2+3n+1)^{-1}$.

7. В условиях примера 10.4 проверить на уровне значимости $p=0{,}05$ параметрическую гипотезу H_0: $\theta_2=0$, считая, что ошибки в наблюдениях независимы и имеют распределение $\mathcal{N}(0;\sigma^2)$.

Указание. Постройте доверительный интервал надежности $q=0{,}95$ для параметра θ_2.

Ответ: H_0 отвергается.

8. По данным примера 10.6 постройте зависимость $X=\theta_1+\theta_2 h$, считая, что:

1) $h=[1,\ldots,12]^\top$;

2) $h=[1954,\ldots,1967]^\top$.

В обоих случаях оцените погрешность найденной реализации оценки и вычислите коэффициент детерминации.

9. Оцените неизвестные параметры α, β в модели потребления человеком некоторого продукта $v=\alpha+\beta\ln w$, где v — расходы на продукт потребления, w — недельный доход (табл. 10.4).

Таблица 10.4

k	1	2	3	4	5	6	7
w_k	0,8	1,2	1,5	2,2	2,3	3,6	3,1
v_k	1,7	2,7	3,6	5,7	6,7	8,1	12,0

Оцените коэффициент эластичности $\sigma = \frac{dv}{dw} \cdot \frac{w}{v}$. Считая, что ошибки в наблюдениях имеют распределение $\mathcal{N}(0; \sigma^2)$, с помощью критерия Фишера проверьте гипотезу H_0: $\beta = 1$ на уровне значимости 0,05.

11. Обобщенная линейная модель регрессии

11.1. Обобщенный метод наименьших квадратов

Естественно ожидать, что при моделировании многих реальных процессов мы можем столкнуться с ситуациями, в которых свойства линейной регрессионной модели, изложенные в предыдущем параграфе, оказываются нарушенными. Так, если в качестве исходных статистических данных мы используем временные или пространственно-временные выборки, то чрезмерно ограничительными становятся, как правило, условия *взаимной некоррелированности и гомоскедастичности ошибок*, которые в терминах ковариационной матрицы ошибок выражаются соотношением $\mathbf{cov}\,(E_n, E_n) = K_{E_n} = \sigma^2 I_n$.

Определение 11.1. Пусть вектор ошибок наблюдений E_n в модели (10.2) имеет нулевое среднее $\mathbf{M}\{E_n\} = 0$ и произвольную невырожденную ковариационную матрицу $K_{E_n} = V_n > 0$. Соответствующая модель наблюдения называется *обобщенной линейной регрессионной моделью.*

В модели обобщенной линейной регрессии можно выделить два частных случая.

1. Линейная модель регрессии с *автокоррелированными* ошибками. В данной модели будем считать, что компоненты вектора E_n являются коррелированными, а значит, матрица V_n не может быть диагональной. Относительно природы зависимости предположим, что она ослабевает по мере взаимного удаления моментов наблюдения друг от друга. Одной из удобных форм реализации этого допущения (при сохранении свойства гомоскедастичности ошибок) является следующая:

$$r(\varepsilon_i, \varepsilon_j) = \rho^{|i-j|}, \tag{11.1}$$

где $r(\varepsilon_i, \varepsilon_j)$ — коэффициент корреляции между ε_i и ε_j, а ρ — некоторое число, по модулю меньшее единицы. Из уравнения (11.1) следует, что ρ — коэффициент корреляции между соседними ошибками.

Уравнение (11.1), в частности, означает:

а) корреляционная связь между ошибками зависит только от меры их «разнесенности», но не зависит от того, к каким именно моментам наблюдения i и j они «привязаны», т.е. $r(\varepsilon_i, \varepsilon_j) = r(\varepsilon_{i+k}, \varepsilon_{j+k})$;

б) корреляционная связь между ε_i и ε_j исчезает при $|i-j| \to \infty$, т.е. при неограниченном удалении ошибок друг от друга;

в) ковариационная матрица вектора E_n имеет вид

$$V_n = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}. \tag{11.2}$$

2. Линейная модель регрессии с *гетероскедастичными* некоррелированными ошибками. В этом случае ковариационная матрица вектора E_n имеет вид

$$V_n = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}, \tag{11.3}$$

где величины $\{\sigma_1^2, \ldots \sigma_n^2\}$ неизвестны.

После рассмотрения частных случаев обобщенной линейной модели регрессии приведем общий вид оценки для произвольной матрицы ошибок.

Теорема 11.1. *Если матрица ковариаций V_n ошибок наблюдений невырождена, тогда справедливо*

1) *НЛН-оценка $\widehat{Y}_n$ вектора Y имеет вид* (10.4), *где*

$$\widehat{\theta}_n = \left(H_n^\top V_n^{-1} H_n\right)^{-1} H_n^\top V_n^{-1} Z_n; \tag{11.4}$$

2) $\widehat{\theta}_n$ *является НЛН-оценкой для* θ;

3) *ковариационная матрица ошибки* $\Delta\widehat{\theta}_n = \widehat{\theta}_n - \theta_n$ *имеет вид*

$$\widehat{K}_n = \left(H_n^\top V_n^{-1} H_n\right)^{-1}. \tag{11.5}$$

Определение 11.2. Оценка (11.4) называется *оценкой обобщенного метода наименьших квадратов* (ОМНК-оценкой).

Как известно, МНК-оценка является результатом минимизации по θ функции потерь $J(\theta) = |Z_n - H_n\theta|^2$.

Приведем вид этой функции для обобщенной линейной модели:

$$\begin{aligned}J(\theta) &= \left[C^{-1}((Z_n - H_n\theta)\right]^{\top}\left[C^{-1}(Z_n - H_n\theta))\right] = \\ &= (Z_n - H_n\theta)^{\top}\left(C^{-1}\right)^{\top}C^{-1}(Z_n - H_n\theta) = \\ &= (Z_n - H_n\theta)^{\top}V_n^{-1}(Z_n - H_n\theta).\end{aligned} \tag{11.6}$$

Поэтому ОМНК-оценка $\widehat{\theta}_n$ может быть также определена, как точка минимума обобщенной функции потерь (11.6).

Если вектор ошибок E_n имеет многомерное нормальное распределение, то можно показать, что оценка вектора θ, получаемая с помощью ОМНК, совпадает с МП-оценкой (естественно, при известной матрице V_n).

Для обобщенной регрессионной модели, в отличие от обычной, коэффициент детерминации

$$R^2 = 1 - \frac{(Z_n - H_n\widehat{\theta}_n)^{\top}(Z_n - H_n\widehat{\theta}_n)}{(Z_n - \overline{Z}_n)^{\top}(Z_n - \overline{Z}_n)}$$

не может служить удовлетворительной мерой качества выбранной модели. В общем случае он даже не принадлежит интервалу $[0;1]$, а добавление или удаление объясняющей переменной не обязательно приводит к его уменьшению или увеличению.

Подчеркнем, что для применения ОМНК необходимо знать матрицу V_n. Если матрица V_n неизвестна, то в общем случае найти «хорошую» оценку ее $\dfrac{n(n+1)}{2}$ элементов не представляется возможным.

11.2. Автокорреляция

Один из наиболее простых способов учета коррелированности ошибок состоит в предположении, что их последовательность $\{\varepsilon_k, k = 1, 2, \ldots, \infty\}$ удовлетворяют рекуррентному соотношению

$$\varepsilon_k = \rho\varepsilon_{k-1} + \nu_k, \tag{11.7}$$

где ν_k — последовательность независимых нормально распределенных СВ с нулевым средним $\mathbf{M}\{\nu_k\} = 0$ и постоянной дисперсией σ_0^2, а ρ — *коэффициент автокорреляции* ($|\rho| < 1$).

Определим основные числовые характеристики вектора ошибок E_n (среднее значение $\mathbf{M}\{E_n\}$ и ковариационную матрицу V_n) для модели (11.7).

Из (11.7) следует:

$$\begin{aligned}\varepsilon_k &= \rho\varepsilon_{k-1} + \nu_k = \rho(\rho\varepsilon_{k-2} + \nu_{k-1}) + \nu_k = \\ &= \rho^2\varepsilon_{k-2} + \rho\nu_{k-1} + \nu_k = \ldots = \\ &= \nu_k + \rho\nu_{k-1} + \rho\nu_{k-2} + \ldots = \sum_{j=0}^{\infty}\rho^j\nu_{k-j}.\end{aligned} \tag{11.8}$$

Из (11.8) с учетом $\mathbf{M}\{\nu_k\} = 0$, $\mathbf{D}\{\nu_k\} = \sigma_0^2$ получаем, что

$$\mathbf{M}\{\varepsilon_k\} = 0,$$

$$\begin{aligned} \sigma_\varepsilon^2 = \mathbf{M}\{\varepsilon_k^2\} = \mathbf{M}\{\nu_k^2\} + \rho^2\mathbf{M}\{\nu_{k-1}^2\} + \\ + \rho^4\mathbf{M}\{\nu_{k-2}^2\} + \ldots = \frac{\sigma_0^2}{1-\rho^2}. \end{aligned} \tag{11.9}$$

Для вычисления ковариаций $\mathbf{cov}(\varepsilon_k, \varepsilon_{k-i}) = \mathbf{M}\{\varepsilon_k\varepsilon_{k-i}\}$, $k = 1, 2, \ldots, n$, $i = 1, 2, \ldots, k-1$ представим произведение $\varepsilon_k\varepsilon_{k-i}$, используя соотношение (11.8), в виде

$$\begin{aligned} \varepsilon_k\varepsilon_{k-i} = \left[\nu_k + \rho\nu_{k-1} + \ldots + \rho^{i-1}\nu_{k-(i-1)} + \rho^i(\nu_{k-i} + \rho\nu_{k-(i-1)} + \ldots)\right]\cdot \\ \cdot[\nu_{k-i} + \rho\nu_{k-(i-1)} + \ldots]. \end{aligned}$$

Тогда, поскольку из взаимной некоррелированности ν_k следует некоррелированность случайных величин $\left(\nu_k + \rho\nu_{k-1} + \ldots + \rho^{i-1}\nu_{k-(i-1)}\right)$ и $\left(\nu_{k-i} + \rho\nu_{k-(i-1)} + \ldots\right)$, получаем:

$$\begin{aligned} \mathbf{cov}(\varepsilon_k, \varepsilon_{k-i}) = \mathbf{M}\left\{\rho^i(\nu_{k-i} + \rho\nu_{k-(i-1)} + \ldots)^2\right\} = \\ = \rho^i\mathbf{M}\{\varepsilon_{k-i}^2\} = \sigma_\varepsilon^2\rho^i, \end{aligned} \tag{11.10}$$

где σ_ε^2 определена выражением (11.9).

Таким образом, ковариационная матрица вектора ошибок размера n имеет вид

$$V_n = \frac{\sigma_0^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}. \tag{11.11}$$

Матрица V_n^{-1} для V_n вида (11.11) вычисляется аналитически:

$$V_n^{-1} = \frac{(1-\rho^2)^2}{\sigma_0^2}\begin{bmatrix} 1 & -\rho & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}. \tag{11.12}$$

11.3. Оценивание в модели с коррелированными ошибками

Если значение ρ в модели (11.7) неизвестно, что встречается очень часто, то его требуется оценить. Для этого существует несколько процедур.

Процедура А. Начальным шагом этой процедуры является вычисление МНК-оценки в исходной линейной регрессии и получение вектора оценок остатков $\widehat{E}_n = \{\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n\}^\top = Z_n - H_n\widehat{\theta}_n$. После этого используется итерационный алгоритм:

1) в качестве приближенного значения $\widehat{\rho}$ берется его МНК-оценка в регрессии (11.7), которая строится по значениям $\widehat{E}_n$:

$$\widehat{\rho} = \frac{\sum\limits_{k=2}^{n} \widehat{\varepsilon}_{k-1}\widehat{\varepsilon}_k}{\sum\limits_{k=2}^{n} \left(\widehat{\varepsilon}_{k-1}\right)^2};$$

2) вычисляется оценка $\widehat{\sigma}_0^2 = \dfrac{1}{n}\sum\limits_{k=1}^{n} \widehat{\varepsilon}_k^2$;

3) находится ОМНК-оценка (11.4) с заменой ρ на $\widehat{\rho}$ и σ_0^2 на $\widehat{\sigma}_0^2$ в выражении (11.12);

4) строится новый вектор оценок остатков $\widehat{E}_n$;

5) процедура повторяется, начиная с шага 1).

Процесс обычно заканчивается, когда очередное приближение $\widehat{\rho}$ мало отличается от предыдущего, хотя возможны и другие критерии остановки процедуры A.

Процедура Б (*Дарбина*) [44]. Строим МНК-оценки параметров следующей регрессии:

$$X_k = h_k^\top \beta^{(1)} + h_{k-1}^\top \beta^{(2)} + \rho X_{k-1} + \xi_k, \quad k = 2, \ldots, n,$$

т.е. X_{k-1} включается в число регрессионных переменных, а ρ — в число оцениваемых параметров. Здесь $\beta^{(1)}$ и $\beta^{(2)}$ — два вектора неизвестных параметров модели размерности p, а ξ_k — случайные ошибки, $k = 2, \ldots, n$.

Полученную оценку $\widehat{\rho}$ используем для получения скорректированных значений

$$X_k' = X_k - \widehat{\rho} X_{k-1}, \quad k = 2, \ldots, n,$$

$$h_{i,k}' = h_{i,k} - \widehat{\rho} h_{i,k-1}, \quad i = 1, \ldots, p, \ \ k = 2, \ldots, n.$$

Далее для скорректированных значений регрессионных переменных и наблюдений строится МНК-оценка неизвестных параметров.

11.4. Критерий Дарбина—Уотсона

Когда проверяют гипотезу о наличии корреляции в ошибках наблюдений для модели линейной регрессии, обычно используют критерий, разработанный Дарбином и Уотсоном [45].

Проверяется гипотеза H_0: $\rho = 0$ против альтернативы H_1: $\rho \neq 0$. При проверке используется статистика D, которая является взвешенной суммой квадратов разностей последовательных остатков:

$$D = \frac{\sum\limits_{k=2}^{n} (\widehat{\varepsilon}_k - \widehat{\varepsilon}_{k-1})^2}{\sum\limits_{k=1}^{n} \widehat{\varepsilon}_k^2}. \tag{11.13}$$

Если реализация статистики D мала, то это означает, что последовательные величины $\widehat{\varepsilon}_k$ близки друг к другу, и может свидетельствовать о существовании положительной корреляции. А когда реализация статистики очень велика, то эти величины сильно отличаются друг от друга, что является признаком отрицательной корреляции.

Раскроем скобки в выражении (11.13) и будем считать (для достаточно больших n), что $\sum\limits_{k=2}^{n} \widehat{\varepsilon}_k^2 = \sum\limits_{k=2}^{n} \widehat{\varepsilon}_{k-1}^2$, тогда

$$D \approx \frac{2\sum\limits_{k=2}^{n} \widehat{\varepsilon}_k^2 - 2\sum\limits_{k=2}^{n} \widehat{\varepsilon}_k \widehat{\varepsilon}_{k-1}}{\sum\limits_{k=1}^{n} \widehat{\varepsilon}_k{}^2} \approx 2 - \frac{2\sum\limits_{k=2}^{n} \widehat{\varepsilon}_k \widehat{\varepsilon}_{k-1}}{\sum\limits_{k=1}^{n} \widehat{\varepsilon}_k{}^2}.$$

Наконец, считая, что слагаемые $\widehat{\varepsilon}_1^2$ и $\widehat{\varepsilon}_n^2$ пренебрежимо малы по сравнению с общей суммой $\sum\limits_{k=1}^{n} \widehat{\varepsilon}_k^2$, окончательно получаем

$$D \approx 2(1 - \widehat{r}), \tag{11.14}$$

где $\widehat{r}$ — выборочный коэффициент корреляции между $\widehat{\varepsilon}_k$ и $\widehat{\varepsilon}_{k-1}$.

Достоинства приближенной формулы (11.14) зависят от относительной величины остатков $\widehat{\varepsilon}_1$ и $\widehat{\varepsilon}_n$, поэтому следует проявлять осмотрительность при использовании приближенной формулы.

Из уравнения (11.14) следует, что реализация D может изменяться от нуля (когда $\widehat{r} = 1$) до 4 (когда $\widehat{r} = -1$). Среднее значение, равное двум, соответствует нулевой корреляции.

При использовании данного критерия границы принятия гипотезы H_0 и отклонения альтернативной гипотезы H_1 не совпадают между собой. Как показано в табл. 11.1, критические значения D позволяют выделить пять областей различных статистических решений. При этом появляются области неопределенности, где невозможно ни принять, ни отвергнуть гипотезу.

Таблица 11.1

Реализация статистики D	Вывод
$0 < d < d_l$	Гипотеза H_0 отвергается, принимается H_1 — есть положительная корреляция
$d_l < d < d_u$	Гипотеза H_0 не принимается и не отвергается
$d_u < d < 4 - d_u$	Гипотеза H_0 принимается
$4 - d_u < d < 4 - d_l$	Гипотеза H_0 не принимается и не отвергается
$4 - d_l < d < 4$	Гипотеза H_0 отвергается, принимается H_1 — есть отрицательная корреляция

В табл. 22.5 (считая, что СВ ε_k, $k = 1, \ldots, n$ имеют гауссовское распределение $\mathcal{N}(0; \sigma^2)$) приведены верхнее d_u и нижнее d_l критические значения статистики D при альтернативе H_1: $\rho \neq 0$ на уровне значимости 0,05, зависящие от числа параметров модели p.

В экономических исследованиях часто в качестве альтернативной гипотезы рассматривают гипотезу о существовании положительной или отрицательной корреляции. Тогда данные в табл. 22.5 соответствуют уровню значимости 0,025.

11.5. Примеры

Пример 11.1. Докажите теорему 11.1.

Решение. Покажем сначала несмещенность оценки (11.4).

$$
\begin{aligned}
\mathbf{M}\Big\{\widehat{\theta}_n\Big\} &= \mathbf{M}\Big\{\big(H_n^\top V_n^{-1} H_n\big)^{-1} H_n^\top V_n^{-1} Z_n\Big\} = \\
&= \mathbf{M}\Big\{\big(H_n^\top V_n^{-1} H_n\big)^{-1} H_n^\top V_n^{-1} (H_n\theta + E_n)\Big\} = \\
&= \mathbf{M}\Big\{\big(H_n^\top V_n^{-1} H_n\big)^{-1} H_n^\top V_n^{-1} H_n\theta\Big\} + \\
&+ \mathbf{M}\Big\{\big(H_n^\top V_n^{-1} H_n\big)^{-1} H_n^\top V_n^{-1} E_n\Big\} = \\
&= \theta + \big(H_n^\top V_n^{-1} H_n\big)^{-1} H_n^\top V_n^{-1} \mathbf{M}\{E_n\} = \theta.
\end{aligned}
$$

Теперь для доказательства утверждения 2) теоремы об оптимальности оценки (11.4) преобразуем ковариационную матрицу V_n к виду $\sigma^2 I_n$, а затем воспользуемся теоремой 10.3.

Выполнить необходимое преобразование позволяет результат из матричной алгебры, в соответствии с которым всякая положительно определенная симметричная $(n \times n)$-матрица A допускает представление в виде

$$A = CC^\top, \tag{11.15}$$

где C — некоторая невырожденная $(n\times n)$-матрица, причем $\left(C^\top\right)^{-1} = \left(C^{-1}\right)^\top$. Воспользуемся этим, чтобы представить матрицу V_n в виде (11.15). Итак, существует матрица C такая, что $V_n = CC^\top$. Тогда

$$C^{-1}V_n\left(C^{-1}\right)^\top = I_n, \tag{11.16}$$

$$\left(C^{-1}\right)^\top C^{-1} = V_n^{-1}, \tag{11.17}$$

где соотношение (11.17) получено с учетом правила обращения произведения квадратных невырожденных матриц $(AB)^{-1} = B^{-1}A^{-1}$.

Теперь, умножив уравнение (10.2) слева на матрицу C^{-1}, получим:

$$Z_n^c = H_n^c\theta + E_n^c, \tag{11.18}$$

где $Z_n^c = C^{-1}Z_n$, $H_n^c = C^{-1}H_n$, $E_n^c = C^{-1}E_n$.

Найдем ковариационную матрицу вектора E_n^c, используя соотношение (11.16).

$$\begin{aligned}&\mathbf{cov}\,(E_n^c, E_n^c) = \mathbf{M}\Big\{C^{-1}E_n\left(C^{-1}E_n\right)^\top\Big\} = \\ &= C^{-1}\mathbf{M}\big\{E_nE_n^\top\big\}\left(C^{-1}\right)^\top = C^{-1}V_n\left(C^{-1}\right)^\top = I_n.\end{aligned}$$

Следовательно, в соответствии с (10.5) НЛН-оценка в модели (11.18) имеет вид

$$\widehat{\theta}_n = \left[(H_n^c)^\top H_n^c\right]^{-1}(H_n^c)^\top Z_n,$$

ковариационная матрица которой равна

$$\widehat{K}_{Z_n^c} = \left[(H_n^c)^\top H_n^c\right]^{-1}.$$

Возвращаясь к исходным обозначениям, получим

$$\begin{aligned}&\widehat{\theta}_n = \left[H_n^\top(C^{-1})^\top(C^{-1}H_n)\right]^{-1}H_n^\top(C^{-1})^\top C^{-1}Z_n = \\ &= \left(H_n^\top V_n^{-1}H_n\right)^{-1}H_n^\top V_n^{-1}Z_n, \\ &\widehat{K}_n = \left(H_n^\top V_n^{-1}H_n\right)^{-1}. \ \blacksquare\end{aligned}$$

Пример 11.2. Рассматривается обобщенная линейная регрессионная модель

$$X_k = \theta_1 + \theta_2 h_k + \varepsilon_k, \quad k = 1, \ldots, 10,$$

где ε_k — центрированные СВ. Наблюдения $\{h_k, X_k\}$, $k = 1, \ldots, 10$, представлены в табл. 11.2.

Таблица 11.2

h_k	8	10	12	16	20	20	24	28	30	36
X_k	6,8	6,9	7,3	7,4	8,6	8,0	8,8	8,0	9,9	10,3

Найдите реализацию ОМНК-оценки вектора $\theta = [\theta_1, \theta_2]^\top$, если дисперсии ошибок ε_k известны и равны 0,04, если $5 \leqslant h_k < 15$, равны 0,16, если $15 \leqslant h_k < 25$ и равны 1,0, если $25 \leqslant h_k \leqslant 40$. Сравните найденную реализацию ОМНК-оценки с реализацией МНК-оценки для той же модели наблюдений, построенной в предположении, что дисперсии ошибок равны, а также реализации ковариационных матриц ошибок найденных оценок.

Решение. Вычислим реализацию ОМНК-оценки по формуле (11.4)

$$\widehat{\theta}^o_n = (H^\top V_n^{-1} H)^{-1} H^\top V_n^{-1} z_n,$$

где $V_n = \mathbf{cov}(E, E)$ — ковариационная матрица вектора ошибок $E = \{\varepsilon_1, \ldots, \varepsilon_{10}\}^\top$. В нашем случае ковариационная матрица имеет вид

$$V_n = \begin{bmatrix} 0{,}04 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}04 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}04 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}16 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}16 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}16 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0{,}16 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Рассматриваемая модель является моделью линейной регрессии с гетероскедастичными некоррелированными ошибками.

Вычисляя

$$\widehat{K}_n = \left(H_n^\top V_n^{-1} H_n\right)^{-1} = \begin{bmatrix} 103 & 1344 \\ 1344 & 20880 \end{bmatrix}^{-1} = \begin{bmatrix} 0{,}061 & -0{,}0039 \\ -0{,}0039 & 0{,}0003 \end{bmatrix};$$

$$H^\top V_n^{-1} Z_n = \begin{bmatrix} 758{,}2 \\ 10302 \end{bmatrix},$$

по формуле (11.4) окончательно получаем

$$\widehat{\theta}^o_n = \begin{bmatrix} 0{,}061 & -0{,}0039 \\ -0{,}0039 & 0{,}0003 \end{bmatrix} \cdot \begin{bmatrix} 758{,}2 \\ 10302 \end{bmatrix} = \begin{bmatrix} 5{,}767 \\ 0{,}122 \end{bmatrix}.$$

Обычная МНК-оценка не зависит от ковариационной матрицы вектора E и ее реализация равна $\widehat{\theta}_n = [5{,}736;\ 0{,}121]^\top$ (процедура ее вычисления подробно описана в примерах 10.4 и 10.6). Вычислим реализацию ковариационной матрицы ошибки МНК-оценки $\Delta\widehat{\theta}_n = \left(H_n^\top H_n\right)^{-1} H_n E$:

$$\widetilde{K}_n = \begin{bmatrix} 0{,}1672 & -0{,}0105 \\ -0{,}0105 & 0{,}0007 \end{bmatrix}.$$

Сравнивая эту матрицу с ковариационной матрицей ошибки ОМНК-оценки $\widehat{K}_n$, видим, что каждая компонента МНК-оценки обладает большей дисперсией (главная диагональ матрицы $\widetilde{K}_n$), чем соответствующая компонента ОМНК-оценки. ■

Пример 11.3. Пусть в модели из примера 11.2 ошибки наблюдений удовлетворяют уравнению авторегрессии первого порядка:

$$\varepsilon_k = \rho\varepsilon_{k-1} + v_k, \quad k = 2, \ldots, n,$$

где v_k — последовательность независимых нормально распределенных СВ с нулевым средним и дисперсией $\sigma_0^2 = 1$. Найдите ОМНК-оценку вектора неизвестных параметров, считая, что коэффициент авторегрессии ρ равен а) $\rho_1 = 0{,}5$; б) $\rho_2 = -0{,}5$; в) $\rho_3 = 0{,}9$; г) $\rho_4 = -0{,}9$.

Решение. Для такой модели ковариационная матрица вектора ошибок E_n вычисляется аналитически по формуле (11.2).

В силу большой размерности матрицы V_n ее выражение не приводится, а ОМНК-оценка вычисляется по формуле (11.4). Тогда реализации оценки и ковариационной матрицы ошибки оценки для $\rho_1 = 0{,}5$ имеют вид:

$$\widehat{\theta} = \begin{bmatrix} 5{,}750 \\ 0{,}122 \end{bmatrix}, \quad \widehat{K} = \begin{bmatrix} 1{,}566 & -0{,}06 \\ -0{,}06 & 0{,}003 \end{bmatrix};$$

для $\rho_2 = -0{,}5$

$$\widehat{\theta} = \begin{bmatrix} 5{,}725 \\ 0{,}121 \end{bmatrix}, \quad \widehat{K} = \begin{bmatrix} 0{,}361 & -0{,}015 \\ -0{,}015 & 0{,}0007 \end{bmatrix};$$

для $\rho_3 = 0{,}9$

$$\widehat{\theta} = \begin{bmatrix} 5{,}868 \\ 0{,}120 \end{bmatrix}, \quad \widehat{K} = \begin{bmatrix} 6{,}700 & -0{,}146 \\ -0{,}146 & 0{,}0068 \end{bmatrix};$$

для $\rho_4 = -0{,}9$

$$\widehat{\theta} = \begin{bmatrix} 5{,}648 \\ 0{,}128 \end{bmatrix}, \quad \widehat{K} = \begin{bmatrix} 0{,}264 & -0{,}01 \\ -0{,}01 & 0{,}0005 \end{bmatrix}. \ ■$$

П р и м е р 11.4. Сравните дисперсии МНК-оценки и ОМНК-оценки параметра θ в модели регрессии вида

$$X_k = h_k\theta + \varepsilon_k, \quad k = 1, \ldots, n,$$

где ошибки ε_k описываются моделью (11.7) с коэффициентом $\rho = 0{,}8$.

Р е ш е н и е. МНК-оценка, вычисленная по формуле (10.5), в данном случае имеет вид

$$\widehat{\theta}_n = \left(H_n^\top H_n\right)^{-1} H_n^\top Z_n = \frac{\sum\limits_{k=1}^{n} h_k X_k}{\sum\limits_{k=1}^{n} h_k^2},$$

где $H_n = [h_1, \ldots, h_n]^\top$.

Вычислим дисперсию ошибки этой оценки, пренебрегая корреляцией ошибок в наблюдениях:

$$D_1 = \frac{\sigma_\varepsilon^2}{\sum\limits_{k=1}^{n} h_k^2}. \tag{11.19}$$

Вычислим дисперсию ОМНК-оценки:

$$D_2 = (H_n^\top H_n)^{-1} H_n^\top V_n H_n (H_n^\top H_n)^{-1} =$$

$$= \frac{\sigma_\varepsilon^2}{\sum\limits_{k=1}^{n} h_k^2} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}^\top \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} \frac{1}{\sum\limits_{k=1}^{n} h_k^2} = \tag{11.20}$$

$$= \frac{\sigma_\varepsilon^2}{\sum\limits_{k=1}^{n} h_k^2} \left(1 + 2\rho \frac{\sum\limits_{k=1}^{n-1} h_k h_{k+1}}{\sum\limits_{k=1}^{n} h_k^2} + \ldots + 2\rho^{n-1} \frac{h_1 h_n}{\sum\limits_{k=1}^{n} h_k^2} \right).$$

Возьмем для сравнения результатов (11.19) и (11.20) их отношение:

$$\frac{D_2}{D_1} = 1 + 2\rho \frac{\sum\limits_{k=1}^{n-1} h_k h_{k+1}}{\sum\limits_{k=1}^{n} h_k^2} + \ldots + 2\rho^{n-1} \frac{h_1 h_n}{\sum\limits_{k=1}^{n} h_k^2}. \tag{11.21}$$

Очевидно, что правая часть (11.21) может существенно превышать единицу, например, для $n = 11$, $\rho = 0{,}8$ и $h = [1, 2, \ldots, 11]^\top$ величина $\dfrac{D_2}{D_1} = 4{,}93$, если $h = [-5, -4, \ldots, 0, 1, \ldots, 5]^\top$, то $\dfrac{D_2}{D_1} = 2{,}24$.

Мы видим, что игнорирование коррелированности ошибок в линейной модели регрессии при подсчете дисперсии ошибки оценки может привести к занижению этой дисперсии более, чем в четыре раза. ■

Пример 11.5. Для потребительской функции США из примера 10.6 проверьте с помощью критерия Дарбина—Уотсона гипотезу H_0: $\rho = 0$ на уровне значимости $\alpha = 0{,}05$. Считать, что ошибки в наблюдениях имеют распределение $\mathcal{N}(0; \sigma^2)$.

Решение. В примере 10.6 была построена линия регрессии $X = -2{,}927 + 0{,}928h$. Найдем оценки ошибок в наблюдениях $\widehat{\varepsilon}_k = X_k - \widehat{X}_k$, $k = 1, \ldots, 12$:

$$\widehat{\varepsilon} = [0{,}52,\ 1{,}82,\ -1{,}87,\ -2{,}72,\ -2{,}99,\ 1{,}31,\ 3{,}25,\ 0{,}26,$$
$$0{,}78,\ 2{,}23,\ -1{,}46,\ -1{,}14]^\top .$$

Далее вычислим реализацию статистики критерия Дарбина—Уотсона

$$D = \frac{\sum\limits_{k=2}^{12} \left(\widehat{\varepsilon}_k - \widehat{\varepsilon}_{k-1}\right)^2}{\sum\limits_{k=1}^{12} \widehat{\varepsilon}_k^2} =$$
$$= \frac{(1{,}82 - 0{,}52)^2 + (-1{,}87 - 1{,}82)^2 + \cdots + (-1{,}14 + 1{,}46)^2}{0{,}52^2 + 1{,}82^2 + \cdots + 1{,}14^2} = \frac{63{,}39}{44{,}79} = 1{,}415.$$

Теперь по табл. 22.5 вычислим значения d_l и d_u на уровне значимости 0,05. Поскольку для $n < 15$ данные в таблице отсутствуют, то экстраполируем нужные значения:

$$d_l(12) = d_l(15) - 0{,}03 - 0{,}03 - 0{,}03 = 1{,}08 - 0{,}09 = 0{,}99,$$
$$d_u(12) = d_u(15) - 0{,}01 - 0{,}01 - 0{,}01 = 1{,}36 - 0{,}03 = 1{,}33.$$

Так как $4 - d_u(12) = 2{,}67$ и $D = 1{,}415 \in (d_u; 4 - d_u) = (1{,}33; 2{,}67)$, то в соответствии с табл. 11.1 гипотеза H_0 принимается на уровне значимости 0,05. ■

Пример 11.6. Пусть наблюдения, представленные в табл. 11.3, описываются уравнением линейной регрессии

$$X_k = \theta_1 + \theta_2 h_k + \varepsilon_k, \quad k = 1, \ldots, n,$$

где $\varepsilon_k \sim \mathcal{N}(0; \sigma^2)$ для любого $k = 1, \ldots, n$.

Таблица 11.3

h_k	0,10	0,15	0,20	0,25	0,3	0,35	0,4	0,45	0,5
X_k	0,019	0,019	0,027	0,051	0,093	0,136	0,171	0,198	0,267
h_k	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
X_k	0,314	0,365	0,396	0,482	0,569	0,627	0,710	0,835	0,913

Являются ли коррелированными ошибки наблюдений? Найдите также оценки неизвестных параметров θ_1 и θ_2.

Решение. Применим критерий Дарбина—Уотсона.

Предположим, что последовательность случайных ошибок $\{\varepsilon_k\}$ удовлетворяет уравнению

$$\varepsilon_k = \rho\varepsilon_{k-1} + e_k, \quad k = 2, \ldots, n,$$

где e_k — центрированные гауссовские СВ с дисперсией $\mathbf{D}\{e_k\} = \sigma_0^2$. Сначала вычислим реализацию МНК-оценки вектора неизвестных параметров:

$$\widehat{\theta} = [-0{,}209, 1{,}053]^T.$$

Используя найденную оценку, вычислим оценку вектора ошибок:

$$\widehat{\varepsilon} = [\widehat{\varepsilon}_1, \ldots \widehat{\varepsilon}_{18}]^T = [-0{,}014, -0{,}024, -0{,}041, -0{,}067, -0{,}05, -0{,}056, \\ -0{,}058, -0{,}079, -0{,}046, -0{,}012, -0{,}006, 0{,}002, 0{,}096, 0{,}122,]^\top.$$

Тогда реализация статистики критерия равна $D = 0{,}238$. По табл. 22.5 для $p = 1$, $\alpha = 0{,}05$ границы принятия решения равны $d_l = 1{,}03$, $d_u = 1{,}26$. Следовательно (см. табл. 11.1), $D \in (0; d_l) = (0; 1{,}03)$ и гипотеза H_0: $\rho = 0$ отвергается на уровне значимости 0,05 и принимается H_1: $\rho > 0$.

Теперь построим оценки неизвестных параметров модели с помощью процедуры А.

Процесс поиска оценки коэффициента ρ носит итерационный характер: сначала вычисляется

$$\widehat{\rho} = \frac{\sum\limits_{k=2}^{n} \widehat{\varepsilon}_{k-1}\widehat{\varepsilon}_k}{\sum\limits_{k=2}^{n} \left(\widehat{\varepsilon}_{k-1}\right)^2}, \tag{11.22}$$

а затем вычисляется ОМНК-оценка вектора θ и строится новый вектор оценок остатков $\widehat{\varepsilon}$. После этого все шаги повторяются заново. Первое значение реализации оценки коэффициента автокорреляции равно $\widehat{\rho} = 0{,}846$. Реализация оценки неизвестной дисперсии σ_0^2 равна:

$$\widehat{\sigma}_0^2 = \frac{1}{18-1}\sum_{k=2}^{18}\left(\widehat{\varepsilon}_k - \widehat{\rho}\widehat{\varepsilon}_{k-1}\right)^2 = 0{,}00090.$$

Далее строится матрица V_n по формуле (11.2) и вычисляется реализация ОМНК-оценки по формуле (11.1):

$$\widehat{\theta}^\circ = \{-0{,}162, 1{,}052\}^T.$$

Найденная оценка используется для построения новой реализации оценки вектора ε, и все шаги повторяются до тех пор, пока значение модуля разности $\Delta = |\widehat{\rho}_i - \widehat{\rho}_{i-1}|$, где i — номер итерации, не станет меньше 0,0005. Результаты вычислений представлены в табл. 11.4.

Таблица 11.4

$\widehat{\rho}$	$\widehat{\sigma}_0^2$	$\widehat{\theta}_1^\circ$	$\widehat{\theta}_2^\circ$	Δ
0,919	0,00094	-0,1403	1,0519	0,073
0,944	0,00095	-0,1285	1,05186	0,025
0,954	0,00096	-0,12293	1,05182	0,01
0,9576	0,00096	-0,12099	1,05182	0,0036
0,9595	0,00096	-0,11980	1,05182	0,0019
0,9602	0,00096	-0,11993	1,05182	0,0007
0,9605	0,00096	-0,11915	1,05182	0,0003

■

11.6. Задачи для самостоятельного решения

1. Найдите ОМНК-оценку и дисперсию ее компонент в случае, когда ковариационная матрица ошибок V_n является диагональной.

2. Получите матрицу вида (11.12).

3. Пусть $X_k = \theta_1 + \theta_2 h_k + \varepsilon_k$, $k = 1, 2, 3$, где вектор ошибок $E = \{\varepsilon_1, \varepsilon_2, \varepsilon_3\}^T$ имеет математическое ожидание $\mathbf{M}\{E\} = 0$, а $\mathbf{cov}(E, E) = \sigma^2 V$,

$$V = \begin{bmatrix} 1 & \rho a & \rho \\ \rho a & a^2 & \rho a \\ \rho & \rho a & 1 \end{bmatrix},$$

числа a и $0 < \rho < 1$ неизвестны, $h_1 = -1$, $h_2 = 0$, $h_3 = 1$. Покажите, что ОМНК-оценка вектора θ имеет вид

$$\widehat{\theta} = \begin{bmatrix} \frac{1}{r}\big((a^2 - a\rho)X_1 + (1 - 2a\rho + \rho)X_2 + (a^2 - a\rho)X_3\big) \\ -0{,}5X_1 + 0{,}5X_3 \end{bmatrix},$$

где $r = 1 + \rho + 2a^2 - 4a\rho$. Найдите такие значения ρ, a, $X_k, k = 1, 2, 3$, при которых линия регрессии лежит целиком ниже и целиком выше всех наблюдавшихся значений.

4. Пусть $X_i = \theta h_i + \varepsilon_i$, $i = 1, 2$, где $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$, $\varepsilon_2 \sim \mathcal{N}(0, 4\sigma^2)$, причем ε_1 и ε_2 независимы. Для двух векторов h: $h^{(1)} = [-1, 1]^\top$, $h^{(2)} = [1, -1]^\top$ найдите ОМНК-оценку и ее дисперсию для двух векторов h.

5. Пусть модель регрессии имеет вид

$$X_k = \theta_1 + \theta_2 h_k + \varepsilon_k, \quad k = 1, \ldots 5,$$

где ε_k — независимые, центрированные СВ с дисперсиями $\sigma_k^2 = h_k^2$. По наблюдениям, приведенным в табл. 11.5 найдите реализации МНК-оценки и ОМНК-оценки неизвестных параметров $[\theta_1, \theta_2]^\top$. Для каждой оценки вычислите ковариационную матрицу ее ошибки.

Таблица 11.5

h_k	1	2	3	4	5
X_k	1	2	2	3	4

6. Покажите, что использование ОМНК для модели простой линейной регрессии с дисперсиями ошибок, равными $\mathbf{D}\{e_k\} = \sigma^2 h_k^2$, равносильно использованию обычного МНК для модели наблюдений вида

$$\frac{X_k}{h_k} = \frac{\theta_1}{h_k} + \theta_2 + \delta_k, \quad k = 1, \ldots, n,$$

где δ_k — независимые СВ с дисперсиями $\mathbf{D}\{\delta_k\} = \sigma^2$ для $k = 1, \ldots, n$.

7. Пусть задана модель регрессии

$$X_k = \theta_1 + \theta_2 h_k + \varepsilon_k, \quad k = 1, \ldots 10,$$

где ошибки ε_k удовлетворяют уравнению авторегрессии первого порядка

$$\varepsilon_k = -0{,}4\varepsilon_{k-1} + v_k, \quad k = 2, \ldots, 10.$$

где v_k — последовательность независимых нормально распределенных случайных величин с нулевым средним и дисперсией $\sigma_0^2 = 1$. Найдите реализацию ОМНК-оценки вектора неизвестных параметров $\theta = [\theta_1, \theta_2]^\top$ по наблюдениям, приведенным в табл. 11.6.

Таблица 11.6

h_k	5,0	2,5	1,8	6,8	9,0	3,8	6,5	9,0	1,0	3,5
X_k	5,0	4,8	3,1	8,2	8,6	5,5	6,5	11,1	2,1	4,5

Сравните полученный вектор с реализацией МНК-оценки, построенной в предположении некоррелированности ошибок ε_k.

8. Рассмотрите модель, связывающую количество вакансий w и уровень безработицы u

$$X_k = \theta_1 + \theta_2 h_k + \varepsilon_k, \quad k = 1, \ldots, 24,$$

где $X_k = \ln w_k$, $h_k = \ln u_k$, $\varepsilon_k \sim \mathcal{N}(0; \sigma^2)$.

Используя данные из табл. 11.7, найдите МНК-оценки параметров θ_1 и θ_2. Проверьте с помощью критерия Дарбина-Уотсона гипотезу H_0: $\rho = 0$ о некоррелированности ошибок наблюдений на уровне значимости $\alpha = 0{,}05$. Если гипотеза H_0 отвергается, то предположите, что последовательность случайных ошибок $\{\varepsilon_k\}$ удовлетворяет уравнению авторегрессии первого порядка

$$\varepsilon_k = \rho\varepsilon_{k-1} + e_k, \quad k = 2, \ldots, n,$$

где e_k — центрированные гауссовские СВ с дисперсией $\mathbf{D}\{e_k\} = \sigma_0^2$. Найдите оценку неизвестного коэффициента ρ в уравнении авторегрессии.

Таблица 11.7

k	w_k	u_k	k	w_k	u_k	k	w_k	u_k
1	1,73	8,65	9	5,06	2,87	17	3,15	4,72
2	1,94	4,82	10	2,81	5,29	18	1,92	7,45
3	3,05	2,67	11	4,43	3,31	19	2,26	6,21
4	4,17	2,67	12	3,19	5,44	20	6,18	2,64
5	2,52	2,58	13	2,23	6,80	21	2,07	8,55
6	1,71	8,07	14	2,06	8,25	22	8,39	2,60
7	1,95	8,83	15	3,33	3,44	23	2,75	6,25
8	2,57	5,54	16	2,12	7,80	24	6,10	2,70

9. Исследуйте потребительскую функцию Японии $X(h)$ за период 1951—1962 гг. на наличие коррелированности ошибок наблюдений. Если ошибки в наблюдениях являются коррелированными, то найдите реализацию ОМНК-оценки вектора неизвестных параметров. Найдите оценку ковариационной матрицы вектора ошибок E. Данные для расчетов приведены в табл. 11.8 (данные взяты из: U. N. Yearbook of National Accounts Statistics, 1963). Здесь X — совокупное индивидуальное потребление (трлн иен), h — совокупные личные доходы (трлн иен).

Таблица 11.8

k	X_k	h_k	k	X_k	h_k
1	2,86	3,52	7	5,89	7,00
2	3,51	4,26	8	6,20	7,32
3	4,22	4,77	9	6,70	8,24
4	4,67	5,25	10	7,51	9,46
5	4,97	5,73	11	8,58	11,10
6	5,43	6,35	12	9,96	12,69

12. Гетероскедастичность

12.1. Гетероскедастичность в ошибках

В предыдущем параграфе была описана линейная модель регрессии с гетероскедастичными ошибками как один из частных случаев обобщенной линейной регрессионной модели, в котором матрица V_n имеет вид (11.3).

Попытаемся конкретизировать общие формулы ОМНК для данного частного случая. С этой целью, во-первых, выпишем конкретный вид матрицы C вспомогательного преобразования (см. пример 11.1):

$$C = \begin{bmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_n^{-1} \end{bmatrix}. \tag{12.1}$$

Легко проверить, что в этом случае выполняются соотношения (11.16) и (11.17). Кроме того, определенная соотношением (11.6) функция потерь ОМНК для V_n вида (11.3) представим в виде

$$J(\theta) = (Z_n - H_n\theta)^\top V_n^{-1}(Z_n - H_n\theta) = \sum_{k=1}^{n} \frac{1}{\sigma_k^2} \left(X_k - h_k^\top\theta\right)^2, \quad (12.2)$$

а ОМНК-оценка имеет вид

$$\widehat{\theta}_n = \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} h_k h_k^\top\right)^{-1} \sum_{k=1}^{n} \frac{1}{\sigma_k^2} h_k^\top X_k. \quad (12.3)$$

Определение 12.1. Оценка (12.3) называется *оценкой метода взвешенных наименьших квадратов.*

Роль «весов» здесь играют диагональные элементы матрицы V_n^{-1}, т.е. σ_k^{-2}.

Если σ_k неизвестны, то необходимо их оценить. Так как число параметров равно n, то без дополнительных предположений о структуре матрицы V_n мы не сможем получить удовлетворительные оценки. Ниже рассматриваются несколько классов моделей с гетероскедастичностью, где наложены такие предположения, благодаря которым удается построить оценку матрицы V_n, а следовательно, ОМНК-оценку $\widehat{\theta}_n$.

1. *Дисперсии ошибок пропорциональны объясняющей переменной.* В некоторых ситуациях априорно можно считать, что среднее квадратическое отклонение ошибок прямо пропорционально одной из объясняющих переменных, например, $h_{k,i}$: $\sigma_k = \sigma h_{k,i}$, $k = 1, \ldots, n$, где $h_{k,i}$ — i-й элемент вектора h_k, $\sigma > 0$, $h_{k,i} > 0$ для всех $k = 1, \ldots, n$. Тогда, разделив уравнение с номером k на $h_{k,i}$ и введя новые переменые $h_{k,j}^*(t) = \dfrac{h_{k,j}}{h_{k,i}}$, $j = 1, \ldots, p$, $X_k^* = \dfrac{X_k}{h_{k,i}}$, $k = 1, \ldots, n$, получим обычную регрессионную модель.

МНК-оценка в модели с преобразованными величинами дает непосредственно оценку параметров исходной модели. Однако следует помнить, что если первый столбец в матрице H_n состоит из единиц, то оценки свободного члена и параметра при $h_{k,j}^*(t) = \dfrac{1}{h_{k,i}}$ в новой модели являются оценками соответственно параметра θ_i и свободного члена в исходной модели.

Ниже будут приведены два статистических критерия проверки гипотезы на гетероскедастичность, а пока ограничимся практическим рекомендациями по применению описанного метода. Если есть предположение о зависимости ошибок от одной из объясняющих переменных, то целесообразно расположить наблюдения в порядке возрастания значений этой переменной, а затем построить обычную регрессию и получить регрессионные остатки. Если размах их колебаний тоже возрастает, то это говорит в пользу сделанного предположения. Тогда нужно провести описанное выше преобразование, вновь построить регрессию и исследовать остатки. Если теперь их колебания имеют неупорядоченный характер, то это указывает на то, что коррекция гетероскедастичности прошла успешно. Естественно, следует сравнивать и другие характеристики регрессии и только тогда принимать окончательное решение о том, какая из моделей более адекватна.

2. *Дисперсия ошибки принимает только два значения.* Пусть известно, что $\sigma_k^2 = w_1^2$ для $k = 1, \ldots, n_1$ и $\sigma_k^2 = w_2^2$ для $k = n_1 + 1, \ldots, n$, но значения w_1^2 и w_2^2 неизвестны. Иными словами, в первых n_1 наблюдениях дисперсия ошибки имеет одно значение, в последующих $n_2 = n - n_1$ — другое. В этом случае необходимо действовать по следующему алгоритму:

1) постройте обычную регрессию (10.5), получите оценку вектора ошибок $\widehat{E}_n$ и разбейте его на два подвектора e_1 и e_2 размера n_1 и n_2 соответственно;

2) постройте оценки $\widehat{w}_1^2 = \dfrac{1}{n_1} e_1^\top e_1$ и $\widehat{w}_2^2 = \dfrac{1}{n_2} e_2^\top e_2$ дисперсий w_1^2 и w_2^2;

3) преобразуйте переменные, разделив первые n_1 уравнений на $\widehat{w}_1$, а последующие — на $\widehat{w}_2$;

4) постройте обычную регрессию для преобразованной модели.

Оценки $\widehat{w}_1$ и $\widehat{w}_2$ являются смещенными, но состоятельными [48].

Очевидно, что модель допускает обобщение на случай, когда дисперсия принимает любое конечное число значений, не зависящее от n.

3. *Состоятельное оценивание дисперсий.* Рассмотрим общий случай (11.3). МНК-оценку можно представить в виде $\widehat{\theta}_n = \theta + W_n^{-1} H_n^\top Z_n$. Выпишем ковариационную матрицу ошибки оценки

$$\widehat{K}_n = \mathbf{M}\{W_n^{-1} H_n^\top E_n E_n^\top H_n W_n^{-1}\} = W_n^{-1} H_n^\top V_n H_n W_n^{-1} = \\ = n W_n^{-1} \left[\frac{1}{n} H_n^\top V_n H_n\right] W_n^{-1}, \text{ где } W_n = H_n^\top H_n.$$

Поскольку V_n диагональная, то $H_n^\top V_n H_n = \sum\limits_{k=1}^{n} \sigma_k^2 h_k h_k^\top$. Тогда, заменяя σ_k^2 на ε_k, где ε_k — k-й элемент вектора остатков $\widehat{E}_n = Z_n - H_n \widehat{\theta}_n$, получаем оценку для $\widehat{K}_n$:

$$\widehat{K}_n^* = n W_n^{-1} \left(\frac{1}{n} \sum_{k=1}^{n} \varepsilon_k^2 h_k h_k^\top \right) W_n^{-1}. \tag{12.4}$$

Можно показать, что $\widehat{K}_n^*$, задаваемая выражением (12.4), является состоятельной оценкой матрицы ковариаций ОМНК-оценки при наличии гетероскедастичности.

Для более сложного случая, когда в матрице V_n ненулевые элементы стоят не только на главной диагонали, тоже существуют оценки матрицы V_n. Однако сходятся они очень медленно и для практического использования малопригодны.

12.2. Тесты на гетероскедастичность

Опишем несколько статистических тестов, которые позволяют определить наличие гетероскедастичности в модели линейной регрессии. Во всех тестах проверяется основная гипотеза H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$ против альтернативы H_1: $\exists 1 \leqslant i \neq j \leqslant n$ такие, что $\sigma_i^2 \neq \sigma_j^2$.

Большинство тестов предполагают, что относительно характера гетероскедастичности есть достоверная априорная информация.

Критерий Бартлетта [42]. Предположим, что мы можем разделить выборку объема n на r независимых подвыборок объема n_i $(n_1 + \ldots + n_r = n)$, причем в каждой из них значения объясняющих переменных или совпадают, или принадлежат одним интервалам.

Вычислим оценки дисперсии для каждой группы

$$\widetilde{S}_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(X_{ki} - \overline{X}_i\right)^2, \quad i = 1, \ldots, r,$$

где X_{ki} — значение наблюдений в i-й подвыборке, а $\overline{X}_i = \dfrac{1}{n_i} \sum\limits_{k=1}^{n_i} X_{ki}$.

Далее вычисляется статистика критерия Бартлетта $\dfrac{Q}{l}$, где

$$Q = n \ln\left(\sum_{i=1}^{r} \frac{n_i}{n} \widetilde{S}_i^2\right) - \sum_{i=1}^{r} n_i \ln \widetilde{S}_i^2;$$

$$l = 1 + \frac{1}{3(r-1)} \left(\sum_{i=1}^{r} \frac{1}{n_i} - \frac{1}{n}\right).$$

Если гипотеза H_0 справедлива и случайные ошибки ε_k независимы и имеют нормальное распределение, то статистика $\dfrac{Q}{l}$ распределена примерно как хи-квадрат с $(r-1)$ степенью свободы, т.е. можно считать, что $\dfrac{Q}{l} \sim \mathcal{H}_{r-1}$.

Критическая область уровня значимости α имеет вид $(k_{1-\alpha}(r-1); +\infty)$, где $k_{1-\alpha}(r-1)$ — квантиль уровня $1-\alpha$ распределения $\mathcal{H}_{r-1}$.

К сожалению, этот критерий слишком чувствителен к любому отклонению от нормальности ошибок ε_k. Значимость статистики

Бартлетта может указывать не на отсутствие гетероскедастичности, а просто на отклонение от нормальности.

Поскольку в практических задачах все условия, описанные выше, редко выполняются, то проверку с помощью теста Бартлетта следует считать приближенной.

Критерий Голдфелда–Куандта [47]. Этот критерий применяется, как правило, когда есть предположение о прямой зависимости дисперсии ошибки от величины некоторой объясняющей переменной. Критерий состоит из следующих шагов:

1) упорядочьте данные по убыванию выбранной переменной;

2) исключите d средних наблюдений из этой последовательности (d выбирается так, чтобы $\frac{n-d}{2} > p$, где p — число оцениваемых параметров);

3) постройте две регрессии по первым и последним $\frac{n-d}{2}$ наблюдениям и вычислите соответствующие вектора оценок остатков e_1 и e_2;

4) вычислите статистику критерия

$$F(n,d) = \frac{S_1}{S_2} = \frac{e_1^\top e_1}{e_2^\top e_2}. \tag{12.5}$$

Если распределение ошибок ε_k является нормальным, а сами ошибки некоррелированные, то статистика $F(n,d)$ имеет F-распределение Фишера $F(m;s)$ с $m = s = \frac{n-d}{2} - p$ степенями свободы;

Критическая область для уровня значимости α имеет вид $(f_{1-\alpha}(m;s); +\infty)$, где $f_{1-\alpha}(m;s)$ — квантиль уровня $1-\alpha$ распределения Фишера $F(m;s)$.

З а м е ч а н и е. Надежность проверки на гетероскедастичность зависит от выбора числа d. Для больших значений d надежность проверки очень мала. Однако, при выборе небольших значений d остаточные дисперсии S_1 и S_2 начинают приближаются друг к другу, что может привести к тому, что различие между ними (в случае гетероскедастичности) выявлено не будет.

12.3. Примеры

П р и м е р 12.1. В табл. 12.1 приведены данные о размерах индивидуального потребления X и личных доходов h (млрд франков) в Бельгии за период с 1951 г. по 1962 г. (данные взяты из: U. N. Yearbook of National Accounts Statistics, 1963). На уровне значимости $\alpha = 0{,}05$ проверьте гипотезу H_0: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2$ об отсутствии гетероскедастичности в модели наблюдений

$$X_k = \theta_1 + \theta_2 h_k + \varepsilon_k, \quad k = 1, \dots, 12,$$

где θ_1, θ_2 — неизвестные параметры модели; $\varepsilon_k \sim \mathcal{N}(0; \sigma_k^2)$.

Таблица 12.1

k	X	h	k	X	h
1	297,0	331,4	7	358,5	398,6
2	303,8	333,2	8	358,0	410,2
3	308,1	338,1	9	378,7	417,7
4	325,2	360,4	10	391,7	445,9
5	339,8	378,2	11	413,1	462,7
6	338,6	375,7	12	432,8	486,8

Решение. Сначала проверим гипотезу H_0 с помощью критерия Бартлетта (см. разд. 12.2). Пусть число групп $r = 2$. Разобьем выборку на две части: в первой группе семь первых наблюдений ($n_1 = 7$), во второй — остальные ($n_2 = 5$).

Вычислим в каждой группе реализацию оценки дисперсии по формуле

$$\widetilde{S}_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(X_{ki} - \overline{X}_i\right)^2, \quad i = 1, 2, \tag{12.6}$$

где X_{ki} — k-е наблюдение в i-й подвыборке, а $\overline{X}_i = \dfrac{1}{n_i} \sum\limits_{k=1}^{n_i} X_{ki}$.

Тогда

$$\overline{X}_1 = \frac{1}{7}(297{,}0 + 303{,}8 + 308{,}1 + 325{,}2 + 339{,}8 + 338{,}6 + 358{,}5) = 324{,}43,$$

Далее по формуле (12.6) получаем:

$$\widetilde{S}_1^2 = \frac{1}{7-1}\big[(297{,}0 - 324{,}43)^2 + (303{,}8 - 324{,}43)^2 + (308{,}1 - 324{,}43)^2 + \\ + (325{,}2 - 324{,}43)^2 + (339{,}8 - 324{,}43)^2 + (338{,}6 - 324{,}43)^2 + \\ + (358{,}5 - 324{,}43)^2\big] = 507{,}176.$$

Аналогично, для второй подвыборки $\widetilde{S}_2^2 = 850{,}483$.

Теперь вычислим реализацию статистики критерия Бартлетта $\dfrac{Q}{l}$, где

$$Q = n \ln\left(\sum_{i=1}^{r} \frac{n_i}{n}\widetilde{S}_i^2\right) - \sum_{i=1}^{r} n_i \ln \widetilde{S}_i^2 = 12 \ln\left(\frac{7}{12}507{,}18 + \frac{5}{12}850{,}48\right) - \\ - (7 \ln 507{,}18 + 5 \ln 850{,}48) = 0{,}379.$$

$$l = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r} \frac{1}{n_i} - \frac{1}{n}\right) = 1 + \frac{1}{3(2-1)}\left(\frac{1}{7} + \frac{1}{5} - \frac{1}{12}\right) = 1{,}086.$$

Таким образом, $\dfrac{Q}{l} = \dfrac{0{,}379}{1{,}086} = 0{,}349$.

По условию $\alpha = 0{,}05$. Тогда критическая область $(k_{1-\alpha}(r-1);+\infty) = (3{,}84;+\infty)$, где $k_{1-\alpha}(r-1) = 3{,}84$ — квантиль уровня 0,95 распределения хи-квадрат с $r-1 = 2-1 = 1$ степенью свободы, найденная по табл. 22.3. В результате получили, что $0{,}349 \notin (3{,}84;+\infty)$, следовательно, гипотеза H_0: $\sigma_1^2 = \sigma_2^2$ принимается на уровне значимости 0,05.

Теперь проверим гипотезу H_0 с помощью критерия Голдфельда—Куандта. Упорядочим наблюдения по убыванию переменной h и исключим два центральных наблюдения, т.е. $d = 2$. В результате получим две группы наблюдений. По наблюдениям в каждой группе нужно вычислить МНК-оценки неизвестных параметров и найти сумму квадратов оценок ошибок $\widehat{\varepsilon}_k^2$. Результаты вычислений приведем в виде табл. 12.2.

Таблица 12.2

Регрессия	h_k	X_k	$\widehat{X}_k$	$\widehat{\varepsilon}_k$	$\widehat{\varepsilon}_k^2$
I: $X = -6{,}93 + 0{,}9h$	486,8	432,8	432,9	-0,14	0,02
	462,7	413,1	411,2	1,94	3,76
	445,9	391,7	396,0	-4,28	18,32
	417,7	378,7	370,5	8,20	67,24
	410,2	358,0	363,7	-5,72	32,72
II: $X = -6{,}21 + 0{,}87h$	375,7	338,6	338,8	-0,17	0,03
	360,4	325,2	325,5	-0,30	0,09
	338,1	308,1	306,2	1,94	3,76
	333,2	303,8	301,9	1,89	3,57
	331,4	297,0	300,4	-3,35	11,22

Суммируя значения в последнем столбце таблицы, получим числитель и знаменатель статистики критерия Голдфельда—Куандта (12.5)

$$F(n;d) = \frac{\sum\limits_{k=1}^{5} \left(\widehat{\varepsilon}_k^2\right)_I}{\sum\limits_{k=1}^{5} \left(\widehat{\varepsilon}_k^2\right)_{II}} = \frac{122{,}1}{18{,}97} = 6{,}55.$$

Критическая область уровня значимости 0,05 имеет вид $(f_{0,95}(m,s);+\infty)$, где $f_{0,95}(m,s) = 9{,}28$ — квантиль уровня 0,95 распределения Фишера $F(m,s)$ с $m = s = \dfrac{n-d}{2} - p = \dfrac{12-2}{2} - 2 = 3$ степенями свободы. Реализация статистики критерия 6,55 не попадает в критическую область и гипотеза H_0 принимается на уровне значимости 0,05. ■

П р и м е р 12.2. По выборке для 30 стран (табл. 12.3) за 1980 г. постройте зависимость государственных расходов на образование (X)

от валового внутреннего продукта (h_1) и численности населения (h_2), т.е. $X = f(h_1, h_2)$. (Данные взяты из статистического ежегодника ЮНЕСКО «Statistical Yearbook» за 1984 г. и из источника Международного валютного фонда «International Financial Statistics» за 1984 г.) Покажите, что модель гетероскедастична. Величины (X) и (h_1) измеряются в миллиардах долларов, а население — в миллионах человек. Оцените параметры модели после устранения гетероскедастичности.

Таблица 12.3

Страна	X	h_1	h_2
Люксембург	0,34	5,67	0,36
Уругвай	0,22	10,1	2,90
Сингапур	0,32	11,3	2,39
Ирландия	1,23	18,9	3,44
Израиль	1,81	20,9	3,87
Венгрия	1,02	22,1	10,7
Новая Зеландия	1,27	23,8	3,10
Гонконг	0,67	27,6	5,07
Чили	1,25	27,6	11,1
Греция	0,75	40,1	9,60
Финляндия	2,80	51,6	4,78
Норвегия	4,90	57,7	4,09
Дания	4,45	66,3	5,12
Турция	1,60	67,0	44,9
Швейцария	5,31	101,7	6,37
Бельгия	7,15	119,5	9,86
Швеция	11,2	124,1	8,31
Австралия	8,66	141,0	14,6
Аргентина	5,56	153,8	27,1
Нидерланды	13,4	169,4	14,1
Мексика	5,46	186,3	67,4
Испания	4,79	211,8	37,4
Бразилия	8,92	249,7	123,0
Канада	18,9	261,4	23,9
Италия	15,9	395,5	57,0
Великобритания	29,9	534,9	55,9
Франция	33,6	655,3	53,7
ФРГ	38,6	815,0	61,6
Япония	61,6	1040,4	116,8
США	181,3	2586,4	227,6

Решение. Вычислим МНК-оценку неизвестных параметров $\theta = \{\theta_1, \theta_2, \theta_3\}^\top$ в модели

$$X_k = \theta_1 + \theta_2 h_{1,k} + \theta_3 h_{2,k} + \varepsilon_k, \quad k = 1, \ldots, 30,$$

где $\{\varepsilon_k\}$ — случайные величины с распределением $\mathcal{N}(0; \sigma_k^2)$. Реализации МНК-оценки и ковариационной матрицы ошибки оценки равны:

$$\widehat{\theta} = [-1{,}696, 0{,}074, -0{,}081]^\top, \tag{12.7}$$

$$\widehat{K} = \begin{bmatrix} 1{,}102 & 6{,}09 \cdot 10^{-4} & -0{,}016 \\ 6{,}09 \cdot 10^{-4} & 1{,}50 \cdot 10^{-5} & -1{,}39 \cdot 10^{-4} \\ -0{,}016 & -1{,}39 \cdot 10^{-4} & 1{,}60 \cdot 10^{-3} \end{bmatrix}. \tag{12.8}$$

Теперь проверим с помощью критерия Бартлетта наличие гетероскедастичности. Для этого разобьем выборку на три подвыборки ($r = 3$) так, чтобы значения переменных h_1 и h_2 в каждой подвыборке были примерно одинаковыми. В результате получим:

$$X1 = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_7 \\ X_8 \\ X_{11} \\ X_{12} \\ X_{13} \end{bmatrix} = \begin{bmatrix} 0{,}34 \\ 0{,}22 \\ 0{,}32 \\ 1{,}23 \\ 1{,}81 \\ 1{,}02 \\ 1{,}27 \\ 0{,}67 \\ 2{,}80 \\ 4{,}90 \\ 4{,}45 \end{bmatrix}, \qquad X2 = \begin{bmatrix} X_6 \\ X_9 \\ X_{10} \\ X_{14} \\ X_{15} \\ X_{16} \\ X_{17} \\ X_{20} \\ X_{22} \\ X_{24} \end{bmatrix} = \begin{bmatrix} 1{,}02 \\ 1{,}25 \\ 0{,}75 \\ 1{,}60 \\ 5{,}31 \\ 7{,}15 \\ 11{,}22 \\ 13{,}41 \\ 4{,}79 \\ 18{,}9 \end{bmatrix},$$

$$X3 = \begin{bmatrix} X_{18} \\ X_{19} \\ X_{21} \\ X_{23} \\ X_{25} \\ X_{26} \\ X_{27} \\ X_{28} \\ X_{29} \\ X_{30} \end{bmatrix} = \begin{bmatrix} 8{,}66 \\ 5{,}56 \\ 5{,}46 \\ 8{,}92 \\ 15{,}95 \\ 29{,}9 \\ 33{,}6 \\ 38{,}6 \\ 61{,}6 \\ 181{,}3 \end{bmatrix}.$$

Далее для каждой подвыборки вычислим выборочные дисперсии по формуле (12.6):

$$\widetilde{S}_1^2 = 2{,}926, \quad \widetilde{S}_2^2 = 38{,}133, \quad \widetilde{S}_3^2 = 2830.$$

Окончательно, статистика критерия Бартлетта равна

$$\frac{Q}{l} = \frac{79{,}29}{1{,}044} = 75{,}92.$$

Поскольку $k_{0,95}(2) = 5{,}99$ и $75{,}92 \in (5{,}99; +\infty)$, следовательно, гипотеза H_0: $\sigma_1^2 = \ldots = \sigma_n^2$ отвергается на уровне значимости 0,05.

Теперь применим к этим же данным тест Голдфельда—Куандта. Расположим наблюдения по убыванию переменной h_1 (ВВП) и исключим десять средних наблюдений, т.е. $d = 10$. Далее найдем МНК-оценки для двух групп наблюдений:

$$\widehat{\theta}_1 = \begin{bmatrix} -11{,}313 \\ 0{,}071 \\ 0{,}028 \end{bmatrix}, \qquad \widehat{\theta}_2 = \begin{bmatrix} 0{,}433 \\ 0{,}022 \\ -0{,}0013 \end{bmatrix}.$$

Используя найденные оценки, найдем для каждой регрессии оценки ошибок в наблюдениях $\widehat{\varepsilon}_k$ и вычислим реализацию статистики Голдфельда–Куандта (12.5) (подробнее см. пример 12.1):

$$F(n; d) = \frac{258{,}18}{1{,}96} = 131{,}53.$$

В нашем случае $m = s = \dfrac{n-d}{2} - p = \dfrac{30-10}{2} - 3 = 7$, поэтому по таблице из [7] $f_{0,95}(7; 7) = 3{,}79$. Поскольку $131{,}53 \in (3{,}79;\ +\infty)$, то гипотеза H_0 отвергается на уровне значимости 0,05.

Попробуем устранить гетероскедастичность. Для этого разделим каждое наблюдение на соответствующее значение переменной h_1 (ВВП). К полученным наблюдениям применим критерий Бартлетта. Также разобьем выборку на три подвыборки и для каждой подвыборки вычислим реализации оценок дисперсий по формуле (12.6):

$$\widetilde{S}_1^2 = 4{,}85 \cdot 10^{-4}, \quad \widetilde{S}_2^2 = 4{,}36 \cdot 10^{-4}, \quad \widetilde{S}_3^2 = 2{,}78 \cdot 10^{-4}.$$

Таким образом, реализация статистики критерия Бартлетта равна

$$\frac{Q}{l} = \frac{0{,}82}{1{,}04} = 0{,}78.$$

Следовательно, $0{,}78 \notin (5{,}99;\ +\infty)$ и гипотеза H_0 принимается на уровне значимости 0,05.

Теперь в «исправленной» регрессии вычислим реализации МНК-оценки неизвестных параметров и ковариационной матрицы ошибок:

$$\widetilde{\theta} = \begin{bmatrix} -0{,}057 \\ 0{,}056 \\ -0{,}018 \end{bmatrix}, \quad \widetilde{K} = \begin{bmatrix} 9{,}83 \cdot 10^{-3} & -2{,}34 \cdot 10^{-4} & -1{,}30 \cdot 10^{-4} \\ -2{,}34 \cdot 10^{-4} & 4{,}02 \cdot 10^{-5} & -1{,}10 \cdot 10^{-4} \\ -1{,}30 \cdot 10^{-4} & -1{,}10 \cdot 10^{-4} & 6{,}08 \cdot 10^{-4} \end{bmatrix}. \quad (12.9)$$

Сравнивая (12.7), (12.8) и (12.9), видим, что полученные реализации оценок параметров различаются достаточно сильно. При этом дисперсии оценок, построенных по преобразованным наблюдениям, значительно меньше, чем дисперсии оценок, построенных по исходным наблюдениям. ■

12.4. Задачи для самостоятельного решения

1. В табл. 12.4 приведены данные об инвестициях (X), государственных расходах (h_1), валовом внутреннем продукте (h_2) и численности населения (h_3) для 28 стран в 1997 г. Величины X, h_1 и h_2 приводятся в миллиардах долларов США, h_3 — в миллионах человек. Постройте зависимость объема инвестиций от объема государственных расходов и ВВП вида

$$X_k = \theta_1 + \theta_2 h_k^1 \theta_3 h_k^2 \theta_4 h_k^3 + \varepsilon_k, \quad k = 1, \ldots, 28,$$

где $\theta_1, \ldots, \theta_4$ — неизвестные параметры, $\varepsilon_k \sim \mathcal{N}(0; \sigma_k^2)$. Проверьте гипотезу H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{28}^2$ с помощью критерия Голдфельда—Куандта на уровне значимости 0,05.

Таблица 12.4

Страна	X	h_1	h_2	h_3	Страна	X	h_1	h_2	h_3
Австралия	94	76	408	18	Нидерланды	73	50	361	16
Австрия	46	39	206	8,1	Филиппины	20	11	82	79
Канада	119	125	631	30	Швейцария	50	39	256	7
Чехия	16	10	52	10	Южная Корея	155	49	442	46
Дания	34	43	169	5,3					
Франция	256	347	1409	57	Германия	422	407	2103	82
Греция	24	18	120	10	Исландия	1,4	1,5	7,5	0,3
Италия	191	190	1145	57	Ирландия	14	10	73	3,7
Япония	1106	376	3901	126	Малайзия	42	11	97	21
Норвегия	35	31	153	4,4	Сингапур	36	9,0	96	3,7
Польша	29	23	136	39	Испания	109	86	532	39
Россия	85	94	436	147	Таиланд	48	15	154	61
Швеция	31	59	228	8,9	Португалия	20	10	82	78
США	1518	1244	8111	268	Финляндия	20	25	122	5,1
Турция	50	23	189	62					

2. Для данных задачи 1 примените тест Бартлетта.

3. Для данных задачи 1 постройте зависимость инвестиций от государственных расходов и валового внутреннего продукта, разделив значения всех переменных на соответствующие значение численности населения. Для нового набора наблюдений примените тест Голдфельда—Куандта.

4. Исследуйте зависимость между величиной добавленной стоимости в обрабатывающей промышленности (X) и валовым внутренним продуктом (h) по выборке из 28 стран за 1994 г., включающей как малые страны, так и большие (UNIDO Yearbook 1997). В табл. 12.5 приведены значения X и h в миллионах долларов США, население (g) — в миллионах человек, данные последних двух столбцов — в долларах США на одного человека. Постройте МНК-оценку параметров θ_1 и θ_2 для модели

$$X_k = \theta_1 + \theta_2 h_k + \varepsilon_k, \quad k = 1, \ldots, 28.$$

Проверьте наличие гетероскедастичности с помощью одного из критериев в предположении, что ошибки в наблюдениях ε_k, $k = 1, \ldots, 28$ некоррелированны и имеют гауссовское распределение. Если гетероскедастичность обнаружена, то устраните ее, разделив каждое наблюдение на значение численности населения. Для этого можно воспользоваться данными из последних двух столбцов табл. 12.5.

Таблица 12.5

Страна	X	h	g	X/g	h/g
Бельгия	44 517	232 006	10,093	4 411	22 987
Канада	112 617	447 203	29,109	3 869	18 798
Чили	13 096	50 919	13,994	936	3 639
Дания	25 927	151 266	5,207	4 979	29 050
Финляндия	21 581	97 624	5,085	4 244	19 199
Франция	256 316	1 330 998	57,856	4 430	23 005
Греция	9392	98 861	10,413	902	9494
Гонконг	11 758	130 823	6,044	1945	21 645
Венгрия	7 227	41 506	10,162	711	4 084
Ирландия	17 572	52 662	3,536	4 970	14 893
Израиль	11 349	74 121	5,362	2 117	13 823
Италия	145 013	1 016 286	57,177	2 536	17 774
Южная Корея	161 318	380 820	44,501	3 625	8 558
Кувейт	2 797	24 848	1,754	1 595	14 167
Малайзия	18 874	72 505	19,695	958	3 681
Мексика	55 073	420 788	89,564	615	4 698
Нидерланды	48 595	334 286	15,382	3 159	21 732
Норвегия	13 484	122 926	4,314	3 126	28 495
Португалия	17 025	87 352	9,824	1 733	8 892
Сингапур	20 648	71 039	3,268	6 318	21 738
Словакия	2 720	13 746	5,325	511	2 581
Словения	4 520	14 386	1,925	2 348	7 473
Испания	80 104	483 652	39,577	2 024	12 221
Швеция	34 806	198 432	8,751	3 977	22 675
Швейцария	57 503	261 388	7,104	8 094	36 794
Сирия	3 317	44 753	13,840	240	3 234
Турция	31 115	135 961	59,903	519	2 270
Великобритания	244 397	1 024 609	58,005	4213	17 664

5. По выборке для 15 стран оцените параметры модели $X = \theta_1 + \theta_2 h_2 + \theta_3 h_3$, где X — доход на душу населения; h_2 — процент рабочей силы, занятой в сельском хозяйстве; h_3 — средний уровень образования в возрасте после 25 лет (число лет, проведенных в учебных заведениях). Проверьте наличие или отсутствие гетероскедастичности в предположении, что ошибки в наблюдениях некоррелированны и имеют гауссовское распределение, при обнаружении попробовать устранить.

Таблица 12.6

k	X	h_2	h_3	k	X	h_2	h_3	k	X	h_2	h_3
1	7	8	9	6	14	4	16	11	11	6	11
2	9	9	13	7	9	5	11	12	12	4	15
3	9	7	11	8	8	5	11	13	9	8	15
4	8	6	11	9	10	6	12	14	10	5	10
5	8	10	12	10	11	7	14	15	12	8	13

13. Оценивание в мультиколлинеарных моделях

13.1. Мультиколлинеарность

Рассмотрим линейную регрессионную модель

$$Z_n = H_n\theta + E_n, \tag{13.1}$$

где $V_n = \mathbf{cov}\{E_n, E_n\} = \sigma^2 I$. В разделе 10 было показано, что МНК-оценка $\widehat{\theta}_n$ вектора параметров θ имеет вид

$$\widehat{\theta}_n = W_n^{-1} H_n^\top Z_n, \text{ где } W_n = H_n^\top H_n. \tag{13.2}$$

Ошибка $\Delta\widehat{\theta}_n = \widehat{\theta}_n - \theta$ этой оценки имеет нулевое среднее, т.е. $\mathbf{M}\left\{\Delta\widehat{\theta}_n\right\} = 0$, и ковариационную матрицу $\widehat{K}_n = \sigma^2 W_n^{-1}$. Далее в качестве меры точности оценивания вектора θ будем использовать среднюю квадратическую погрешность (с.к.-погрешность) оценки $\widehat{\theta}_n$:

$$\Delta_n = \mathbf{M}\left\{|\Delta\widehat{\theta}_n|^2\right\} = \sigma^2 \operatorname{tr}\left[W_n^{-1}\right], \tag{13.3}$$

где $\operatorname{tr}[\cdot]$ — след матрицы.

Из (13.3) следует, что оценка $\widehat{\theta}_n$ тем точнее, чем меньше величина Δ_n. С.к.-погрешность Δ_n зависит от σ^2 и от $tr\left[W_n^{-1}\right]$. Величина σ^2 определяется точностью наблюдений, а величина $g_n = tr\left[W_n^{-1}\right]$ зависит только от матрицы регрессионных переменных H_n. Если матрица W_n близка к вырожденной, то величина g_n будет большой. Поэтому с.к.-погрешность Δ_n может оказаться неприемлемо большой даже при весьма малой дисперсии ошибок наблюдений σ^2.

Если матрица W_n близка к вырожденной, то говорят, что регрессионная модель (13.1) является *мультиколлинеарной*.

На практике используются различные меры мультиколлинеарности. Наиболее распространенными являются меры:

1) *число обусловленности* $\operatorname{cond}(W_n) = \dfrac{\lambda_{\max}(W_n)}{\lambda_{\min}(W_n)}$, где $\lambda_{\max}(W_n)$ и $\lambda_{\min}(W_n)$ — максимальное и минимальное собственные значения матрицы W_n соответственно;

2) *максимальная парная сопряженность* $r = \max\limits_{i,j}|r_{ij}|$, $i \neq j$, где $r_{ij} = \dfrac{w_{ij}(n)}{\sqrt{w_{ii}(n)w_{jj}(n)}}$, $w_{ij}(n)$ — элемент матрицы W_n, стоящий на пересечении i-й строки и j-го столбца.

Если $\operatorname{cond}(W_n) \gg 1$ или r близка к 1, то модель (13.1) следует признать мультиколлинеарной.

Рассмотрим способы уменьшения погрешности Δ_n при мультиколлинеарности модели (13.1).

13.2. Ридж-оценки

В общем случае *ридж-оценка* вектора θ имеет вид

$$\overline{\theta}_n(k) = (W_n + kT_n)^{-1} H_n^\top Z_n, \tag{13.4}$$

где $k \geqslant 0$ — заданный параметр, а T_n — некоторая положительно определенная матрица размера $(p \times p)$. Обычно T_n выбирают диагональной, причем ее главная диагональ совпадает с главной диагональю матрицы W_n, т.е. $T_n = \operatorname{diag}\left[w_{11}(n), \ldots, w_{pp}(n)\right]$.

Пусть $D_n = T_n^{1/2} = \operatorname{diag}\left[\sqrt{w_{11}(n)}, \ldots, \sqrt{w_{pp}(n)}\right]$. Сделаем замену переменных в модели (13.1):

$$\beta = D_n\theta, \quad G_n = H_n D_n^{-1}.$$

Тогда

$$Z_n = G_n\beta + E_n. \tag{13.5}$$

Модель (13.5) называется *стандартизованной моделью* линейной регрессии. Далее будем предполагать, что (13.1) приведена к виду (13.5).

Нетрудно видеть, что МНК-оценка $\widehat{\beta}_n$ вектора β в модели (13.5) имеет вид

$$\widehat{\beta}_n = \overline{W}_n^{-1} G_n^\top Z_n, \ \text{ где } \ \overline{W}_n = G_n^\top G_n. \tag{13.6}$$

По построению матрица $\overline{W}_n$ имеет единичную диагональ, поэтому ридж-оценка $\overline{\beta}_n(k)$ для β в силу (13.4) принимает следующий вид:

$$\overline{\beta}_n(k) = \left(\overline{W}_n + kI\right)^{-1} G_n^\top Z_n. \tag{13.7}$$

Ридж-оценка $\overline{\beta}_n(k)$ однозначно определяет ридж-оценку $\overline{\theta}_n(k)$ в исходной модели, так как

$$\overline{\theta}_n(k) = D_n^{-1}\overline{\beta}_n(k).$$

Кратко сформулируем основные свойства ридж-оценки:

1) $\overline{\beta}_n(k)$ при $k = 0$ является МНК-оценкой:

$$\overline{\beta}_n(0) = \widehat{\beta}_n;$$

2) $\overline{\beta}_n(k)$ является линейным преобразованием МНК-оценки $\widehat{\beta}_n$:

$$\overline{\beta}_n(k) = B_n(k)\widehat{\beta}_n, \ \text{ где } \ B_n(k) = \left(I + k\overline{W}_n^{-1}\right)^{-1};$$

3) при любом $k > 0$ оценка $\overline{\beta}_n(k)$ является смещенной:

$$\mathbf{M}\{\overline{\beta}_n(k)\} = B_n(k)\mathbf{M}\left\{\widehat{\beta}_n\right\} = B_n(k)\beta \neq \beta,$$

так как $B_n(k) \neq I$ при $k \neq 0$;

4) в классе оценок с фиксированной длиной ридж-оценка минимизирует сумму квадратов отклонений:

$$\overline{\beta}_n(k) = \arg \min_{\beta: |\beta|^2 = c} |Z_n - G_n\beta|^2,$$

где $k = k(c)$, $c > 0$ — заданная величина.

Рассмотрим теперь вопрос о точности ридж-оценки. Пусть $\Delta_n(k) = \mathbf{M}\{|\overline{\beta}_n(k) - \beta|^2\}$ — с.к.-погрешность оценки $\overline{\beta}_n(k)$ вектора параметров β стандартизованной модели (13.5). Используя (13.7), можно показать, что

$$\Delta_n(k) = k^2\beta^\top \left(G_n^\top G_n + kI\right)^{-2}\beta + \sigma^2 \operatorname{tr}\left[G_n\left(G_n^\top G_n + kI\right)^{-2} G_n^\top\right]. \quad (13.8)$$

Из выражения (13.8) следует, что $\Delta_n(k)$ зависит не только от параметра k, но также и от вектора неизвестных параметров β. Поэтому при каждом k величину $\Delta_n(k)$ можно только лишь оценить, заменив в (13.8) β на $\widehat{\beta}_n$ или $\overline{\beta}_n(k)$. Если величина σ^2 также неизвестна, то ее можно заменить на оценку

$$\widehat{\sigma}^2 = \frac{1}{n}|Z_n - G_n\widehat{\beta}_n|^2, \quad (13.9)$$

которая является достаточно точной даже при наличии мультиколлинеарности.

Заметим, что при $k = 0$ из (13.8) следует, что

$$\Delta_n(0) = \sigma^2 \operatorname{tr}\left[\left(G_n^\top G_n\right)^{-1}\right],$$

т.е. совпадает с с.к.-погрешностью МНК-оценки $\widehat{\beta}_n$. Анализируя (13.8), можно показать, что для каждого β существует

$$k^* = \arg \min_{k \geqslant 0} \Delta_n(k). \quad (13.10)$$

Очевидно, что $\overline{\beta}_n(k^*)$ является с.к.-оптимальной ридж-оценкой, так как $\Delta_n(k^*) \leqslant \Delta_n(k)$ для любого $k \geqslant 0$. В частности, $\Delta_n(k^*) < < \Delta_n(0)$, т.е. ридж-оценка $\overline{\beta}_n(k^*)$ точнее МНК-оценки $\widehat{\beta}_n$, причем величина $\Delta_n(k^*)$ может быть намного меньше $\Delta_n(0)$. Именно в этом и состоит смысл перехода от МНК-оценки $\widehat{\beta}_n$ к ридж-оценке $\overline{\beta}_n(k^*)$ для мультиколлинеарной модели.

На практике для приближенного определения k^* можно воспользоваться следующими двумя методами.

1. Итерационное определение k^*:

а) полагаем $k = 0$ и вычисляем $\widehat{\sigma}^2$ по формуле (13.9);

б) вычисляем оценку $\overline{\beta}_n(k)$ по формуле (13.7);

в) находим k, минимизируя $\Delta_n(k)$ из (13.8), где β заменено на $\overline{\beta}_n(k)$, а σ^2 на $\widehat{\sigma}^2$.

Повторяем пункты б) и в) до тех пор, пока итерационный процесс не сойдется (т.е. до выполнения условия $|k_i - k_{i-1}| < \gamma$, где i — номер итерации, а γ — точность определения k^*). Результат последней итерации принимаем за приближенное значение k^*.

2. Метод Макдональда и Галарню.

Вычисляются МНК-оценка $\widehat{\beta}_n$ вектора β и соответствующая оценка $\widehat{\sigma}_n^2$ для σ^2. Далее находится величина

$$C = \widehat{\beta}_n^\top \widehat{\beta}_n - \widehat{\sigma}_n^2 \operatorname{tr}\left[\overline{W}_n^{-1}\right].$$

Если $C \leqslant 0$, то полагаем $k^* = 0$. Если же $C > 0$, то k^* находится из уравнения

$$|\overline{\beta}_n(k)|^2 = C,$$

где $\overline{\beta}_n(k)$ — ридж-оценка вида (13.7). Обоснованием метода является свойство 4) ридж-оценки с учетом того, что C в данном случае является оценкой величины $|\beta|^2$.

Свойства ридж-оценок многократно исследовались методом статистического моделирования на ЭВМ. При наличии мультиколлинеарности в подавляющем числе случаев ридж-оценки оказывались существенно точнее МНК-оценки (даже, если величина k^* определялась лишь приближенно с использованием описанных выше методов).

13.3. Метод редукции

Пусть модель наблюдения имеет вид (13.1), тогда МНК-оценка $\widehat{\theta}_n$ имеет вид (13.2).

Известно, что $\mathbf{M}\left\{\widehat{\theta}_n - \theta\right\} = 0$, т.е. у оценки $\widehat{\theta}_n$ нет смещения, а ковариационная матрица ошибки оценки имеет $K_{\widehat{\theta}_n} = \sigma^2 W_n^{-1}$.

С.к.-погрешность произвольной оценки $\widetilde{\theta}_n$ вектора θ по определению имеет вид

$$\Delta_n = \mathbf{M}\left\{|\widetilde{\theta}_n - \theta|^2\right\} = l_n + d_n, \tag{13.11}$$

где

$$l_n = \left|\mathbf{M}\left\{\widetilde{\theta}_n\right\} - \theta\right|^2, \tag{13.12}$$

$$d_n = \mathbf{M}\left\{\left|\widetilde{\theta}_n - \mathbf{M}\left\{\widetilde{\theta}_n\right\}\right|^2\right\}. \tag{13.13}$$

Величина l_n характеризует систематическую погрешность оценки $\widetilde{\theta}_n$ (т.е. ее смещение), а d_n — случайную погрешность оценки, вызванную наличием ошибок наблюдений E_n.

Для МНК-оценки $\widehat{\theta}_n$ соответствующие параметры точности равны

$$l_n = 0, \quad d_n = \sigma^2 \operatorname{tr}\left[W_n^{-1}\right]. \tag{13.14}$$

Если модель (13.1) мультиколлинеарна, то вместо $\widehat{\theta}_n$ можно построить ридж-оценку $\overline{\theta}_n(k)$:

$$\overline{\theta}_n(k) = \left(W_n + \sigma^2 kI\right)^{-1} H_n^\top Z_n, \quad k \geqslant 0. \tag{13.15}$$

Нетрудно проверить, что для оценки (13.15) параметры $l_n = l_n(k)$ и $d_n = d_n(k)$ имеют следующий вид:

$$l_n(k) = \sigma^4 k^2 \left(\theta^{(0)}\right)^\top \left(W_n + \sigma^2 kI\right)^{-2} \theta^{(0)}, \tag{13.16}$$

где $\theta^{(0)}$ — истинное значение вектора параметров θ, а

$$d_n(k) = \sigma^2 \operatorname{tr}\left[\left(W_n + \sigma^2 kI\right)^{-2} W_n\right]. \tag{13.17}$$

Для определения подходящего значения k проще всего воспользоваться *методом редукции*.

1. Вычисляем $d_n = d_n(0) = \sigma^2 \operatorname{tr}[W_n^{-1}]$ (для мультиколлинеарной модели (13.1) $d_n(0)$ будет очень велико).

2. Задаем некоторое ε: $0 \leqslant \varepsilon \ll d_n(0)$.

3. Находим k_ε, решая относительно k уравнение

$$d_n(k) = \varepsilon, \tag{13.18}$$

где $d_n(k)$ имеет вид (13.17).

4. Вычисляем ридж-оценку $\overline{\theta}_n(k_\varepsilon)$ по формуле (13.15).

5. Находим приближенное значение с.к.-погрешности оценки $\overline{\theta}_n(k_\varepsilon)$:

$$\Delta_n(k_\varepsilon) = l_n(k_\varepsilon) + d_n(k_\varepsilon), \tag{13.19}$$

причем в формуле (13.16) неизвестный вектор $\theta^{(0)}$ заменяем на $\overline{\theta}_n(k_\varepsilon)$.

Варьируя ε, можно добиться существенного снижения погрешности $\Delta_n(k_\varepsilon)$ по сравнению с $\Delta_n(0)$.

В заключение заметим, что уравнение (13.18) является нелинейным алгебраическим уравнением с одним неизвестным k. Для его решения можно использовать численные методы (метод дихотомии, метод золотого сечения и др. подробно описанные в [10]).

13.4. Примеры

Пример 13.1. В табл. 13.1 приведены данные о результатах наблюдения в модели с двумя параметрами θ_1 и θ_2:

$$z_m = h_{1,m}\theta_1 + h_{2,m}\theta_2 + \varepsilon_m, \quad m = 1, \ldots, n. \tag{13.20}$$

Таблица 13.1

m	$h_{1,m}$	$h_{2,m}$	$z_m^{(0)}$	ε_m	$z_m = z_m^{(0)} + \varepsilon_m$
1	0	0,01	0,01	0,561	0,571
2	0,5	−0,24	0,26	−0,220	0,040
3	1	−0,49	0,51	0,646	1,156
4	1,5	0,74	0,76	0,271	1,031
5	2	−0,99	1,01	−0,831	0,179
6	2,5	−1,24	1,26	−1,109	0,151
7	3	−1,49	1,51	−0,328	1,182
8	3,5	−1,74	1,76	0,521	2,281
9	4	−1,99	2,01	−0,450	1,560
10	4,5	−2,24	2,26	−0,557	1,703

В четвертом столбце табл. 13.1 приведены *точные значения измеряемой зависимости*:

$$z_m^{(0)} = h_{1,m}\theta_1^{(0)} + h_{2,m}\theta_2^{(0)}, \quad m = 1, \ldots, n,$$

где $n = 10$ — количество наблюдений; $\theta_1^{(0)}$ и $\theta_2^{(0)}$ — истинные (неизвестные) значения параметров θ_1 и θ_2. Требуется исследовать модель (13.20) на мультиколлинеарность.

Решение. Вычислим матрицу W_n, где $h_m = \left[h_{1,m}, h_{2,m}\right]^\top$, $m = 1, \ldots, n$:

$$W_n = \sum_{m=1}^{n} h_m^\top h_m = \begin{bmatrix} \sum\limits_{m=1}^{n} (h_{1,m})^2 & \sum\limits_{m=1}^{n} h_{1,m}h_{2,m} \\ \sum\limits_{m=1}^{n} h_{2,m}h_{1,m} & \sum\limits_{m=1}^{n} (h_{2,m})^2 \end{bmatrix} = \tag{13.21}$$
$$= \begin{bmatrix} 71{,}25 & -35{,}4 \\ -35{,}4 & 17{,}59 \end{bmatrix}.$$

1. Найдем число обусловленности матрицы W_n:

$$\mathrm{cond}(W_n) = \frac{\lambda_{\max}(W_n)}{\lambda_{\min}(W_n)}. \tag{13.22}$$

Для этого вычислим собственные значения матрицы W_n, решая уравнение

$$\det[W_n - \lambda I] = (71{,}25 - \lambda)(17{,}59 - \lambda) - 35{,}4^2 = 0. \tag{13.23}$$

Корнями квадратного уравнения (13.23) являются числа

$$\lambda_1 = 88{,}538 = \lambda_{\max}(W_n), \quad \lambda_2 = 0{,}00025 = \lambda_{\min}(W_n).$$

Теперь из (13.22) следует

$$\operatorname{cond}(W_n) = \frac{88{,}538}{0{,}00025} = 354\,153{,}2.$$

Видим, что матрица W_n плохо обусловлена, так как $\operatorname{cond}(W_n) \gg 1$.

2. Максимальная парная сопряженность r в данном случае равна

$$r = |r_{12}| = \frac{|w_{12}(n)|}{\sqrt{w_{11}(n) \cdot w_{22}(n)}} = \frac{35{,}4}{\sqrt{71{,}25 \cdot 17{,}59}} \approx 0{,}99998,$$

где $\{w_{ij}(n)\}$ — элементы матрицы W_n. Итак, $\operatorname{cond}(W_n) \gg 1$, а величина r близка к единице, поэтому модель (13.20) с данными из табл. 13.1 следует признать мультиколлиниарной. ■

Пример 13.2. В условиях примера 13.1 с помощью МНК найдите $\theta^{(0)}$ (используя данные $\{z_m^{(0)}\}$) и найдите МНК-оценку $\widehat{\theta}_n$ (используя данные $\{z_m\}$). Проанализировать точность МНК-оценки, если известно, что $\varepsilon_m \sim \mathcal{N}(0; 0{,}25)$, $m = 1, \ldots, n$, а ошибки ε_m и ε_i независимы, если $m \neq i$.

Решение. В примере 13.1 показано, что получаемая модель мультиколлиниарна. Из теоретических соображений следует, что оценка $\widehat{\theta}_n$ должна быть весьма неточной. Убедимся в этом посредством прямых вычислений.

1. Найдем точные значения $\theta_1^{(0)}$, $\theta_2^{(0)}$ параметров модели. Из МНК следует, что

$$\theta^{(0)} = W_n^{-1} H_n^\top Z_n^{(0)}, \tag{13.24}$$

где

$$W_n^{-1} = \begin{bmatrix} 71{,}25 & -35{,}4 \\ -35{,}4 & 17{,}59 \end{bmatrix}^{-1} = \begin{bmatrix} 852{,}776 & 1716{,}364 \\ 1716{,}364 & 3454{,}545 \end{bmatrix},$$

$$H_n^\top Z_n^{(0)} = \begin{bmatrix} \sum\limits_{m=1}^{n} h_{1,m} z_m^{(0)} \\ \sum\limits_{m=1}^{n} h_{2,m} z_m^{(0)} \end{bmatrix} = \begin{bmatrix} 35{,}85 \\ -17{,}812 \end{bmatrix}.$$

Теперь из (13.24) следует, что

$$\theta_1^{(0)} = \theta_2^{(0)} = 1. \tag{13.25}$$

2. Для вычисления $\widehat{\theta}_n$ воспользуемся формулой (13.24), где вместо $Z_n^{(0)}$ используется вектор реальных наблюдений Z_n, приведенный в табл. 13.1. В этом случае

$$H_n^\top Z_n = \begin{bmatrix} \sum\limits_{m=1}^{n} h_{1,m} z_m \\ \sum\limits_{m=1}^{n} h_{2,m} z_m \end{bmatrix} = \begin{bmatrix} 28{,}891 \\ -14{,}347 \end{bmatrix}.$$

Теперь, используя в (13.24) вектор $H_n^\top Z_n$ вместо $H_n^\top Z_n^{(0)}$, получаем

$$\widehat{\theta}_n = W_n^{-1} H_n^\top Z_n = \begin{bmatrix} 12{,}944 \\ 25{,}236 \end{bmatrix}.$$

Итак, $\widehat{\theta}_1 = 12{,}944$, $\widehat{\theta}_2 = 25{,}236$, т.е. оценка $\widehat{\theta}_n$ совершенно не похожа на $\theta^{(0)}$.

Ковариационная матрица $K_{\widehat{\theta}_n} = \sigma W_n^{-1}$ ошибки оценки $\Delta\widehat{\theta}_n = \widehat{\theta}_n - \theta^{(0)}$ с учетом того, что $\sigma^2 = 0{,}25$ по условию, равна

$$K_{\widehat{\theta}_n} = \begin{bmatrix} 213{,}1939 & 429{,}091 \\ 429{,}091 & 863{,}636 \end{bmatrix}. \tag{13.26}$$

Из (13.26) следует, что

$$d_n = \operatorname{tr}\left[K_{\widehat{\theta}_n}\right] = 1076{,}83.$$

При этом $l_n = 0$, так как $\widehat{\theta}_n$ — несмещенная оценка. Отсюда получаем окончательное значение для с.к.-погрешности $\Delta_n = \mathbf{M}\Big\{|\widehat{\theta}_n - \theta^{(0)}|^2\Big\}$:

$$\Delta_n = l_n + d_n = 1076{,}83 \gg 1. \tag{13.27}$$

Кроме того, из (13.26) следует, что с.к.о. σ_1 и σ_2 ошибок оценок параметров θ_1 и θ_2 равны

$$\sigma_1 = \sqrt{213{,}1939} \approx 14{,}6; \quad \sigma_2 = \sqrt{863{,}636} \approx 29{,}39.$$

Таким образом, оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$ «имеют право» быть неточными. Действительно,

$$|\widehat{\theta}_1 - \theta_1^{(0)}| = 11{,}944 < \sigma_1, \quad |\widehat{\theta}_2 - \theta_2^{(0)}| = 24{,}236 < \sigma_2,$$

т.е. реализации оценок $\widehat{\theta}_1 = 12{,}944$ и $\widehat{\theta}_2 = 25{,}236$ не так уж и плохи, если учесть с.к.о. их ошибок σ_1 и σ_2.

Полученные результаты показывают, что в рассматриваемом случае получить приемлемые оценки параметров θ_1 и θ_2 методом наименьших квадратов просто невозможно, что объясняется мультиколлинеарностью модели. ■

Пример 13.3. В условиях примеров 13.1 и 13.2 построить ридж-оценки параметров модели (13.1) методом редукции.

Решение. В примере 13.2 мы построили МНК-оценку $\widehat{\theta}_n$, которая является ридж-оценкой $\overline{\theta}_n(k_\varepsilon)$ для $k_\varepsilon = 0$. Для нахождения значения k_ε при заданном $\varepsilon > 0$ в силу (13.17), где $\sigma^2 = 0{,}25$ по условию, нам следует решить уравнение

$$\operatorname{tr}\left[(W_n + 0{,}25kI)^{-2} W_n\right] = 4\varepsilon. \tag{13.28}$$

Уравнение (13.28) — нелинейное алгебраическое уравнение с одним неизвестным k, которое можно решить, например, методом дихотомии [10]. В табл. 13.2 приведены результаты вычисления k_ε, $\overline{\theta}_n(k_\varepsilon)$ и характеристик точности оценивания $l_n(k_\varepsilon)$, $d_n(k_\varepsilon)$ и $\Delta_n(k_\varepsilon)$ для нескольких значений ε.

Таблица 13.2

ε	k_ε	$\overline{\theta}_n(k_\varepsilon)$		$l_n(k_\varepsilon)$	$d_n(k_\varepsilon)$	$\Delta_n(k_\varepsilon)$
1076,83	0	12,94	25,24	0	1076,83	1076,83
4,17	0,0140	1,11	1,42	1,58	4,17	5,75
3,00	0,0167	0,99	1,18	1,61	3,00	4,61
2,12	0,0201	0,86	0,91	1,64	2,12	3,76
1,62	0,0230	0,81	0,82	1,66	1,62	3,28

Из табл. 13.2 видно, что при $\varepsilon = 3$ коэффициент $k_\varepsilon = 0{,}0167$. Поэтому из (13.15) следует, что

$$\overline{\theta}_n(0{,}0167) = (W_n + 0{,}00418I)^{-1} H_n^\top Z_n = \begin{bmatrix} 0{,}99 \\ 1{,}18 \end{bmatrix}.$$

Заметим, что матрица $\overline{W}_n(k_\varepsilon) = W_n + 0{,}00418I$ чрезвычайно мало отличается от W_n. Тем не менее соответствующая ридж-оценка $\overline{\theta}_n(0{,}0167)$ значимо отличается от $\overline{\theta}_n(0) = \widehat{\theta}_n$ (первая строка табл. 13.2). При этом $\overline{\theta}_n(0{,}0167) - \theta^{(0)} = \begin{bmatrix} -0{,}01 \\ 0{,}18 \end{bmatrix}$, т.е. ошибка ридж-оценки весьма мала по сравнению с $\Delta\widehat{\theta}_n = \widehat{\theta}_n - \theta^{(0)} = \begin{bmatrix} 11{,}94 \\ 24{,}24 \end{bmatrix}$.

Что касается с.к.-погрешности ридж-оценки, то при $k_\varepsilon = 0{,}0167$ мы имеем $\Delta_n(k_\varepsilon) = 4{,}61$, т.е. $\Delta_n(k_\varepsilon)$ уменьшилась более чем в 230 раз по сравнению с с.к.-погрешностью $\Delta_n(0) = 1076{,}83$ МНК-оценки $\widehat{\theta}_n$. Итак, ридж-оценки $\overline{\theta}_n(k)$ существенно точнее $\widehat{\theta}_n$ для $k \in [0{,}014; 0{,}023]$, что вызвано существенной мультиколлинеарностью рассматриваемой модели. Заметим также, что все ридж-оценки — смещенные, так как $l_n(k_\varepsilon) \neq 0$, если $k_\varepsilon > 0$. ■

13.5. Задачи для самостоятельного решения

1. Выведите формулы (13.11) — (13.13) для вычисления с.к.-погрешности произвольной оценки $\widetilde{\theta}_n$ вектора θ.

2. Покажите, что для произвольной ридж-оценки $\overline{\theta}_n(k)$ величины $l_n(k)$ и $d_n(k)$ определяются соотношениями, соответственно, (13.16) и (13.17).

3. Покажите, что при $k = 0$ выполнено $\overline{\theta}_n(k) = \widehat{\theta}_n$, $l_n(k) = 0$, $d_n(k) = \sigma^2 \operatorname{tr}\left[W_n^{-1}\right]$.

4. Рассмотрите предельные значения ридж-оценки $\overline{\theta}_n(k)$ и ее характеристик $l_n(k)$, $d_n(k)$ и $\Delta_n(k)$ при $k \to \infty$. К чему приводит использование в алгоритме ридж-оценивания слишком больших значений параметра регуляризации k?

5. Используя данные табл. 13.1, постройте $\overline{\theta}_n(k)$ и $\Delta_n(k)$ для $k = 0{,}05$, $k = 0{,}5$ и $k = 5$. Выберите из этих оценок лучшую по критерию минимума с.к.-погрешности $\Delta_n(k)$.

6. Пусть $\widetilde{\theta}_n$ равна $\widehat{\theta}_n$ или $\overline{\theta}_n(0{,}0167)$, $\widetilde{z}_m = \widetilde{h}_m\widetilde{\theta}_n$ — соответствующая оценка точного значения измеряемой зависимости $z_m^{(0)} = h_m\theta^{(0)}$, $m = 1, \ldots, n$. Пусть также $\delta_n = \sum\limits_{m=1}^{n} \left(\widetilde{z}_m - z_m^{(0)}\right)^2$. Для какой из оценок величина δ_n будет меньше? Сильно ли изменяется величина δ_n при переходе от МНК-оценки $\overline{\theta}_n(0)$ к ридж-оценке $\overline{\theta}_n(0{,}0167)$? Верно ли утверждение: МНК не позволяет построить «хорошие» оценки параметров $\theta^{(0)}$, но позволяет весьма точно оценить измеряемую зависимость $\{z_m^{(0)},\ m = 1, \ldots, n\}$?

7. Используя данные из табл. 13.3, найдите оценки параметров модели производственной функции $Y = F(K, L)$, если:

1) $\dfrac{Y_m}{L_m} = \theta_1 \left(\dfrac{K_m}{L_m}\right)^{\theta_2}$, $\quad m = 1, \ldots, n$, т.е. без учета технического прогресса;

2) $\dfrac{Y_m}{L_m} = \theta_1 e^{\theta_3 t} \left(\dfrac{K_m}{L_m}\right)^{\theta_2}$, $\quad m = 1, \ldots, n$, т.е. с учетом технического прогресса.

Таблица 13.3

Год	Y	K	L	Год	Y	K	L
1899	100	100	100	1911	153	216	145
1900	101	107	105	1912	177	226	152
1901	112	114	110	1913	184	236	154
1902	122	122	118	1914	169	244	149
1903	124	131	123	1915	189	266	154
1904	122	138	116	1916	225	298	182
1905	143	149	125	1917	227	335	196
1906	152	163	133	1918	223	366	200
1907	151	176	138	1919	218	387	193
1908	126	185	121	1920	231	407	193
1909	155	198	140	1921	179	417	147
1910	159	208	144	1922	240	431	161

Приведенные данные были использованы Ч. Коббом и П. Дугласом при построении производственной функции Кобба—Дугласа [6] для обрабатывающей промышленности США за период 1899—1922 гг. (данные 1899 г. приняты за 100%).

Является ли какая-нибудь из этих моделей мультиколлинеарной? Если да, то примените для оценивания параметров θ метод редукции и проанализируйте полученные результаты.

14. Устойчивые методы регрессионного анализа

Рассмотрим линейную регрессионную модель

$$X_k = h_k^\top \theta + \varepsilon_k, \quad k = 1, \ldots, n. \tag{14.1}$$

Относительно ошибок $\{\varepsilon_k\}$ выше предполагалось, что они независимы и распределены по закону $\mathcal{N}(0; \sigma^2)$. В этих условиях было показано, что наименьшую с.к.-погрешность имеет МНК-оценка

$$\widehat{\theta}_n = W_n^{-1} \sum_{k=1}^{n} h_k X_k, \text{ где } W_n = \sum_{k=1}^{n} h_k h_k^\top. \tag{14.2}$$

С.к.-погрешность оценки $\widehat{\theta}_n$ вычисляется аналитически:

$$\Delta_n = \mathbf{M}\Big\{|\widehat{\theta}_n - \theta|^2\Big\} = \sigma^2 \operatorname{tr}\left[W_n^{-1}\right], \tag{14.3}$$

а сама оценка $\widehat{\theta}_n$ при каждом $n \geqslant 1$ имеет гауссовское распределение: $\widehat{\theta}_n \sim \mathcal{N}\left(\theta; \sigma^2 W_n^{-1}\right)$.

В разд. 13 были рассмотрены методы оценивания, позволяющие уменьшить Δ_n, если модель (14.1) мультиколлинеарна. В данном разделе мы рассмотрим проблему уменьшения Δ_n в случае, если $\sigma^2 = \mathbf{D}\{\varepsilon_k\}$ оказывается слишком большой из-за того, что часть наблюдений в (14.1) являются аномальными.

14.1. Модель аномальных наблюдений

Практический анализ рядов данных показывает, что среди наблюдений весьма часто (с вероятностью до 0,05) встречаются *аномальные наблюдения* (*выбросы*). Появление выбросов обычно связано с возникающими большими ошибками при регистрации данных и передаче их по информационным каналам. Если k-е наблюдение является аномальным, то это может означать, что распределение ошибки ε_k существенно отличается от $\mathcal{N}(0; \sigma^2)$. Для математического моделирования процесса наблюдения с выбросами была предложена следующая схема. Пусть $\delta \geqslant 0$ — вероятность того, что очередное наблюдение будет аномальным, $p(x;\sigma) = \dfrac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\dfrac{x^2}{2\sigma^2}\right\}$ — плотность распределения $\mathcal{N}(0;\sigma^2)$, σ_0^2 — номинальная дисперсия ошибки наблюдения, а σ_1^2 — аномальная дисперсия ошибки наблюдения, причем $0 < \sigma_0 \leqslant \sigma_1$. Тогда плотность вероятности $p(x)$ ошибки наблюдения ε_k можно представить в виде

$$p(x) = (1-\delta)p(x;\sigma_0) + \delta p(x;\sigma_1). \tag{14.4}$$

Нетрудно проверить, что дисперсия σ^2 ошибки ε_k в этом случае будет иметь вид

$$\sigma^2 = (1-\delta)\sigma_0^2 + \delta\sigma_1^2. \tag{14.5}$$

Из (14.5) следует:

1) если аномальные наблюдения отсутствуют (т.е. $\delta = 0$ или $\sigma_0 = \sigma_1$), то $\sigma^2 = \sigma_0^2$;

2) если $\delta > 0$, а $\sigma_1 \gg \sigma_0$, то σ^2 может быть намного больше σ_0^2. Например, если $\delta = 0{,}05$, а $\sigma_1 = 10\sigma_0$, то $\sigma^2 = 0{,}95\sigma_0^2 + 0{,}05\sigma_1^2 = 5{,}95\sigma_0^2$. Таким образом, дисперсия ошибок измерения σ^2 и, следовательно, с.к.-погрешность Δ_n из (14.3) будет почти в 6 раз больше, чем их номинальные значения $\sigma^2 = \sigma_0^2$, и $\Delta_n = \sigma_0^2 \operatorname{tr}\left[W_n^{-1}\right]$, несмотря на то что выбросом является в среднем лишь одно из 20 наблюдений.

Для снижения влияния выбросов на точность оценивания параметров θ в модели (14.1) используются следующие приемы:

1) предварительная отбраковка аномальных наблюдений;

2) модификация метода оценивания для придания ему свойства устойчивости (нечувствительности) к наличию аномальных наблюдений.

Рассмотрим более подробно оба указанных выше подхода.

14.2. Отбраковка аномальных наблюдений

Пусть $\mathbb{Y}_n = = [Y_1, \ldots, Y_n]^\top$ — выборка порожденная СВ Y.

Определение 14.1. *Выборочной медианой* выборки $\mathbb{Y}_n$ называется величина $\operatorname{med}\{\mathbb{Y}_n\} = \operatorname{med}\{Y_1, \ldots, Y_n\}$, определяемая следующим выражением:

$$\operatorname{med}\{\mathbb{Y}_n\} = \begin{cases} Y_{\left(\frac{n+1}{2}\right)}, & \text{если } n \text{ — нечетное число;} \\ \dfrac{1}{2}\left[Y_{\left(\frac{n}{2}\right)} + Y_{\left(\frac{n}{2}+1\right)}\right], & \text{если } n \text{ — четное число.} \end{cases} \tag{14.6}$$

Из определения 14.1 следует, что выборочная медиана совпадает со средним элементом вариационного ряда $\mathbb{Y}_{(n)}$ (n — нечетное) или с полусуммой средних элементов (n — четное). Например, $\operatorname{med}\{\mathbb{Y}_6\} = \frac{Y_{(3)} + Y_{(4)}}{2}$, а $\operatorname{med}\{\mathbb{Y}_7\} = Y_{(4)}$.

Нетрудно видеть, что $\operatorname{med}\{\mathbb{Y}_n\}$ обладает следующим характеристическим свойством: половина из всех наблюдений $[Y_1, \ldots, Y_n]$ не меньше, чем $\operatorname{med}\{\mathbb{Y}_n\}$, а остальные — не больше. Поэтому выборочная медиана является разумной оценкой медианы случайной величины Y (см. определение 21.14).

Замечание. Выборочную медиану для четного числа n не обязательно определять как полусумму средних элементов вариационного ряда. Ее можно определить неоднозначно, как любую случайную величину из отрезка $\left[Y_{\left(\frac{n}{2}\right)}; Y_{\left(\frac{n}{2}+1\right)}\right]$.

Важнейшим свойством оценки $\operatorname{med}\{\mathbb{Y}_n\}$ является ее устойчивость к наличию выбросов среди наблюдений $\{Y_1, \ldots, Y_n\}$ (см. пример 14.1).

Рассмотрим алгоритм отбраковки выбросов в модели (14.1), использующий понятие выборочной медианы.

Алгоритм отбраковки выбросов

1. Найдите МНК-оценку $\widehat{\theta}_n$ по формуле (14.2).

2. Постройте ряд остатков $\{\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n\}$, где остаток k-го наблюдения вычисляется по формуле

$$\widehat{\varepsilon}_k = X_k - h_k^\top \widehat{\theta}_n, \quad k = 1, \ldots, n. \tag{14.7}$$

Заметим сразу, что остаток $\widehat{\varepsilon}_k$ является оценкой ошибки ε_k k-го наблюдения.

3. Найдите выборочную медиану ряда из абсолютных значений найденных остатков:

$$\rho_n = \operatorname{med}\{|\widehat{\varepsilon}_1|, \ldots, |\widehat{\varepsilon}_n|\}. \tag{14.8}$$

4. Вычислите устойчивую оценку номинального значения σ_0:

$$\widehat{\sigma}_0 = \frac{\rho_n}{0{,}675}.$$

5. Найдите аномальные наблюдения среди $\{X_1, \ldots, X_n\}$, пользуясь следующим правилом: если $|\widehat{\varepsilon}_k| > 2\widehat{\sigma}_0$, то X_k — выброс.

Заметим, что в п. 5 алгоритма мы воспользовались известным «правилом двух сигм»: если $X \sim \mathcal{N}(0; \sigma^2)$, то $\mathbf{P}(|X| > 2\sigma) \approx 0{,}05$.

Теперь найденные аномальные наблюдения следует исключить из рассмотрения, а по оставшимся наблюдениям построить новую МНК-оценку $\widehat{\theta}_m^{(1)}$, где $m = n - n_a$, n_a — количество отбракованных аномальных наблюдений.

Для с.к.-погрешности $\Delta_m^{(1)}$ оценки $\widehat{\theta}_m^{(1)}$ можно использовать ее достаточно точную оценку:

$$\Delta_m^{(1)} = \widehat{\sigma}_0^2 \operatorname{tr}\left[W_m^{-1}\right].$$

Если доля отбракованных наблюдений невелика (т.е. $n_a \ll n$), то $\operatorname{tr}\left[W_m^{-1}\right] \approx \operatorname{tr}\left[W_n^{-1}\right]$, поэтому оценка $\widehat{\theta}_m^{(1)}$ будет значительно точнее исходной оценки $\widehat{\theta}_n$, если $\sigma_0^2 \ll \sigma^2 = (1-\delta)\sigma_0^2 + \delta\sigma_1^2$.

В заключение заметим, что оценка $\widehat{\sigma}_0^2$ номинальной дисперсии σ_0^2 устойчива к выбросам, так как в ее основе лежит медианная оценка (14.8).

14.3. Метод наименьших модулей

При наличии аномальных наблюдений МНК-оценка не обладает достаточной точностью из-за ее неустойчивости по отношению к выбросам. Поэтому вместо МНК-оценки мы рассмотрим оценку *метода наименьших модулей* (МНМ-оценку).

МНМ-оценка вектора параметров θ в модели (14.1), обозначаемая далее $\widetilde{\theta}_n$, определяется как решение следующей задачи:

$$\widetilde{\theta}_n = \arg\min_{\theta} \sum_{k=1}^{n} |X_k - h_k^\top \theta|. \tag{14.9}$$

К сожалению, в общем случае $\widetilde{\theta}_n$ не имеет явного аналитического выражения, поэтому для ее практического вычисления разработаны различные итерационные методы [29]. Ниже мы рассмотрим один из них, который называется *методом вариационно-взвешенных наименьших квадратов* (ВВНК-метод).

Пусть m — номер итерации, $\theta(m)$ — приближение к $\widetilde{\theta}_n$, построенное на m-й итерации. Пусть также $\alpha^+ = \begin{cases} \alpha^{-1}, & \text{если } \alpha \neq 0 \\ 0, & \text{если } \alpha = 0. \end{cases}$

Алгоритм ВВНК-метода

1. Положите $m = 0$ и $p_k = 1$, $k = 1, \ldots, n$.

2. Вычислите оценку $\theta(m)$ по формуле метода взвешенных наименьших квадратов:

$$\theta(m) = \left(\sum_{k=1}^{n} p_k h_k h_k^\top\right)^{-1} \sum_{k=1}^{n} p_k h_k X_k. \tag{14.10}$$

3. Вычислите новые веса $\{p_k\}$ по формуле

$$p_k = |X_k - h_k^\top \theta(m)|^+, \quad k = 1, \ldots, n. \tag{14.11}$$

Если $m = 0$, то положите $m = 1$ и перейдите к п. 2.

4. Увеличьте m на единицу и перейдите к п. 2.

В качестве условия окончания итераций можно выбрать

$$\frac{|\theta(m) - \theta(m-1)|}{|\theta(m-1)|} < \gamma,$$

где $\gamma > 0$ выбирается заранее.

Можно доказать, что полученная последовательность приближений $\{\theta(m)\}$ сходится к $\widetilde{\theta}_n$:

1) если $\widetilde{\theta}_n$ — единственная МНМ-оценка (т.е. задача (14.9) имеет единственное решение), то $\theta(m) \to \widetilde{\theta}_n$, $m \to \infty$;

2) если (14.9) имеет множество решений Θ_n, то $\theta(m)$ при $m \to \infty$ сходится к некоторому решению из Θ_n.

Таким образом, если $\lim\limits_{m\to\infty} \theta(m) = \widetilde{\theta}_n$, то

$$\mathcal{L}(\widetilde{\theta}_n) = \sum_{k=1}^{n} |X_k - h_k^\top \widetilde{\theta}_n| \leqslant \sum_{k=1}^{n} |X_k - h_k^\top \theta| = \mathcal{L}(\theta)$$

для любого допустимого θ.

Формулы (14.10) и (14.11) показывают, что для вычисления МНМ-оценки на каждой итерации вычисляется вспомогательная оценка $\theta(m)$. Последнее объясняется тем, что для минимизируемой в (14.9) функции $\mathcal{L}(\theta)$ на каждом шаге алгоритма справедливо:

$$\mathcal{L}(\theta) = \sum_{k=1}^{n} |X_k - h_k^\top \theta| = \sum_{k=1}^{n} |X_k - h_k^\top \theta|^{+} (X_k - h_k^\top \theta)^2 =$$
$$= \sum_{k=1}^{n} p_k (X_k - h_k^\top \theta)^2 = \widetilde{\mathcal{L}}(\theta).$$

Таким образом, минимизация по θ квадратичной функции $\widetilde{\mathcal{L}}(\theta)$ одновременно приводит к минимизации исходной функции $\mathcal{L}(\theta)$, что и требуется в (14.9).

Для иллюстрации различия между МНК-оценкой и МНМ-оценкой рассмотрим наблюдения частного вида (см. пример 14.2):

$$X_k = \theta + \varepsilon_k, \quad k = 1, \ldots, n. \tag{14.12}$$

Применяя к (14.12) МНК, получаем

$$\widehat{\theta}_n = \arg\min_{\theta} \sum_{k=1}^{n} (X_k - \theta)^2 = \frac{1}{n} \sum_{k=1}^{n} X_k = \overline{X}_n.$$

Если же для оценивания θ в (14.12) применить МНМ, то

$$\widehat{\theta}_n = \arg\min_{\theta} \sum_{k=1}^{n} |X_k - \theta| = \mathrm{med}\{X_1, \ldots, X_n\}.$$

Как было указано выше, выборочной медиане следует отдать предпочтение, если среди $\{X_1, \ldots, X_n\}$ есть аномальные наблюдения.

Точное распределение ошибки $\Delta\widetilde{\theta}_n = \widetilde{\theta}_n - \theta$ МНМ-оценки в общем случае найти сложно (подробнее см. гл. 5 в [29]). Однако если $n \gg \gg 1$, модель выбросов имеет вид (14.4) и $\sigma_1 \geqslant \sigma_0$, то в практических расчетах можно полагать, что

$$\widetilde{\theta}_n \sim \mathcal{N}\left(\theta; \kappa W_n^{-1}\right), \tag{14.13}$$

где $\kappa \leqslant \kappa_{\max} = \dfrac{\pi}{2}\sigma_0^2$.

Заметим, что $\kappa_{\max}$ не зависит ни от вероятности δ появления выброса, ни от его дисперсии σ_1^2. Поэтому, $\kappa_{\max} \ll \sigma^2$, если $\delta > 0$ и $\sigma_1 \gg \sigma_0$. Если же выбросы отсутствуют, т.е. $\delta = 0$ и $\sigma_1 = \sigma_0$, то МНМ-оценка лишь немного проигрывает по точности МНК-оценке:

$$\frac{\widetilde{\Delta}_n}{\widehat{\Delta}_n} = \frac{\mathbf{M}\left\{|\widetilde{\theta}_n - \theta|^2\right\}}{\mathbf{M}\left\{|\widehat{\theta}_n - \theta|^2\right\}} \leqslant \frac{\kappa_{\max} \operatorname{tr}\left[W_n^{-1}\right]}{\sigma_0^2 \operatorname{tr}\left[W_n^{-1}\right]} = \frac{\pi}{2}. \tag{14.14}$$

Практическое использование МНМ показывает, что оценки $\widetilde{\theta}_n$ и $\widehat{\theta}_n$ различаются мало, если ошибки наблюдений в модели (14.1) имеют гауссовское распределение $\mathcal{N}(0; \sigma_0^2)$. В этом случае следует отдать предпочтение МНК-оценке $\widehat{\theta}_n$, так как она обладает несколько большей точностью, а вычисляется существенно проще. Если же некоторые наблюдения являются аномальными, то более точной становится оценка МНМ $\widetilde{\theta}_n$, причем ее преимущество над $\widehat{\theta}_n$ тем больше, чем сильнее аномальные ошибки наблюдений отличаются от номинальных.

14.4. Примеры

П р и м е р 14.1. Пусть задана выборка $\mathbb{Y}_n$. Вычислите выборочное среднее и выборочную медиану по реализации $\{1, 4, 2, 5, 3\}$ и в случае, когда к реализации выборки добавлено аномальное наблюдение $y_6 = 75$.

Р е ш е н и е. Построим вариационный ряд выборки $\mathbb{Y}_n$: $\mathbb{Y}_{(n)} = \{1, 2, 3, 4, 5\}$ и, следовательно, $\operatorname{med}\{\mathbb{Y}_5\} = 3$. Выборочное среднее $\overline{Y}_5 = \dfrac{1}{5}(1 + 2 + 3 + 4 + 5) = 3$, т.е. в данном случае $\operatorname{med}\{\mathbb{Y}_n\} = \overline{Y}_n$.

Добавим к ряду наблюдений аномальное наблюдение $y_6 = 75$ и рассмотрим ряд $\mathbb{Y}_{(n+1)} = \{1, 4, 2, 5, 3, 75\}$. Из (14.6) следует, что $\operatorname{med}\{\mathbb{Y}_6\} = \dfrac{y_{(3)} + y_{(4)}}{2} = \dfrac{3 + 4}{2} = 3{,}5$. В то же время $\overline{Y}_6 = \dfrac{1}{6}(1 + 2 + 3 + 4 + 5 + 75) = \dfrac{90}{6} = 15$. Таким образом, величина

$\mathrm{med}\{\mathbb{Y}_6\}$ изменилась мало по сравнению с $\mathrm{med}\{\mathbb{Y}_5\}$, в то время как $\overline{Y}_6$ стала в 5 раз больше, чем $\overline{Y}_5$.

Рассмотренный пример показывает, что при оценивании среднего выборочной медиане следует отдать предпочтение по сравнению с выборочным средним в случае, когда наблюдения содержат выбросы. ■

Пример 14.2. Пусть $\{X_1, \ldots, X_n\}$ — ряд наблюдений, $\widehat{\theta}_n = \overline{X}_n$ — выборочное среднее, а $\widetilde{\theta}_n = \mathrm{med}\{X_1, \ldots, X_n\}$ — выборочная медиана. Покажите, что $\widehat{\theta}_n$ и $\widetilde{\theta}_n$ являются решениями следующих задач оптимизации:

$$\widehat{\theta}_n = \arg\min_{\theta} \sum_{k=1}^{n} (X_k - \theta)^2, \tag{14.15}$$

$$\widetilde{\theta}_n = \arg\min_{\theta} \sum_{k=1}^{n} |X_k - \theta|. \tag{14.16}$$

Решение. 1. Задача (14.15) — частный случай МНК, где модель наблюдения имеет вид

$$X_k = h_k\theta + \varepsilon_k, \quad , k = 1, \ldots, n, \tag{14.17}$$

причем $h_k = 1$, $k = 1, \ldots, n$. Отсюда

$$\widehat{\theta}_n = \frac{\sum\limits_{k=1}^{n} h_k X_k}{\sum\limits_{k=1}^{n} h_k^2} = \frac{1}{n} \sum_{k=1}^{n} X_k = \overline{X}_n. \tag{14.18}$$

2. Покажем теперь, что выборочная медиана $\widetilde{\theta}_n$ действительно является точкой минимума функции потерь

$$\varphi_n(\theta) = \sum_{k=1}^{n} |X_k - \theta|.$$

Для определенности положим, что $n = 2m+1$, $m \geqslant 1$. Тогда известно, что

$$\widetilde{\theta}_n = X_{(m)}, \tag{14.19}$$

где $X_{(m)}$ — m-й (средний) элемент вариационного ряда $\{X_{(1)}, \ldots, X_{(n)}\}$, построенного по выборке $Z_n = \{X_1, \ldots, X_n\}$.

По определению $\varphi_n(\widetilde{\theta}_n) = \sum\limits_{k=1}^{n} |X_{(k)} - X_{(m)}| = \sum\limits_{k=1}^{m-1} |X_{(k)} - X_{(m)}| +$

$$+ \sum_{m+1}^{n} |X_{(k)} - X_{(m)}|.$$

Пусть теперь $\theta \in \left(X_{(m)}, X_{(m+1)}\right]$. Покажем, что $\varphi_n(\theta) \geqslant \varphi_n(\widetilde{\theta}_n)$. Действительно,

$$\begin{aligned}
\varphi_n(\theta) &= \sum_{k=1}^{m-1} |X_{(k)} - \theta| + |X_{(m)} - \theta| + \sum_{k=m+1}^{n} |X_k - \theta| = \\
&= \left\{ \sum_{k=1}^{m-1} |X_{(k)} - \widetilde{\theta}_n| + (m-1)|X_{(m)} - \theta| \right\} + |X_{(m)} - \theta| + \\
&+ \left\{ \sum_{k=m+1}^{n} |X_k - \theta| - (m-1)|X_{(m)} - \theta| \right\} = \\
&= \varphi_n(\widetilde{\theta}_n) + |X_{(m)} - \theta| \geqslant \varphi_n(\widetilde{\theta}_n).
\end{aligned}$$

Аналогичным образом можно показать, что $\varphi_n(\theta) \geqslant \varphi_n(\widetilde{\theta}_n)$ при произвольно выбранном θ.

В случае когда число наблюдений четно, т.е. $n = 2m$, $m \geqslant 1$, минимум функции $\varphi_n(\theta)$ достигается в каждой точке отрезка $\left[X_{(m)}, X_{(m+1)}\right]$, т.е.

$$\varphi_n(\theta_1) = \varphi_n(\theta_2) \ \text{ для любых } \ \theta_1, \theta_2 \in \left[X_{(m)}, X_{(m+1)}\right].$$

Последнее означает, что выборочной медианой является любая точка указанного отрезка, т.е. в выборе $\widetilde{\theta}_n$ имеется некоторый произвол. Обычно в данной ситуации полагают $\widetilde{\theta}_n = \frac{1}{2}\left(X_{(m)} + X_{(m+1)}\right)$, хотя возможно также положить $\widetilde{\theta}_n = X_{(m)}$ или $\widetilde{\theta}_n = X_{(m+1)}$.

Проведенные рассмотрения позволяют сделать следующий вывод: для оценивания среднего значения θ ряда данных $\{X_1, \ldots, X_n\}$ можно применить не только метод наименьших квадратов (МНК) (14.15), но и метод наименьших модулей (МНМ) (14.16). В первом случае оценкой среднего будет $\widehat{\theta}_n = \overline{X}_n$ — выборочное среднее, а во втором — $\widetilde{\theta}_n = \mathrm{med}\{X_1, \ldots, X_n\}$ — выборочная медиана. ■

Пример 14.3. В табл. 14.1 приведены данные Госкомстата РФ (сентябрь 2005 г.) о среднедушевых доходах и средней начисленной зарплате в 18 регионах РФ, образующих Центральный федеральный округ.

Таблица 14.1

Регион (область)	Средний душевой доход, руб.	Средняя начисленная зарплата, руб.
Белгородская	5507	6473
Брянская	4569	4989
Владимирская	3886	5917
Воронежская	4958	5283
Ивановская	3123	5086
Калужская	5089	6637
Костромская	4367	5575
Курская	5283	5205
Липецкая	5330	6709
г. Москва	23 274	12 969
Московская	7234	9047
Орловская	4780	5175
Рязанская	4450	5926
Смоленская	5437	5971
Тамбовская	5327	4805
Тверская	5500	6280
Тульская	5061	6083
Ярославская	6438	7013

Требуется построить оценки среднедушевого дохода в ЦФО методом наименьших квадратов и методом наименьших модулей, определить точность этих оценок и проанализировать данные на наличие выбросов.

Решение. Пусть θ — искомый среднедушевой доход в ЦФО РФ. МНК-оценкой для θ является выборочное среднее $\overline{X}_n$, где $n = 18$. Используя данные, приведенные во втором столбце табл. 14.1, находим:

$$\widehat{\theta}_n = \overline{X}_n = \frac{1}{18}\sum_{k=1}^{18} X_k = 6089{,}61. \tag{14.20}$$

Оценкой для среднего квадратического отклонения (с.к.о.) σ является величина $\widehat{\sigma}_n$:

$$\widehat{\sigma}_n = \sqrt{\frac{1}{18}\sum_{k=1}^{18}(X_k - \widehat{\theta}_n)^2} = 4257{,}91. \tag{14.21}$$

Так как $\mathbf{D}\left\{\widehat{\theta}_n\right\} = \dfrac{\sigma^2}{n}$, с.к.о. ошибки оценки $\widehat{\theta}_n$ равно

$$\widehat{\sigma}_{\theta_n} = \sqrt{\mathbf{D}\left\{\widehat{\theta}_n\right\}} = \frac{\widehat{\sigma}_n}{\sqrt{18}} \approx 1003{,}60. \tag{14.22}$$

Видно, что величина (14.22) чрезвычайно велика, что указывает на то, что оценка $\widehat{\theta}_n$ может быть весьма неточной. Например, полагая, что $\widehat{\theta}_n \sim N\left(\theta; \widehat{\sigma}^2_{\theta_n}\right)$, из (14.20) и (14.22) следует, что истинное значение θ с надежностью близкой к 0,95 лежит в интервале $\left[\widehat{\theta}_n - 1{,}96\widehat{\sigma}_{\theta_n}; \widehat{\theta}_n + 1{,}96\widehat{\sigma}_{\theta_n}\right] = [4122{,}56; 8056{,}06] = I_n^{(1)}$. Видно, что интервал $I_n^{(1)}$ накрывает практически все результаты наблюдений. Последнее указывает на неинформативность полученных результатов оценивания.

Рассмотрим теперь результаты оценивания по МНМ. Так как $n = 18$ — четное число, то МНМ-оценка для θ, которая является выборочной медианой $\widetilde{\theta}_n$, будет равна

$$\widetilde{\theta}_n = \frac{X_{(9)} + X_{(10)}}{2}. \tag{14.23}$$

Для вычисления $\widetilde{\theta}_n$ построим вариационный ряд $\left[X_{(1)}, \ldots, X_{(9)}, X_{(10)}, \ldots, X_{(18)}\right]^\top = [3123, \ldots, 5089, 5283, \ldots, 23\,274]^\top$.

Выборочной медианой $\widetilde{\theta}_n$ является любая точка промежутка $\left[X_{(9)}, X_{(10)}\right] = [5089; 5283]$. В качестве окончательной оценки $\widetilde{\theta}_n$ выберем (14.23), т.е.

$$\widetilde{\theta}_n = \frac{5089 + 5283}{2} = 5186. \tag{14.24}$$

Сравнивая (14.24) и (14.20), видим, что оценки $\widehat{\theta}_n$ и $\widetilde{\theta}_n$ различаются очень существенно:

$$\delta\% = \frac{|\widehat{\theta}_n - \widetilde{\theta}_n|}{\widetilde{\theta}_n} \cdot 100\% = 17{,}42\%. \tag{14.25}$$

Теперь построим устойчивую оценку параметра разброса σ. Для этого вычислим абсолютные значения отклонений наблюдений X_k от $\widetilde{\theta}_n$, $k = 1, \ldots, n$:

$$v_k = |X_k - \widetilde{\theta}_n|, \quad k = 1, \ldots, n.$$

После этого найдем выборочную медиану выборки $\{v_k\}$:

$$\overline{\sigma}_n = \operatorname{med}\{v_1, \ldots, v_n\}.$$

Приведем реализацию вариационного ряда $V_{(n)} = \left[v_{(1)}, \ldots, v_{(n)}\right]^\top$:

$$V_{(n)} = [97; 97; 125; 141; 144; 228; 251; 314; 321;$$
$$406; 617; 736; 819; 1252; 1300; 2048; 2063; 18\,088]^\top.$$

Отсюда видно, что $v_{(9)} = 321$, $v_{(10)} = 406$, т.е.

$$\overline{\sigma}_n = \frac{321 + 406}{2} = 363{,}5. \qquad (14.26)$$

Теперь искомая оценка $\widetilde{\sigma}_n$ для σ получается посредством «подправления» оценки (14.26):

$$\widetilde{\sigma}_n = \frac{\overline{\sigma}_n}{0{,}675} = 538{,}52. \qquad (14.27)$$

Из (14.21) и (14.27) следует, что $\widetilde{\sigma}_n$ практически в 8 раз меньше, чем $\widehat{\sigma}_n$. Последнее явно указывает на присутствие аномальных наблюдений.

Дисперсию ошибки оценки $\widetilde{\theta}_n$ можно приближенно вычислить по формуле

$$\mathbf{D}\left\{\widetilde{\theta}_n\right\} \approx \frac{\pi}{2} \cdot \frac{\sigma}{n}.$$

Отсюда, заменяя σ на $\widetilde{\sigma}_n$, получаем

$$\widetilde{\sigma}_{\theta_n} = \sqrt{\mathbf{D}\left\{\widetilde{\theta}_n\right\}} \approx \sqrt{\frac{\pi}{2n}} \cdot \widetilde{\sigma}_n = 159{,}08. \qquad (14.28)$$

Теперь с надежностью близкой к 0,95 можно утверждать, что θ лежит в промежутке

$$I_n^{(2)} = \left[\widetilde{\theta}_n - 1{,}96\widetilde{\sigma}_{\theta_n}; \widetilde{\theta}_n + 1{,}96\widetilde{\sigma}_{\theta_n}\right] = [4874; 5497{,}8].$$

Длина интервала $I_n^{(2)}$ намного меньше длины $I_n^{(1)}$, что указывает на значимое преимущество по точности оценки $\widetilde{\theta}_n$ над $\widehat{\theta}_n$. ■

П р и м е р 14.4. Используя результаты примера 14.3, проведите отбраковку аномальных наблюдений и постройте оценки $\widetilde{\theta}_n$ и $\widehat{\theta}_n$ по урезанной выборке.

Р е ш е н и е. Отбраковку проведем с использованием «правила трех сигм», а именно: наблюдение X_k отбраковывается, если

$$v_k = |X_k - \widetilde{\theta}_n| > 3\widetilde{\sigma}_n, \quad k = 1, \ldots, n. \qquad (14.29)$$

Сравнивая реализацию вариационного ряда $V_{(n)}$ с величиной $l = 3\widetilde{\sigma}_n = 1615{,}56$, находим, что отбраковке подлежат наблюдения $X_5 = 3123$, $X_{10} = 23\,274$ и $X_{11} = 7234$. Исключая указанные наблюдения из исходной выборки объема $n = 18$, получаем новую выборку объема $n = 15$.

Теперь, повторяя все вычисления, описанные в примере 14.3, находим новые реализации оценок:

$$\widehat{\theta}_n = 5065{,}47; \quad \widehat{\sigma}_n = 588{,}18;$$

$$\widetilde{\theta}_n = 5089{,}0; \quad \widetilde{\sigma}_n = 515{,}56.$$

Видно, что реализации оценок $\widehat{\theta}_n$, $\widetilde{\theta}_n$ и $\widehat{\sigma}_n$, $\widetilde{\sigma}_n$ теперь различаются несущественно. Это указывает на однородность выборки после отбраковки и отсутствие аномальных наблюдений.

Заметим также, что

$$\max\{|X_k - \widetilde{\theta}_n|,\ k = 1, \ldots, n\} = 1349{,}0 < 3\widetilde{\sigma}_n = 1546{,}67.$$

Это означает, что выборка, полученная после первой отбраковки, более не содержит аномальных наблюдений. ■

Пример 14.5. В условиях примера 14.3 вычислите МНМ-оценку $\widetilde{\theta}_n$ с помощью алгоритма вариационно-взвешенных наименьших квадратов (ВВНК-метод).

Решение. Пусть $\{p_k,\ k = 1, \ldots, n\}$ — некоторые положительные веса. Рассмотрим задачу построения взвешенной МНК-оценки для θ:

$$\widehat{\theta}_n = \arg\min_{\theta} \sum_{k=1}^{n} p_k (X_k - \theta)^2. \tag{14.30}$$

Из (14.30) следует, что

$$\widehat{\theta}_n = \frac{\sum\limits_{k=1}^{n} p_k X_k}{\sum\limits_{k=1}^{n} p_k}. \tag{14.31}$$

Пусть $m \geqslant 1$ — номер итерации ВВНК-метода, а $\theta(m)$ — соответствующее приближение к $\widetilde{\theta}_n$. Вычислим веса $\{p_k^{(m)}\}$, которые будут использованы для построения $\theta(m)$, по следующей формуле:

$$p_k^{(m)} = \begin{cases} |X_k - \theta(m-1)|^{-1}, & \text{если } |X_k - \theta(m-1)| > 0; \\ 0, & \text{если } |X_k - \theta(m-1)| = 0. \end{cases} \tag{14.32}$$

Тогда $\theta(m)$ вычисляется по формуле (14.31), где веса p_k заменяются на $p_k^{(m)}$ из (14.32):

$$\theta(m) = \frac{\sum\limits_{k=1}^{n} p_k^{(m)} X_k}{\sum\limits_{k=1}^{n} p_k^{(m)}}. \tag{14.33}$$

Задав некоторое начальное приближение $\theta(0)$, по рекуррентным формулам (14.32), (14.33) получаем последовательность приближений $\{\theta(m),\ m = 1, 2, \ldots\}$, сходящуюся к искомой МНМ-оценке $\widetilde{\theta}_n$.

В табл.14.2 приведены результаты итераций по формулам (14.32) и (14.33) для двух начальных значений: $\theta(0) = 3200$ и $\theta(0) = 6089$.

Таблица 14.2

m	$\theta^{(1)}(m)$	$\theta^{(2)}(m)$	m	$\theta^{(1)}(m)$	$\theta^{(2)}(m)$
0	3200	6089	6	5018	5305
1	3913	5453	7	5066	5293
2	4180	5395	8	5073	5282
3	4700	5333	9	5084	5282
4	4867	5321	10	5091	5282
5	4978	5315	11	5091	5282

Из таблицы видно, что для случая $\theta^{(1)}(0) = 3200$ процесс сошелся за десять итераций, а для $\theta^{(2)}(0) = 6089$ — за восемь. Видно, что обе предельные точки 5091 и 5282 принадлежат промежутку $[X_{(9)}, X_{(10)}] = [5089; 5283]$, т.е. любое решение является искомой оценкой $\widetilde{\theta}_n$. В качестве $\widetilde{\theta}_n$ в данном случае можно положить также $\widetilde{\theta}_n = \dfrac{5091 + 5282}{2} = 5186{,}5$, что полностью согласуется с оценкой (14.24). ■

П р и м е р 14.6. Связь между данными $\{x_k\}$ и $\{y_k\}$, приведенными в табл. 14.3, описывается линейной регрессионной моделью

$$y_k = \theta_1 + \theta_2 x_k + \varepsilon_k, \quad k = 1, \ldots, n, \tag{14.34}$$

где $\{\theta_1, \theta_2\}^\top$ — оцениваемые параметры, $n = 10$, $\varepsilon_k \sim N(0; 0{,}25)$ — независимые ошибки наблюдений.

Таблица 14.3

k	y_k	x_k	k	y_k	x_k
1	0,850	1,0	6	6,867	3,5
2	1,361	1,5	7	5,908	4,0
3	3,122	2,0	8	7,883	4,5
4	4,638	2,5	9	9,548	5,0
5	5,599	3,0	10	9,457	5,5

1. По данным табл. 14.3 постройте МНК-оценку $\widehat{\theta}_n$ и МНМ-оценку $\widetilde{\theta}_n$ вектора $\theta = \{\theta_1, \theta_2\}^\top$, найдите их с.к.-погрешности и сравните оценки с точными значениями параметров $\theta^{(0)} = \{-1; 2\}^\top$.

2. Проделайте те же вычисления для случая, когда по техническим причинам наблюдение y_{10} «пропало»: $y_{10} = 0$. Сделайте выводы об устойчивости оценок $\widehat{\theta}_n$ и $\widetilde{\theta}_n$.

Р е ш е н и е.

1. МНК-оценка $\widehat{\theta}_n$ является решением системы уравнений

$$\begin{cases} n\theta_1 + \sum\limits_{k=1}^{n} x_k \theta_2 = \sum\limits_{k=1}^{n} y_k, \\ \sum\limits_{k=1}^{n} x_k \theta_1 + \sum\limits_{k=1}^{n} (x_k)^2 \theta_2 = \sum\limits_{k=1}^{n} x_k y_k. \end{cases} \tag{14.35}$$

Используя данные табл. 14.3, находим МНК-оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$ параметров θ_1 и θ_2 из (14.35):

$$\widehat{\theta}_1 = -0{,}92; \quad \widehat{\theta}_2 = 1{,}98. \tag{14.36}$$

Найдем с.к.о. ошибок оценок $\widehat{\theta}_1$ и $\widehat{\theta}_2$. Известно (см. теорему 10.2), что ковариационная матрица ошибок МНК-оценок имеет вид

$$K_{\widehat{\theta}} = \sigma_\varepsilon^2 W_n^{-1}, \tag{14.37}$$

где $\sigma_\varepsilon^2 = 0{,}25$, а $W_n = \begin{bmatrix} 10 & 32{,}5 \\ 32{,}5 & 12625 \end{bmatrix}$. Отсюда следует, что

$$K_{\widehat{\theta}} = 0{,}25 \begin{bmatrix} 0{,}612 & -0{,}158 \\ -0{,}158 & 0{,}048 \end{bmatrix}. \tag{14.38}$$

С.к.о. ошибок оценок, таким образом, равны

$$\begin{aligned} \sigma_1 &= \sqrt{\mathbf{D}\left\{\widehat{\theta}_1\right\}} = 0{,}5 \cdot \sqrt{0{,}612} = 0{,}39; \\ \sigma_2 &= \sqrt{\mathbf{D}\left\{\widehat{\theta}_2\right\}} = 0{,}5 \cdot \sqrt{0{,}048} = 0{,}11. \end{aligned} \tag{14.39}$$

Из условия примера известны точные значения параметров θ_1 и θ_2: $\theta_1^{(0)} = -1$, $\theta_2^{(0)} = 2$. Видим, что из (14.36) и (14.39) следует:

$$|\widehat{\theta}_1 - \theta_1^{(0)}| = 0{,}08 < \sigma_1; \quad |\widehat{\theta}_2 - \theta_2^{(0)}| = 0{,}02 < \sigma_2.$$

Таким образом, мы получили достаточно точные оценки параметров модели (14.34) с помощью МНК. Это и не удивительно, так как регрессионная модель (14.34) — гауссовская, и, следовательно, МНК-оценки являются эффективными (с.к.-оптимальными на классе всех несмещенных оценок).

Найдем теперь МНМ-оценки $\widetilde{\theta}_n$ методом вариационно-взвешенных наименьших квадратов. В данном случае на m-й итерации решается система

$$\begin{cases} \sum\limits_{k=1}^{n} p_k\theta_1 + \sum\limits_{k=1}^{n} p_k x_k \theta_2 = \sum\limits_{k=1}^{n} p_k y_k, \\ \sum\limits_{k=1}^{n} p_k x_k \theta_1 + \sum\limits_{k=1}^{n} p_k (x_k)^2 \theta_2 = \sum\limits_{k=1}^{n} p_k x_k y_k, \end{cases} \tag{14.40}$$

где $p_k = |y_k - \theta_1(m-1) - \theta_2(m-1)x_k|^{-1}$, $\theta_1(m)$, $\theta_2(m)$ — приближения к $\widetilde{\theta}_1$ и $\widetilde{\theta}_2$ на m-й итерации, являющиеся решением системы (14.40).

В качестве начального приближения естественно взять МНК-оценки, т.е.

$$\widetilde{\theta}_1(0) = \widehat{\theta}_1 = -0{,}92; \quad \widetilde{\theta}_2(0) = \widehat{\theta}_2 = 1{,}98. \tag{14.41}$$

Проведя описанные соотношениями (14.40) и (14.41) итерации, находим МНМ-оценки:

$$\widetilde{\theta}_1 = -0{,}69; \quad \widetilde{\theta}_2 = 1{,}91. \tag{14.42}$$

Видим, что оценки (14.42) несколько хуже по точности МНК-оценок (14.36).

С.к.о. $\widetilde{\sigma}_1$ и $\widetilde{\sigma}_2$ ошибок оценок $\widetilde{\theta}_1$, $\widetilde{\theta}_2$ вычисляются с помощью (14.13). Действительно, $\widetilde{\sigma}_m = \sqrt{\dfrac{\pi}{2}}\sigma_m$, $m = 1, 2$. Отсюда получаем, что

$$\widetilde{\sigma}_1 = 0{,}49; \quad \widetilde{\sigma}_2 = 0{,}14.$$

По-прежнему, реальные ошибки оценок согласуются с характеристиками их точности:

$$|\widetilde{\theta}_1 - \theta_1^{(0)}| = 0{,}31 < \widetilde{\sigma}_1; \quad |\widetilde{\theta}_2 - \theta_2^{(0)}| = 0{,}09 < \widetilde{\sigma}_2. \tag{14.43}$$

2. Для исследования устойчивости оценок заменим в табл. 14.3 «пропавшее» измерение $y_{10} = 9{,}457$ на $y_{10} = 0$ и повторим все вычисления, подробно описанные выше. Полученные результаты для МНК-алгоритма выглядят так:

$$\begin{gathered} \widehat{\theta}_1 = 1{,}48; \quad \widehat{\theta}_2 = 0{,}95; \\ \sigma_1 = 0{,}39, \quad \sigma_2 = 0{,}11. \end{gathered} \tag{14.44}$$

Видно, что оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$ совершенно не похожи на $\theta_1^{(0)}$ и $\theta_2^{(0)}$. Более того,

$$|\widehat{\theta}_1 - \theta_1^0| = 2{,}48 > 6\sigma_1; \quad |\widehat{\theta}_2 - \theta_2^0| = 1{,}05 > 9\sigma_2. \tag{14.45}$$

Из (14.45) следует, что фактические ошибки оценок $\widehat{\theta}_1$ и $\widehat{\theta}_2$ не согласуются с их теоретическими характеристиками.

Теперь вычислим $\widetilde{\theta}_1$ и $\widetilde{\theta}_2$ итерационным методом, выбрав за начальные значения $\theta_1(0) = 1{,}48$; $\theta_2(0) = 0{,}95$. Оказалось, что оценки ВВНК-метода практически не изменились:

$$\begin{gathered} \widetilde{\theta}_1 = -0{,}68; \quad \widetilde{\theta}_2 = 1{,}91; \\ \widetilde{\sigma}_1 = 0{,}49; \quad \widetilde{\sigma}_2 = 0{,}14. \end{gathered}$$

Отсюда следует, что соотношения (14.43) остаются в силе, т.е. ошибки МНМ-оценок по-прежнему согласуются со своими теоретическими характеристиками.

Сформулируем некоторые выводы:

1) при отсутствии аномальных наблюдений МНК-оценки параметров модели (14.34) несколько превосходят по точности МНМ-оценки, хотя их различие весьма незначительно;

2) при наличии аномального измерения МНК-оценки стали чрезвычайно сильно отличаться от истинных значений параметров, в то время как МНМ-оценки на появление выброса практически никак не отреагировали.

Проведенные численные эксперименты подтверждают вывод о чувствительности МНК-оценок к присутствию аномальных наблюдений и устойчивости МНМ-оценок. ■

14.5. Задачи для самостоятельного решения

1. Используя данные, приведенные в табл. 14.1, оцените среднюю начисленную зарплату в ЦФО с помощью методов наименьших квадратов и наименьших модулей. Проанализируйте данные на наличие аномальных значений, проведите их отбраковку.

2. Рассмотрите линейную регрессионную модель

$$y_k = \theta_1 + \theta_2 x_k + \varepsilon_k, \quad k = 1, \ldots, n,$$

где y_k — среднедушевой доход в k-м регионе ЦФО, а x_k — средняя начисленная зарплата в этом же регионе (табл. 14.2). Найдите оценки параметров θ_1 и θ_2 с помощью алгоритмов МНК и МНМ.

3. В условиях примера 14.3 проанализируйте данные на наличие выбросов, используя алгоритм отбраковки аномальных данных.

4. Используя данные примера 14.6, постройте графики оценок зависимости $y(x) = \theta_1 + \theta_2 x$ для различных видов оценок параметров. Постройте прогнозные значения для $y(x)$ в точках $x = 6; 6{,}5; 7$ и сравните их с точными значениями $y(x) = -1 + 2x$ в указанных точках. Проанализируйте влияние выбросов на точность полученных прогнозов.

5. В табл.14.4 приведены индексы объема производства Y_k, капитальных затрат K_k и затрат труда L_k для $n = 10$ лет (США, 1913–1922 гг.).

Таблица 14.4

k	Y_k	K_k	L_k	k	Y_k	K_k	L_k
1	184	236	154	6	223	366	200
2	169	244	149	7	218	387	193
3	189	266	154	8	231	407	193
4	225	298	182	9	179	417	147
5	227	335	196	10	240	431	161

Предполагается, что связь между указанными переменными имеет вид

$$\ln\left(\frac{Y_k}{L_k}\right) = \theta_1 + \theta_2 \ln\left(\frac{K_k}{L_k}\right) + \varepsilon_k, \quad k = 1, \ldots, 10.$$

Найдите оценки параметров θ_1 и θ_2 с помощью МНК и МНМ. Найдите устойчивую оценку параметра $\sigma^2 = \mathbf{D}\{\varepsilon_k\}$. Проанализируйте данные на наличие выбросов и проведите их отбраковку.

15. Нелинейные регрессионные модели

В этом разделе мы рассмотрим нелинейные регрессионные модели, в которых наблюдаемые переменные, параметры и ошибки наблюдений связаны некоторой нелинейной зависимостью:

$$X_k = \varphi(h_k, \theta, \varepsilon_k), \quad k = 1, \ldots, n. \tag{15.1}$$

Для оценивания вектора параметров θ по наблюдениям $Z_n = \{X_1, \ldots, X_n\}^\top$ используются различные точные и приближенные методы, причем выбор конкретного метода в значительной степени определяется видом функции $\varphi(\cdot)$ и имеющейся информацией о параметрах θ и ошибках $\{\varepsilon_k, k = 1, \ldots, n\}$.

15.1. Нелинейные модели, приводящиеся к линейным

Предположим, что найдется такое преобразование $\psi(\cdot)$, что

$$\psi\left(\varphi(h_k, \theta, \varepsilon_k)\right) = f_1^\top(h_k)\theta + f_2(h_k)\varepsilon_k. \tag{15.2}$$

Тогда, применяя преобразование $\psi(\cdot)$ к обеим частям уравнения (15.1), получаем

$$\psi(X_k) = f_1^\top(h_k)\theta + f_2(h_k)\varepsilon_k, \quad k = 1, \ldots, n. \tag{15.3}$$

Обозначая $\widetilde{X}_k = \psi(X_k)$, $f_k = f_1(h_k)$, $\sigma_k = f_2(h_k)$, приводим модель (15.3) к виду

$$\widetilde{X}_k = f_k^\top\theta + \sigma_k\varepsilon_k, \quad k = 1, \ldots, n. \tag{15.4}$$

Модель (15.4) является линейной регрессионной моделью с гетероскедастичными ошибками наблюдений. Применяя теперь к (15.4) метод взвешенных наименьших квадратов, находим оценку $\widehat{\theta}_n$ вектора θ по формуле (12.3) из раздела 12.1:

$$\widehat{\theta}_n = \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} f_k f_k^\top\right)^{-1} \sum_{k=1}^{n} \frac{1}{\sigma_k^2} f_k \widetilde{X}_k. \tag{15.5}$$

Естественно, предполагается, что $\sigma_k^2 \neq 0$, $k = 1, \ldots, n$, а обратная матрица в (15.5) существует.

В качестве примера рассмотрим задачу оценивания параметров *производственной функции Кобба—Дугласа* [6]. Модель наблюдения в данном случае имеет вид

$$X_k = e^{\theta_1} K_k^{\theta_2} L_k^{\theta_3} \varepsilon_k, \quad k = 1, \ldots, n, \tag{15.6}$$

причем предполагается также, что ε_k имеет *логнормальное распределение*: $\ln \varepsilon_k \sim \mathcal{N}(0; \sigma^2)$. В этом случае для приведения модели (15.6) к линейному виду можно воспользоваться преобразованием $\psi(x) = \ln x$:

$$\ln X_k = \theta_1 + \theta_2 \ln K_k + \theta_3 \ln L_k + \ln \varepsilon_k. \tag{15.7}$$

Если обозначить $\widetilde{X}_k = \ln X_k$, $f_k^\top = \{1, \ln K_k, \ln L_k\}$, $\widetilde{\varepsilon}_k = \ln \varepsilon_k$, то (15.7) приводится к виду

$$\widetilde{X}_k = f_k^\top \theta + \widetilde{\varepsilon}_k, \quad k = 1, \ldots, n. \tag{15.8}$$

Видим, что (15.8) является стандартной линейной регрессионной моделью, причем $\widetilde{\varepsilon}_k \sim \mathcal{N}(0; \sigma^2)$.

Практически всегда такое сведение исходной модели к линейной приводит к появлению мультиколлинеарности.

15.2. Линеаризуемые модели

Предположим теперь, что в (15.1) функция $\varphi(\cdot)$ непрерывно дифференцируема по переменным θ и ε_k, и известно некоторое опорное (приближенное) значение вектора θ, а дисперсия σ^2 ошибок наблюдения $\{\varepsilon_k\}$ достаточно мала. В этом случае линеаризация $\varphi(h_k, \theta, \varepsilon_k)$ по θ и ε_k около опорных значений $\theta = \theta_0$ и $\varepsilon_k = 0$ дает приближенное соотношение

$$\varphi(h_k, \theta, \varepsilon_k) \approx f_k^0 + f_k^\top (\theta - \theta_0) + \sigma_k \varepsilon_k, \tag{15.9}$$

где

$$f_k = \left. \frac{\partial \varphi(h_k, \theta, \varepsilon_k)}{\partial \theta} \right|_{\theta=\theta_0, \varepsilon_k=0},$$

$$\sigma_k = \left. \frac{\partial \varphi(h_k, \theta, \varepsilon_k)}{\partial \varepsilon_k} \right|_{\theta=\theta_0, \varepsilon_k=0},$$

$$f_k^0 = \varphi(h_k, \theta_0, 0).$$

Заменим в (15.1) функцию $\varphi(h_k, \theta, \varepsilon_k)$ на ее приближение в виде правой части уравнения (15.9):

$$X_k = f_k^0 + f_k^\top (\theta - \theta_0) + \sigma_k \varepsilon_k. \tag{15.10}$$

Наконец, обозначая $\widetilde{X}_k = X_k - f_k^0 + f_k^\top \theta_0$, приводим модель (15.10) к виду (15.4). Теперь искомая оценка $\widehat{\theta}_n$ вектора θ определяется по формуле (15.5).

Пусть, например, модель (15.1) имеет вид

$$X_k = \theta_1 \sin(\theta_2 h_k) + \varepsilon_k, \quad k = 1, \ldots, n.$$

Пусть $\theta_0 = \{\widetilde{\theta}_1, \widetilde{\theta}_2\}^\top$, $\{h_k\}$ — известная числовая последовательность. Тогда из (15.9) следует, что $f_k^0 = \widetilde{\theta}_1 \sin\left(\widetilde{\theta}_2 h_k\right)$, $f_k^\top = \{\sin\left(\widetilde{\theta}_2 h_k\right); \widetilde{\theta}_1 h_k \cos\left(\widetilde{\theta}_2 h_k\right)\}$, $\widetilde{X}_k = X_k - \widetilde{\theta}_1 \widetilde{\theta}_2 h_k \cos\left(\widetilde{\theta}_2 h_k\right)$.

15.3. Метод Гаусса—Ньютона

Предполагается, что ошибки наблюдения в модели (15.1) являются аддитивными:

$$X_k = g_k(\theta) + \varepsilon_k, \quad k = 1, \ldots, n, \tag{15.11}$$

где $g_k(\theta) = \varphi(h_k, \theta, 0)$. Функции $g_k(\theta)$ нелинейным образом зависят от параметров θ, но являются непрерывно дифференцируемыми по θ для всех $k = 1, \ldots, n$.

Метод Гаусса—Ньютона является итерационным, поэтому мы через m обозначим номер итерации, а через $\theta(m)$ — приближение на m-й итерации к оценке $\widehat{\theta}_n$ вектора θ в модели (15.11), получаемой методом наименьших квадратов:

$$\widehat{\theta}_n = \arg\min_{\theta} \sum_{k=1}^{n} \left(X_k - g_k(\theta)\right)^2. \tag{15.12}$$

Заменим в модели (15.11) функцию $g_k(\theta)$ на ее линеаризированное значение, где в качестве опорного вектора θ_0 используется $\theta(m)$:

$$X_k = f_k^{(m)} + \left(F_k^{(m)}\right)^\top (\theta - \theta(m)) + \varepsilon_k, \tag{15.13}$$

где $f_k^{(m)} = g_k(\theta(m))$, $F_k^{(m)} = \left.\dfrac{\partial g_k(\theta)}{\partial \theta}\right|_{\theta = \theta(m)}$. Если обозначить $X_k^{(m)} = X_k - f_k^{(m)} + \left(F_k^{(m)}\right)^\top \theta(m)$, то (15.13) принимает вид линейной по θ регрессионной модели:

$$X_k^{(m)} = \left(F_k^{(m)}\right)^\top \theta + \varepsilon_k, \quad k = 1, \ldots, n. \tag{15.14}$$

Новое приближение $\theta(m+1)$ вычисляем, применяя к (15.14) метод наименьших квадратов:

$$\theta(m+1) = \left(\sum_{k=1}^{n} F_k^{(m)} \left(F_k^{(m)}\right)^\top\right)^{-1} \sum_{k=1}^{n} F_k^{(m)} X_k^{(m)}. \tag{15.15}$$

Так как $\sum\limits_{k=1}^{n} F_k^{(m)} X_k^{(m)} = \sum\limits_{k=1}^{n} F_k^{(m)}(X_k - f_k^m) + \sum\limits_{k=1}^{n} F_k^{(m)} \left(F_k^{(m)}\right)^\top \theta(m)$, то из (15.15) окончательно получаем:

$$\theta(m+1) = \theta(m) + \left(W_n^{(m)}\right)^{-1} \sum_{k=1}^{n} F_k^{(m)}(X_k - f_k^m), \tag{15.16}$$

где $W_n^{(m)} = \sum\limits_{k=1}^{n} F_k^{(m)} \left(F_k^{(m)}\right)^\top$.

Итерации по формуле (15.16) продолжаются до достижения сходимости. Например, в качестве критерия окончания итераций можно использовать $\dfrac{|\theta(m+1) - \theta(m)|}{|\theta(m)|} < \gamma$, где $\gamma > 0$ — заранее выбранная величина относительной погрешности. Если итерации прекратились при $m = m^*$, то мы полагаем $\widehat{\theta}_n = \theta(m^*)$. Для начала итерационного процесса следует выбрать начальное приближение $\theta(0)$, которое должно быть как можно ближе к искомой величине θ. Можно показать, что при определенных условиях итерационный процесс Гаусса—Ньютона сходится, т.е. $\theta(m) \to \widehat{\theta}_n$, $m \to \infty$.

Вычислить точно характеристики нелинейной оценки метода наименьших квадратов (15.12) в общем случае не представляется возможным. Однако если число наблюдений n достаточно велико, то обычно используется асимптотическое приближение

$$\widehat{\theta}_n \sim \mathcal{N}(\theta_{\text{И}}; K_n), \tag{15.17}$$

где $\theta_{\text{И}}$ — истинное значение вектора параметров θ,

$$K_n = \left(\sum_{k=1}^{n} F_k F_k^\top\right)^{-1}, \quad F_k = \left.\frac{\partial g_k(\theta)}{\partial \theta}\right|_{\theta=\theta_{\text{И}}}. \tag{15.18}$$

Вычислить точно вектор F_k невозможно, так как истинное значение $\theta_{\text{И}}$ вектора θ неизвестно. Поэтому при практическом использовании метода Гаусса—Ньютона обычно $\theta_{\text{И}}$ заменяется на $\widehat{\theta}_n$. Если оценка $\widehat{\theta}_n$ *сильно состоятельна*, (т.е. $\widehat{\theta}_n \to \theta_{\text{И}}$, $n \to \infty$ с вероятностью 1), то такая замена представляется обоснованной, если число n имеющихся наблюдений достаточно велико.

15.4. Примеры

Пример 15.1. По группе однотипных предприятий торговли имеется информация о связи между продолжительностью эксплуатации торгового оборудования и затратами на его ремонт. Соответствующие данные приведены в табл. 15.1, где y_k — затраты на ремонт (тыс. у.е.), x_k — возраст оборудования (лет) для k-го предприятия.

Таблица 15.1

k	1	2	3	4	5	6	7	8	9	10
x_k	4	5	5	6	8	10	8	7	11	6
y_k	1,5	2,0	1,4	2,3	2,7	4,0	2,3	2,5	6,6	1,7

Для решения проблемы нормирования расхода средств на ремонт оборудования необходимо построить модель зависимости между затратами на ремонт и возрастом оборудования.

Решение. Пусть $y(x)$ описывает реальную связь между переменными x и y. Рассмотрим вначале простейший вид связи между переменными

$$y(x) = \theta_1 + \theta_2 x, \tag{15.19}$$

где (θ_1, θ_2) — неизвестные параметры модели. В этом случае модель наблюдений, приведенных в табл. 15.1, имеет следующий вид:

$$y_k = \theta_1 + \theta_2 x_k + \varepsilon_k, \tag{15.20}$$

где k — номер предприятия, (y_k, x_k) — соответствующие данные о возрасте оборудования и затратах на ремонт, ε_k — случайная ошибка k-го наблюдения с распределением $\mathcal{N}(0; \sigma^2)$, $k = 1, \ldots, 10$.

Для оценивания параметров θ_1 и θ_2 в модели (15.20) применим метод наименьших квадратов:

$$(\widehat{\theta}_1, \widehat{\theta}_2)^\top = \arg\min_{\theta_1, \theta_2} \sum_{k=1}^{10} (y_k - \theta_1 - \theta_2 x_k)^2\,. \tag{15.21}$$

Используя данные табл. 15.1, находим реализацию МНК-оценки параметров:

$$\widehat{\theta}_1 = -1{,}576; \quad \theta_2 = 0{,}611. \tag{15.22}$$

Теперь из (15.19) и (15.22) следует, что оценку зависимости переменной y от x можно представить в виде

$$y(x) = -1{,}576 + 0{,}611x. \tag{15.23}$$

Для оценивания качества найденной аппроксимации (15.23) вычислим величину усредненной остаточной суммы квадратов ошибок:

$$\Delta^2 = \frac{1}{10} \sum_{k=1}^{10} (y_k - \widehat{y}(x_k))^2\,. \tag{15.24}$$

Данные для вычисления Δ^2 приведены в табл. 15.2.

Таблица 15.2

k	y_k	x_k	$\widehat{y}(x_k)$	$(y_k - \widehat{y}(x_k))^2$
1	1,5	4	0,868	0,399
2	2,0	5	1,479	0,271
3	1,4	5	1,479	0,006
4	2,3	6	2,090	0,044
5	2,7	8	3,312	0,374
6	4,0	10	4,534	0,285
7	2,3	8	3,312	1,024
8	2,5	7	2,700	0,040
9	6,6	11	5,145	2,117
10	1,7	6	2,090	0,152

Суммируя данные последнего столбца таблицы и деля полученную сумму на 10, находим

$$\Delta^2 = 0{,}471.$$

Теперь предпримем попытку улучшить оценку зависимости $y(x)$, переходя к нелинейной модели следующего вида:

$$y(x) = \theta_1 \cdot \theta_2^x. \tag{15.25}$$

Для оценивания параметров θ_1 и θ_2 в (15.25) воспользуемся методом приведения модели к линейной. Действительно, логарифмируя (15.25), получаем

$$\ln y(x) = \ln \theta_1 + \ln \theta_2 x. \tag{15.26}$$

Вводя новые переменные и параметры $\widetilde{y}(x) = \ln y(x)$, $a_1 = \ln \theta_1$, $a_2 = \ln \theta_2$, приходим к линейной (по параметрам a_1, a_2) модели зависимости

$$\widetilde{y}(x) = a_1 + a_2 x. \tag{15.27}$$

К модели (15.27) можно применить обычный МНК. Соответствующая система уравнений имеет вид

$$\begin{cases} 10a_1 + \sum\limits_{k=1}^{10} x_k a_2 = \sum\limits_{k=1}^{10} \ln y_k, \\ \sum\limits_{k=1}^{10} x_k a_1 + \sum\limits_{k=1}^{10} x_k^2 a_2 = \sum\limits_{k=1}^{10} x_k \ln y_k. \end{cases} \tag{15.28}$$

Используя данные табл. 15.1, находим

$$\widehat{a}_1 = 0{,}0840; \quad \widehat{a}_2 = -0{,}205. \tag{15.29}$$

Теперь из (15.29) получаем реализации оценок $\widehat{\theta}_1$, $\widehat{\theta}_2$ параметров θ_1, θ_2 исходной нелинейной модели (15.25):

$$\widehat{\theta}_1 = \exp\{\widehat{a}_1\} = 0{,}624; \quad \widehat{\theta}_2 = \exp\{\widehat{a}_2\} = 1{,}213.$$

Итак, окончательный вид подобранной нелинейной зависимости модели (15.25) таков:

$$\widehat{y}(x) = 0{,}624 \cdot 1{,}213^x. \qquad (15.30)$$

Используя (15.30) и данные табл. 15.1, составим табл. 15.3, аналогичную табл. 15.2.

Таблица 15.3

k	y_k	x_k	$\widehat{y}(x_k)$	$(y_k - \widehat{y}(x_k))^2$
1	1,5	4	1,36	0,0196
2	2,0	5	1,64	0,130
3	1,4	5	1,64	0,058
4	2,3	6	1,99	0,096
5	2,7	8	2,93	0,053
6	4,0	10	4,31	0,096
7	2,3	8	2,93	0,397
8	2,5	7	2,41	0,008
9	6,6	11	5,23	1,877
10	1,7	6	1,99	0,084

Сравнивая данные об ошибках оценивания, приведенных в последних столбцах табл. 15.2 и 15.3, можно видеть, что нелинейная модель превосходит по точности линейную. При этом среднее значение остаточной суммы квадратов Δ^2 для модели (15.30) имеет вид

$$\Delta^2 = \frac{1}{10} \sum_{k=1}^{10} (y_k - \widehat{y}(x_k))^2 = 0{,}282.$$

Сравнивая $\Delta^2 = 0{,}471$ для линейной модели (15.20) и $\Delta^2 = 0{,}282$ для нелинейной модели (15.25), заключаем, что искомая модель зависимости затрат на ремонт оборудования от срока его эксплуатации описывается зависимостью (15.30) лучше, чем зависимостью (15.23). ■

Пример 15.2. По группе наблюдений за показателями Y (объем производства), K (объем основных фондов) и L (объем труда), приведенных в табл. 15.4 и соответствующих периоду интенсивного развития экономики СССР с 1950 по 1969 г., требуется оценить параметры производственной функции с постоянной эластичностью замещения σ труда капиталом (подробнее об этих функциях см. [6]):

$$Y(K, L) = Ae^{\lambda t} \left(\delta K^{-\rho} + (1-\delta)L^{-\rho}\right)^{-\frac{1}{\rho}}, \qquad (15.31)$$

где A — параметр масштаба, λ — параметр технического прогресса, $\rho = \dfrac{1-\sigma}{\sigma}$, где σ — эластичность замещения, δ — параметр распределения.

Таблица 15.4

i	Y_i	K_i	L_i	i	Y_k	K_i	L_i
1	33,15	33,77	79,92	11	100,00	100,00	100,00
2	38,20	37,59	84,52	12	109,56	111,59	99,54
3	42,56	42,35	87,60	13	120,27	123,93	102,81
4	47,35	46,96	91,41	14	131,21	138,10	106,44
5	53,46	52,18	95,83	15	141,78	154,31	110,91
6	60,34	58,64	97,61	16	153,41	170,42	115,65
7	66,33	65,64	96,32	17	167,09	186,13	118,66
8	73,01	72,50	96,60	18	183,65	201,73	122,55
9	81,29	80,45	99,14	19	199,91	217,68	126,62
10	90,56	89,67	100,66	20	215,20	235,10	130,61

В табл. 15.4 все данные приведены в процентах к базовому 1960 г. ($i = 11$).

Решение. Обозначим $\theta_1 = A$, $\theta_2 = \lambda$, $\theta_3 = \rho$, $\theta_4 = \delta$, $t = i - 1$ (дискретное время); Y_i — валовой объем выпуска, K_i — объем основных фондов, L_i — объем труда в i-м году, $i = 1, \ldots, 20$. В этих обозначениях модель (15.31) для каждого $i = 1, \ldots, 20$ принимает вид

$$Y_i = \theta_1 \exp\{\theta_2(i-1)\} \left(\theta_4 K_i^{-\theta_3} + (1-\theta_4) L_i^{-\theta_3}\right)^{-\frac{1}{\theta_3}}. \qquad (15.32)$$

Обычно модель (15.32) используют после логарифмирования, хотя эта операция и не приводит (15.32) к линейному виду. Обозначим $g_i(\theta) = \ln Y_i$, тогда

$$g_i(\theta) = \ln \theta_1 + \theta_2(i-1) - \frac{1}{\theta_3} \ln\left(\theta_4 K_i^{-\theta_3} + (1-\theta_4) L_i^{-\theta_3}\right), \qquad (15.33)$$

где $\theta = \{\theta_1, \ldots, \theta_4\}^\top$ — вектор неизвестных параметров. МНК-оценка $\widehat{\theta}_n$ для θ по наблюдениям $\{Y_i, K_i, L_i\}$, $i = 1, \ldots, n$, где $n = 20$ имеет следующий вид:

$$\widehat{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{n} \left(\ln Y_i - g_i(\theta)\right)^2, \qquad (15.34)$$

где $g_i(\theta)$ определяется формулой (15.33).

Для нахождения решения задачи (15.34) можно применить итерационный алгоритм Гаусса–Ньютона (см. разд. 15.3):

$$\theta(m+1) = \theta(m) + \left(W_n^{(m)}\right)^{-1} \sum_{i=1}^{n} F_i^{(m)} (\ln Y_i - f_i^{(m)}),$$

где $F_i^{(m)} = \left.\dfrac{\partial g_i(\theta)}{\partial \theta}\right|_{\theta=\theta(m)}$, $f_i^{(m)} = g_i\left(\theta(m)\right)$, $W_n^{(m)} = \sum\limits_{i=1}^{n} F_i^{(m)} \left(F_i^{(m)}\right)^\top$, $\theta(m)$ — приближение к оценке $\widehat{\theta}_n$, вычисленной на m-й итерации.

Производные $F_i^{(m)}$ могут быть вычислены аналитически с использованием формулы (15.33). Однако на ЭВМ эти производные обычно вычисляются с помощью метода конечных разностей:

$$\frac{\partial g_i(\theta)}{\partial\theta} = \frac{g_1(\theta_1, \ldots, \theta_k + h, \ldots, \theta_p) - g_i(\theta_1, \ldots, \theta_k - h, \ldots, \theta_p)}{2h},$$

где p — число неизвестных параметров, $0 < h \ll 1$ — шаг численного дифференцирования, $k = 1, \ldots, p$.

В табл. 15.5 приведено описание итерационного процесса вычисления $\widehat{\theta}_n$. При проведении вычислений были использованы следующие параметры алгоритма: $\theta(0) = \{1; 0{,}05; 1; 0{,}5\}^\top$ — начальное приближение; $h = 0{,}0001$ — шаг численного дифференцирования. На каждом шаге вычислялась также остаточная сумма квадратов $RSS(m)$ по формуле

$$RSS(m) = \sum_{i=1}^{n} (\ln Y_i - g_i(\theta(m)))^2 .$$

Таблица 15.5

m	$\theta_1(m)$	$\theta_2(m)$	$\theta_3(m)$	$\theta_4(m)$	$RSS(m)$
0	1	0,05	1	0,5	5,6170
1	0,795	0,0192	1,358	0,652	0,0128
2	0,803	0,0206	1,438	0,638	0,0031
3	0,803	0,0206	1,478	0,638	0,0031
4	0,805	0,0204	1,482	0,640	0,0031
5	0,804	0,0205	1,483	0,639	0,0031
6	0,804	0,0205	1,484	0,639	0,0031
7	0,804	0,0205	1,484	0,640	0,0031

Видно, что итерационный процесс сходится достаточно быстро. В качестве реализаций оценок неизвестных параметров выберем результаты последней итерации:

$$\widehat{\theta}_n = \theta(7) = \{0{,}804; 0{,}0205; 1{,}484; 0{,}640\}^\top.$$

Используя найденные оценки параметров, построим оценку производственной функции $Y(K, L)$, подставив в (15.31) вместо точных значений параметров их оцененные значения:

$$Y(K, L) = 0{,}804 e^{0{,}0205t} \left(0{,}64 K^{-1{,}484} + 0{,}36 L^{-1{,}484}\right)^{-0{,}674} . \ \blacksquare$$

15.5. Задачи для самостоятельного решения

1. Используя данные табл. 15.4 методом приведения модели к линейной, оцените параметры модели производства Кобба—Дугласа [6]:

а) $Y(K, L) = Ae^{\lambda t} K^{\alpha} L^{1-\alpha}$;
б) $Y(K, L) = Ae^{\lambda t} K^{\alpha} L^{\beta}$.

2. Используя данные табл. 15.1, оцените параметры модели

$$y(x) = \theta_1 \cdot \theta_2^x$$

методом Гаусса—Ньютона. Сравните с результатами примера 15.1.

3. В условиях примера 15.2 найдите точные выражения для компонент вектора $F_i^{(m)}$.

4. Зависимость переменной y от x имеет вид дробно-рациональной функции

$$y(x) = \frac{\theta_1 + \theta_2 x}{\theta_3 + \theta_4 x}.$$

Разработайте алгоритм оценивания параметров $\{\theta_1, \ldots, \theta_4\}^\top$ по наблюдениям

$$y_k = y(x_k; \theta) + \varepsilon_k, \quad k = 1, \ldots, n$$

методом Гаусса—Ньютона.

16. Квантильная регрессия

16.1. Теоретические положения

В последнее время большой популярностью пользуется модель квантильной регрессии, рассматриваемая в этом разделе.

Задача квантильной регрессии отличается тем, что вместо поиска минимума квадратичной функции потерь (10.6), рассматривается другой критерий в виде суммы кусочно-линейных функций.

Рассмотрим модель линейной регрессии

$$X_k = h_k{}^\top \theta + e_k, \quad k = 1, \ldots, n. \tag{16.1}$$

или в векторно-матричной форме

$$Z_n = H_n \theta + E_n.$$

Рассмотрим функцию потерь вида

$$J_\alpha(\theta) = \sum_{k=1}^{n} g_\alpha \left(X_k - h_k^\top \theta \right), \tag{16.2}$$

где $\alpha \in (0; 1)$ — заданное число, а

$$g_\alpha(u) = u(\alpha - I(u < 0)), \tag{16.3}$$

где $I(u) = \begin{cases} 1, & u \in A, \\ 0, & u \notin A \end{cases}$ — индикаторная функция (если $A = \{u : u < 0\}$, то $I(u < 0) = I_A(u)$).

Пример кусочно-линейной функции $g_\alpha(u)$ изображен на рис. 16.1.

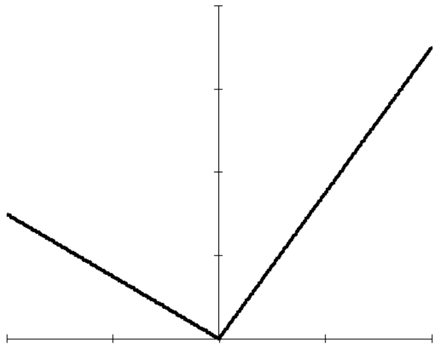


Рис. 16.1. Кусочно-линейная функция $g_{0,7}(u)$

Определение 16.1. Оценкой вектора θ для модели (16.1) в *задаче квантильной регрессии* [43] называется

$$\widehat{\theta}(\alpha) = \arg\min_{\theta \in \mathbb{R}^p} J_\alpha(\theta), \tag{16.4}$$

где $\alpha \in (0; 1)$, $J_\alpha(\theta)$ — функция вида (16.2).

Функцию $Q_X(\alpha|h) = h^\top \widehat{\theta}(\alpha)$ называют *функцией условной квантили.*

Задача (16.4) сводится к решению задачи линейного программирования (ЗЛП) [10] вида

$$\begin{aligned} &\alpha \sum_{k=1}^{n} u_k + (1-\alpha) \sum_{k=1}^{n} v_k \to \min_{\theta \in \mathbb{R}^p, u \in \mathbb{R}^n, v \in \mathbb{R}^n} \\ &h_k{}^\top \theta + u_k - v_k = X_k, \quad k = 1, \ldots, n, \\ &u_k \geqslant 0,\ v_k \geqslant 0, \quad k = 1, \ldots, n. \end{aligned} \tag{16.5}$$

Заметим, что для $\alpha = 0{,}5$ оценка $\widehat{\theta}(0{,}5)$ совпадает с оценкой метода наименьших модулей (см. разд. 14.3).

Рассмотрим частный случай оценки $\widehat{\theta}(\alpha)$ для тривиальной модели регрессии

$$X_k = \theta + e_k, \quad k = 1, \ldots, n.$$

В этом случае $\widehat{\theta}(\alpha)$ определяется как решение задачи

$$\sum_{k=1}^{n} g_\alpha(X_k - x_\alpha) \to \min_{x_\alpha \in \mathbb{R}^1}. \tag{16.6}$$

Решение задачи (16.6) можно свести к задаче оценивания квантили случайной величины по выборке. Пусть задана выборка $Z_n = \{X_1, \ldots, X_n\}^\top$, соответствующая СВ X с функцей распределения $F(x)$. Рассмотрим задачу

$$\begin{aligned} &\mathbf{M}\{g_\alpha(X - x_\alpha)\} = (\alpha - 1) \int_{-\infty}^{x_\alpha} (x - x_\alpha) dF(x) + \\ &+ \alpha \int_{x_\alpha}^{\infty} (x - x_\alpha) dF(x) \to \min_{x_\alpha \in \mathbb{R}^1}. \end{aligned} \tag{16.7}$$

Решение этой задачи совпадает с квантилью СВ X.

Теорема 16.1. *Если функция $F(x)$ монотонна, то решение x_α задачи* (16.7) *будет единственным и совпадает с квантилью уровня α функции $F(x)$.*

Если в задаче (16.7) функцию $F(x)$ заменить на выборочную функцию распределения $\widehat{F}_n(x)$, построенную по выборке Z_n, то мы получим задачу

$$\int_{-\infty}^{\infty} g_\alpha(x - x_\alpha)\, d\widehat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} g_\alpha(X_k - x_\alpha) \to \min_{x_\alpha \in \mathbb{R}^1}. \tag{16.8}$$

Решение задач (16.6) и (16.8) совпадают.

Замечание. Фактически, рассматривая задачу (16.8), мы вместо построения вариационного ряда для поиска квантили x_α предлагаем решать сравнительно простую задачу оптимизации, которая сводится к решению ЗЛП (16.5) вида

$$\begin{aligned} &\alpha \sum_{k=1}^{n} u_k + (1 - \alpha) \sum_{k=1}^{n} v_k \to \min_{x_\alpha \in \mathbb{R}^1, u \in \mathbb{R}^n, v \in \mathbb{R}^n} \\ &x_\alpha + u_k - v_k = X_k, \quad k = 1, \ldots, n, \\ &u_k \geqslant 0,\ v_k \geqslant 0, \quad k = 1, \ldots, n. \end{aligned} \tag{16.9}$$

Ниже приведены теоремы, раскрывающие свойства решения $\widehat{\theta}(\alpha)$ задачи квантильной регрессии.

Теорема 16.2. *Пусть P, N, Z — число положительных, отрицательных и нулевых элементов вектора остатков $\widehat{E}_n = Z_n - \widehat{\theta}(\alpha) H_n$. Если модель регрессии* (16.1) *содержит параметр сдвига,* *т.е. матрица* H_n *содержит столбец, состоящий из единиц, тогда для любого решения* $\widehat{\theta}(\alpha)$ *задачи* (16.4) *выполнены соотношения*

$$N \leqslant \alpha n \leqslant N + Z;$$

$$P \leqslant (1-\alpha)n \leqslant P + Z.$$

Следствие 16.1. *Рассмотрим модель* (16.1), *где матрица* H *имеет вид*

$$H = \begin{bmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{bmatrix},$$

вектор наблюдений также разбит на две части $X = \left[X_1{}^\top, X_2{}^\top\right]^\top$, $n = n_1 + n_2$. *Тогда решение* (16.4) *имеет вид* $\widehat{\theta}(\alpha) = \left[\widehat{\theta}_1(\alpha), \widehat{\theta}_2(\alpha)\right]^\top$, *где* $\widehat{\theta}_1(\alpha)$, $\widehat{\theta}_2(\alpha)$ *— решения задачи* (16.4) *по* n_1 *наблюдениям за параметром* θ_1 *и* n_2 *наблюдениям за параметром* θ_2.

Замечание. Если $Z = 0$, то линия квантильной регрессии $f(H) = H\widehat{\theta}(\alpha)$ делит наблюдаемые значения X_k, $k = 1, \ldots, n$ в соответствии со значением α: $\alpha \cdot 100\%$ наблюдений будет лежать выше построенной линии квантильной регрессии, а $(1-\alpha)\cdot 100\%$ наблюдений будет лежать ниже линии квантильной регрессии. При выборе $\alpha = 0{,}5$ наблюдения будут разделены линией регрессии пополам.

Обозначим решение задачи (16.4), построенное по наблюдениям Z_n и матрице H_n через $\widehat{\theta}(\alpha; Z_n, H_n)$.

Теорема 16.3. *Пусть* $A \in \mathbb{R}^{p\times p} > 0$, $\gamma \in \mathbb{R}^p$, $a > 0$ *заданные матрица, вектор и число. Тогда для любого* $\alpha \in (0,1)$:

1) $\widehat{\theta}(\alpha; aZ_n, H_n) = a\widehat{\theta}(\alpha; Z_n, H_n)$;

2) $\widehat{\theta}(\alpha; -aZ_n, H_n) = -a\widehat{\theta}(1-\alpha; Z_n, H_n)$;

3) $\widehat{\theta}(\alpha; Z_n + H_n\gamma, H_n) = \widehat{\theta}(\alpha; Z_n, H_n) + \gamma$;

4) $\widehat{\theta}(\alpha; Z_n, H_nA) = A^{-1}\widehat{\theta}(\alpha; Z_n, H_n)$.

По поводу вида закона распределения случайного вектора $\widehat{\theta}(\alpha)$ можно сказать, что он является асимптотически нормальным. Кроме того, сама оценка $\widehat{\theta}(\alpha)$ является асимптотически несмещенной оценкой вектора параметров θ. Более подробно эти вопросы освещены в монографии [43] (см. разд. 3.2, 4.1, 4.2).

16.2. Примеры

Пример 16.1. Докажите теорему 16.1.

Решение. Поскольку функция $g_\alpha(X - x_\alpha)$ является выпуклой по x_α, то и $\mathbf{M}\{g_\alpha(X - x_\alpha)\}$ выпуклая функция по x_α, следовательно, необходимым и достаточным условием минимума функции $\mathbf{M}\{g_\alpha(X - x_\alpha)\}$ является условие $\dfrac{d\mathbf{M}\{g_\alpha(X - x_\alpha)\}}{dx_\alpha} \geqslant 0$ (см. [10]).

Вычислим производную функции (16.7), воспользовавшись формулой дифференцирования интеграла по параметру:

$$\frac{d\mathbf{M}\{g_\alpha(X-x_\alpha)\}}{dx_\alpha}=\frac{d}{dx_\alpha}\left[(\alpha-1)\int\limits_{-\infty}^{x_\alpha}(x-x_\alpha)dF(x)+\right.$$

$$\left.+\,\alpha\int\limits_{x_\alpha}^{\infty}(x-x_\alpha)dF(x)\right]=(1-\alpha)\int\limits_{-\infty}^{x_\alpha}dF(x)-\alpha\int\limits_{x_\alpha}^{\infty}dF(x)=$$

$$=\int\limits_{-\infty}^{x_\alpha}dF(x)-\alpha\int\limits_{-\infty}^{x_\alpha}dF(x)-\alpha\int\limits_{x_\alpha}^{\infty}dF(x)=$$

$$=F(x_\alpha)-\alpha\int\limits_{-\infty}^{\infty}dF(x)=F(x_\alpha)-\alpha.$$

Таким образом, необходимым и достаточным условием минимума функции $\mathbf{M}\{g_\alpha(X-x_\alpha)\}$ будет $F(x_\alpha)\geqslant\alpha$. По определению квантили монотонной функции $F(x)$ это условие приводит к тому, что точкой минимума, а следовательно, и решением задачи (16.7) является квантиль x_α функции $F(x)$ уровня α. ■

Пример 16.2. Найдите оценку медианы (т.е. оцените квантиль уровня 0,5) по двум наблюдениям $X_1=-1$ и $X_2=1$.

Решение. Запишем ЗЛП (16.8) для $\alpha=0{,}5$ и $n=2$

$$\begin{aligned}
&u_1+u_2+v_1+v_2\to\min_{x_\alpha\mathbb{R},u\in\mathbb{R}^2,v\in\mathbb{R}^2}\\
&x_\alpha+u_1-v_1=X_1,\\
&x_\alpha+u_2-v_2=X_2,\\
&u_1\geqslant0,\ v_1\geqslant0\ u_2\geqslant0,\ v_2\geqslant0.
\end{aligned}$$

Эта задача не имеет аналитического решения, поэтому для ее решения можно воспользоваться стандартными пакетами типа Matlab, Mathcad, Maple, Mathematica. Решая задачу для $X_1=-1$ и $X_2=1$, получаем: $x_{0,5}=-1$. На самом деле решением этой задачи является отрезок $[-1;1]$, однако, в силу численного способа ее решения мы находим одно из возможных оптимальных решений, в данном случае левую границу отрезка. ■

Пример 16.3. Связь между данных $\{X_k\}$ и $\{h_k\}$ описывается моделью

$$X_k=\theta_1+\theta_2h_k+\theta_2{h_k}^2+\varepsilon_k,\quad k=1,\ldots,30,$$

где ε_k — независимые СВ с распределением $\mathcal{N}(0;2)$. Числовые данные, представлены в табл. 16.1.

Таблица 16.1

h_k X_k	$-1{,}0$ $3{,}62$	$-0{,}9$ $2{,}76$	$-0{,}8$ $2{,}81$	$-0{,}7$ $1{,}53$	$-0{,}6$ $-0{,}23$	$-0{,}5$ $2{,}96$	$-0{,}44$ $2{,}40$	$-0{,}3$ $3{,}55$
h_k X_k	$-0{,}2$ $6{,}64$	$-0{,}1$ $3{,}73$	$0{,}0$ $3{,}97$	$0{,}1$ $3{,}644$	$0{,}2$ $3{,}69$	$0{,}3$ $3{,}18$	$0{,}4$ $-0{,}25$	$0{,}5$ $2{,}01$
h_k X_k	$0{,}6$ $0{,}43$	$0{,}7$ $3{,}43$	$0{,}8$ $1{,}80$	$0{,}9$ $1{,}03$	$1{,}0$ $1{,}05$	$1{,}1$ $1{,}68$	$1{,}2$ $4{,}08$	$1{,}3$ $2{,}74$
h_k X_k	$1{,}4$ $3{,}72$	$1{,}5$ $6{,}40$	$1{,}6$ $2{,}83$	$1{,}7$ $4{,}64$	$1{,}8$ $7{,}27$	$1{,}9$ $7{,}30$		

Найдите реализацию оценки $\widehat{\theta}(\alpha)$ (см. задачу (16.4)) вектора $\theta = [\theta_1, \theta_2]^\top$ для α, равного 0,1, 0,5, 0,9, и сравните ее с реализацией МНК-оценки, построенной по этим же наблюдениям.

Сравните полученные результаты с истинным значением вектора параметров $\theta = [2, -1, 1{,}5]^\top$.

Решение. Решая ЗЛП (16.5), по имеющимся наблюдениям, для α, равного 0,1, 0,5, 0,9, и $n = 30$ получим реализации:

$$\widehat{\theta}(0{,}1) = \begin{bmatrix} -0{,}393 \\ -1{,}860 \\ 0{,}963 \end{bmatrix}, \quad \widehat{\theta}(0{,}5) = \begin{bmatrix} -1{,}613 \\ -1{,}768 \\ 1{,}203 \end{bmatrix}, \quad \widehat{\theta}(0{,}9) = \begin{bmatrix} 3{,}574 \\ -2{,}619 \\ 2{,}623 \end{bmatrix}.$$

Реализация МНК-оценки (10.5) равна $\widehat{\theta}_{\text{мнк}} = [2{,}168;\ -0{,}332;\ 1{,}186]^\top$.

Соответствующие реализации наблюдений и линий регрессии представлены на рис. 16.2.

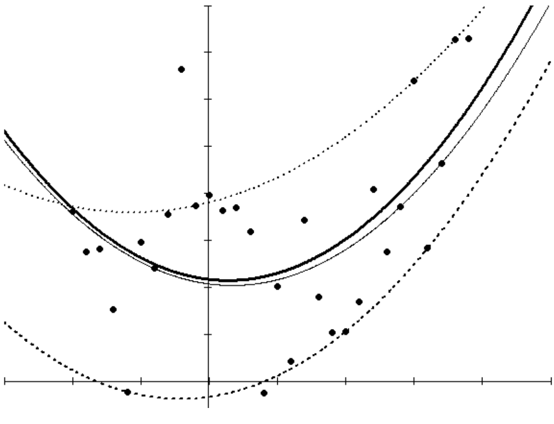


Рис. 16.2: • наблюдения; — $\widehat{\theta}_{\text{мнк}}$; — $\widehat{\theta}(0{,}5)$; --- $\widehat{\theta}(0{,}1)$; ··· $\widehat{\theta}(0{,}9)$

На рис. 16.2 видно (точнее можно сказать, если посмотреть на значения вектора $\widehat{E}_n = Z_n - \widehat{\theta}(\alpha)$), что для $\widehat{\theta}(\alpha)$ число положительных Z, отрицательных N и нулевых P элементов вектора $\widehat{E}_n$ равны

$$\begin{aligned} &\text{при } \alpha = 0{,}1: \ \ Z = 1,\ N = 3,\ P = 26; \\ &\text{при } \alpha = 0{,}5: \ \ Z = 3,\ N = 13,\ P = 14; \\ &\text{при } \alpha = 0{,}9: \ \ Z = 1,\ N = 27,\ P = 2. \end{aligned}$$

Таким образом, для $\alpha = 0{,}9$ линия регрессии проходит через одно наблюдение, два наблюдения лежат выше линии и 27 — ниже линии регрессии (см. рис. 16.2). Этот результат полностью согласуется с утверждением теоремы 16.2.

Вычислим реализации суммы квадратов отклонений линии регрессии от наблюдаемых значений $S = \sum\limits_{k=1}^{30} (X_k - h^\top{}_k \widehat{\theta})^2$, где вместо $\widehat{\theta}$ подставим $\widehat{\theta}(0{,}5)$ и $\theta_{\text{МНК}}$:

$$S_{\text{МНК}} = 79{,}65, \quad S_{0{,}5} = 81{,}09.$$

Хотя квантильная регрессия и не является решением задачи минимизации суммы квадратов S, тем не менее значение этой суммы при $\alpha = 0{,}5$ вполне можно сравнить с значением $S_{\text{МНК}}$ для МНК-оценки, поскольку $\widehat{\theta}(0{,}5)$ является также МНМ-оценкой (подробнее см. разд. 14.3). Видно, что разница между этими значениями в нашем случае невелика.

Вычисление величины S для квантильной регрессии со значениями $\alpha = 0{,}1$ или $0{,}9$ не имеет особенного смысла, поскольку они строятся скорее с целью получения аналога доверительного интервала для неизвестной зависимости, чем с целью уменьшения отклонения линии регрессии от наблюдаемых значений. Однако эти величины могут представлять некоторый интерес, поэтому приведем также и их: $S_{0{,}1} = 255{,}24$, $S_{0{,}9} = 169{,}7$.

В заключение для всех оценок $\widehat{\theta}(\alpha)$ приведем реализации значения квантильного критерия $\widetilde{S}_\alpha = \sum\limits_{k=1}^{30} g_\alpha\Big(X_k - h_k^\top \widehat{\theta}(\alpha)\Big)$:

$$\widetilde{S}_{0{,}1} = 35{,}74, \quad \widetilde{S}_{0{,}5} = 19{,}80, \quad \widetilde{S}_{0{,}9} = 27{,}36.$$

Рассмотренный пример показывает, что квантильная регрессия представляет собой новый аппарат построения параметрических оценок неизвестной зависимости с учетом предпочтений исследователя относительно ее положения по отношению к наблюдаемым значениям.
■

16.3. Задачи для самостоятельного решения

1. Покажите, что $\mathbf{M}\{g_\alpha(X - x_\alpha)\}$ является выпуклой функцией по x_α.

2. Покажите, что решения задач из примеров 14.5 и 14.6, полученные с помощью ВВНК-метода, совпадают с решениями задачи (16.4) для $\alpha = 0{,}5$, построенными по тем же наблюдениям.

3. Для модели простой регрессии

$$X_k = \theta_1 + \theta_2 h_k + \varepsilon_k, \quad k = 1, \ldots, n,$$

где ε_k — независимые центрированные СВ с одинаковой дисперсией, решите задачу (16.4) по трем наблюдениям $(h_k, X_k) = [(-1, -1); (0, 2); (1; 0)]$ при различных значениях $\alpha \in (0, 1)$. Проанализируйте полученные результаты, используя теорему 16.2.

4. По данным из примера 11.6 постройте график, на котором изобразите наблюдения и решение задачи (16.4) для $\alpha = [0{,}05, 0{,}1, \ldots, 0{,}95]$.

Г Л А В А IV

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

---

17. Временны́е ряды

17.1. Определение временно́го ряда

Предположим, что функция $X(t)$ описывает зависимость от времени t некоторого показателя или параметра изучаемой системы на временно́м промежутке $[0, T]$. Например, $X(t)$ может характеризовать зависимость от времени таких показателей, как урожайность зерновых культур, объем продаж некоторого товара, обменный курс валюты, объем остатков на счетах физических лиц в банке, количество продукции, произведенной некоторой фирмой, число дорожно-транспортных происшествий, скорость ветра и т.д.

Изменение во времени изучаемого показателя вызвано не только вполне определенными закономерностями неслучайной природы, но также и постоянно действующими неконтролируемыми факторами случайного характера. Поэтому функция $X(t)$ является *случайной*, т.е. $X(t)$ является случайной величиной при каждом фиксированном $t \in [0, T]$.

Обычно функция $X(t)$ доступна наблюдению на $[0, T]$ в моменты времени t_k, $k = 1, \ldots, n$, которые расположены в хронологическом порядке, т.е. $t_1 < t_2 < \ldots < t_n$. Кроме того, предполагается, что наблюдения проводятся через промежутки времени фиксированной длины Δt, так что

$$t_k = t_{k-1} + \Delta t, \quad k = 2, \ldots, n.$$

О п р е д е л е н и е 17.1. Совокупность случайных величин $Z_n = \{X_1, \ldots, X_n\}$, где $X_k = X(t_k)$, $k = 1, \ldots, n$ называется *временны́м рядом* (ВР). Случайная величина X_k называется k-м элементом временно́го ряда.

Таким образом, изучение динамики показателя $X(t)$ на практике сводится к изучению ВР Z_n.

17.2. Цели и задачи анализа временны́х рядов

Изучение свойств временны́х рядов методами математического и стохастического анализа называется *анализом временны́х рядов*.

Анализ ВР преследует следующие основные цели:

1) определение основных вероятностно-статистических характеристик ряда;

2) подбор подходящей математической модели ряда;

3) прогнозирование значений ряда по имеющимся к настоящему моменту времени T данным (т.е. предсказание величины $X(t)$ для некоторого $t > T$);

4) управление процессом $X(t)$, порождающим ВР Z_n.

Практический анализ ВР обычно подразделяется на несколько этапов, число и сложность которых определяют глубину и полноту эконометрического анализа динамики показателя $X(t)$. Отметим наиболее типичные этапы анализа ВР:

1) графическое представление ВР и визуальный анализ графических данных;

2) выделение и анализ детерминированной компоненты ВР;

3) удаление детерминированной компоненты;

4) анализ и идентификация случайной компоненты ВР;

5) прогнозирование ВР с использованием найденных моделей для детерминированной и случайной компонент.

Визуальный анализ графического представления ВР позволяет сформулировать некоторые исходные предположения о самых общих характеристиках ВР, исследование которых проводится далее количественными методами. Например, анализ графика ВР позволяет обычно сделать предварительное заключение о наличии или отсутствии некоторой основной (долговременной) тенденции (тренда) в динамике ряда, сезонной или циклической компоненты, свойства стационарности ряда и т.д. Если графический анализ ВР дает основание предполагать наличие существенной детерминированной компоненты, то далее проводятся действия по ее выделению (оцениванию) и анализу ее свойств. После чего производится статистический анализ полученного остатка ВР. Построенную в итоге модель ВР можно использовать для оценивания его будущих значений, т.е. для решения задачи прогнозирования ВР. Указанная задача является важнейшей проблемой эконометрического анализа, решение которой служит основанием для дальнейшего принятия тех или иных управленческих решений.

Более подробно указанные выше этапы анализа ВР будут описаны в разд. 18.

17.3. Модель временно́го ряда

Многочисленные прикладные исследования рядов экономических данных показывают, что в большинстве случаев временной ряд X_t, $t = 1, \ldots, n$ можно достаточно точно описать с помощью одной из следующих моделей:

а) *аддитивная модель*:

$$X_t = d_t + \varepsilon_t, \quad t = 1, 2, \ldots, n; \tag{17.1}$$

б) *мультипликативная модель*:

$$X_t = d_t \varepsilon_t, \quad t = 1, 2, \ldots, n. \tag{17.2}$$

В соотношениях (17.1) и (17.2) через d_t обозначена детерминированная компонента ВР, а через ε_t — его случайная компонента. Компонента d_t описывает некоторые закономерные (точно предсказуемые) изменения ВР X_t, а компонента ε_t предназначена для описания непредсказуемых (хаотических) изменений ВР X_t, вызванных наличием случайных неконтролируемых факторов, воздействующих в каждый момент времени t на изучаемый показатель.

Далее мы всегда будем предполагать, что СП $\{\varepsilon_t\}$ — центрированная, т.е. $\mathbf{M}\{\varepsilon_t\} = 0$, $t = 1, \ldots, n$. Поэтому компонента d_t является математическим ожиданием (средним значением) ВР X_t, описываемого моделью (17.1), т.е. $\mathbf{M}\{X_t\} = d_t$.

Если имеет место модель (17.2), а $\mathbf{D}\{\varepsilon_t\} = \sigma^2 > 0$, то компонента d_t связана с дисперсией ВР X_t следующим образом: $d_t^2 = \dfrac{\mathbf{D}\{X_t\}}{\sigma^2}$.

17.4. Модели детерминированной компоненты

В исследованиях, связанных с моделированием экономических и других показателей, обычно предполагается, что d_t можно представить в виде трех компонент, имеющих самостоятельный смысл:

$$d_t = \varphi_t + s_t + c_t, \quad t = 1, \ldots, n. \tag{17.3}$$

Компонента φ_t называется *трендом* и описывает достаточно плавные, долговременные тенденции в изменении анализируемого показателя X_t. Весьма часто тренд φ_t является монотонной функцией времени t.

Компонента s_t называется *сезонной* и отражает периодические колебания показателя X_t в течение годового цикла. Обычно s_t описывает периодические изменения X_t, связанные с какой-либо существенной зависимостью этого показателя от времени года. Например, если X_t описывает объем потребления прохладительных напитков как функцию времени, то наличие сезонной компоненты s_t в модели данного временно́го ряда представляется очевидным. Заметим, однако, что многие экономические параметры никак не связаны с какими-либо сезонными явлениями, поэтому соответствующие временные ряды не будут содержать компоненту s_t. Последнее справедливо, например, если X_t описывает такие показатели, как уровень производительности труда, валютообменные курсы, объем выплавки стали, количество регистрируемых изобретений и т.п.

Компонента c_t, называемая *циклической*, также описывает некоторые периодические изменения функции d_t, которые не связаны с явлениями сезонного характера. Иногда такие изменения называют *конъюнктурными*, так как они вызваны наличием периодов относительного подъема и последующего спада, характерными для большинства экономических показателей. Таким образом, присутствие циклической компоненты обусловлено конъюнктурой спроса и предложения на товары и услуги, колебаниями деловой активности, демографическими процессами и т.д. Совокупное влияние всех этих факторов приводит к тому, что модель циклической компоненты крайне трудно построить с помощью формальных статистических методов, основываясь только на данных изучаемого ВР без привлечения содержательной дополнительной информации о природе исследуемого показателя.

Рассмотренная выше модель (17.3) аддитивна по всем компонентам φ_t, s_t и c_t. В ряде практически важных случаев более подходящей оказывается *мультипликативно-аддитивная модель* следующего вида:

$$d_t = \varphi_t s_t + c_t, \quad t = 1, \ldots, n. \tag{17.4}$$

В модели (17.4) тренд φ_t выступает в роли меняющейся во времени «амплитуды» колебательного процесса s_t.

17.5. Модели тренда

Для математического описания тренда φ_t можно использовать типовые модели:

а) *линейная модель:*

$$\varphi_t = \theta_0 + \theta_1 t, \tag{17.5}$$

где θ_0, θ_1 — неизвестные постоянные параметры;

б) *полиномиальная модель* порядка $m \geqslant 0$:

$$\varphi_t = \theta_0 + \theta_1 t + \cdots + \theta_m t^m, \tag{17.6}$$

где $\theta = \{\theta_0, \ldots, \theta_m\}^\top$ — вектор параметров модели. Очевидно, что при $m = 0$ тренд отсутствует, т.е. $\varphi_t = \theta_0$. Если же $m = 1$, то модель (17.6) превращается в (17.5);

в) *логарифмическая модель:*

$$\ln \varphi_t = \theta_0 + \theta_1 t. \tag{17.7}$$

Модель (17.7) используется в случае, когда предварительный анализ данных показывает, что φ_t имеет свойство *постоянства темпа роста*, т.е.

$$\frac{\varphi_2}{\varphi_1} = \frac{\varphi_3}{\varphi_2} = \cdots = \frac{\varphi_n}{\varphi_{n-1}} = \text{const}.$$

Действительно, из (17.7) следует, что $\varphi_t = \exp\{\theta_0 + \theta_1 t\}$, поэтому

$$\frac{\varphi_{t+1}}{\varphi_t} = \frac{\exp\{\theta_0 + \theta_1(t+1)\}}{\exp\{\theta_0 + \theta_1 t\}} = \frac{\exp\{\theta_0 + \theta_1 t\}\exp\{\theta_1\}}{\exp\{\theta_0 + \theta_1 t\}} = e^{\theta_1} = \text{const};$$

г) *логистическая модель:*

$$\varphi_t = \frac{\theta_0}{1 + \theta_1 e^{-\theta_2 t}}, \quad \theta_2 > 0; \tag{17.8}$$

д) *модель Гомперца:*

$$\ln \varphi_t = \theta_0 + \theta_1 \alpha^t, \tag{17.9}$$

где $\alpha \in (0, 1)$.

Модели г) и д) описывают процессы с переменными темпами роста (в начале процесса темп роста возрастает, а к концу — затухает).

Модели (17.5) — (17.9) можно представить в виде

$$\varphi_t = f(t; \theta), \tag{17.10}$$

где θ — вектор неизвестных параметров.

Модели вида (17.10) называются *параметрическими моделями*, причем модель (17.10) называется *линейной*, если $f(t; \theta)$ при каждом t зависит от θ линейным образом. В остальных случаях она описывает *нелинейную параметрическую модель*. Заметим, что модели (17.5) и (17.6) — линейные, а (17.7) – (17.9) — нелинейные. Кроме того, только модель (17.5) зависит от t линейным образом, а все остальные нелинейны по t.

17.6. Модели периодических компонент

К периодическим относятся компоненты s_t и c_t. Для описания сезонной компоненты s_t в простейшем случае используют параметрическую модель гармонических колебаний:

$$s_t = \theta_0 \sin\left(\frac{2\pi}{\Omega} t + \theta_1\right), \quad t = 1, \ldots, n, \tag{17.11}$$

где Ω — период колебаний (т.е. $s_{t+\Omega} = s_t$ для всех $t \geqslant 0$), а θ_0 и θ_1 — неизвестные параметры модели. Конкретное значение Ω зависит от длины временного промежутка Δt между соседними элементами ВР. Например, если период сезонных колебаний равен одному году, а элементы X_t и X_{t+1} ВР отстоят друг от друга по времени на $\Delta t = 1$ месяц, то $\Omega = 12$. Если же наблюдения $\{X_1, \ldots, X_n\}$ получены с частотой 1 раз в день, то $\Omega = 365$ и т.д.

Для моделирования циклической компоненты c_t можно использовать формальные соотношения, обобщающие (17.11). Например,

$$c_t = \sum_{k=1}^{r} a_k \sin\Big(\frac{2\pi}{\Omega_k}\, t + b_k\Big), \tag{17.12}$$

где $\theta = \{a_1, \ldots, a_r, b_1, \ldots, b_r\}^\top$ — вектор параметров модели.

Заметим, что при практическом анализе временны́х рядов зачастую удается учесть или, наоборот, устранить влияние сезонных и циклических компонент без использования моделей (17.11) и (17.12), опираясь лишь на свойство периодичности, являющееся характеристическим для компонент s_t и c_t.

17.7. Модели случайной компоненты

Относительно случайной компоненты ε_t, присутствующей в моделях (17.1) и (17.2) временного ряда X_t, обычно предполагается, что СП $\{\varepsilon_t, t = 0, \pm 1, \pm 2, \ldots\}$ обладает свойствами:

а) центрированность:

$$\mathbf{M}\{\varepsilon_t\} = 0; \tag{17.13}$$

б)

$$\mathbf{M}\big\{\varepsilon_t^2\big\} < \infty; \tag{17.14}$$

в) стационарность в широком смысле:

$$\mathbf{cov}(\varepsilon_t, \varepsilon_k) = \mathbf{cov}(\varepsilon_{t+m}, \varepsilon_{k+m}) \tag{17.15}$$

для любых целых значений t, k и m.

Случайную последовательность $\{\varepsilon_t\}$, обладающую свойствами (17.13) — (17.15), далее будем называть *стационарной случайной последовательностью* (ССП).

Из (17.15) следует, что для любых целых t и k

$$\mathbf{cov}(\varepsilon_t, \varepsilon_{t+k}) = \mathbf{cov}(\varepsilon_0, \varepsilon_k).$$

Определение 17.2. Функция $R_\varepsilon(k) = \mathbf{cov}(\varepsilon_0, \varepsilon_k)$, $k = 0, \pm 1, \pm 2, \ldots$ называется *ковариационной функцией* ССП $\{\varepsilon_t\}$.

Определение 17.3. Функция $D_\varepsilon(t) = \mathbf{cov}(\varepsilon_t, \varepsilon_t)$, $t = 0, \pm 1, \pm 2, \ldots$ называется *дисперсией* СП $\{\varepsilon_t\}$.

Из условия стационарности ССП следует, что дисперсия ССП $D_\varepsilon(t)$ постоянна (не зависит от времени), так как $D_\varepsilon(t) = \mathbf{cov}(\varepsilon_0, \varepsilon_0) = R_\varepsilon(0) = \mathrm{const}$ для любого t. Заметим, что постоянство математического ожидания и дисперсии стационарной случайной последовательности является ее важным характеристическим свойством. В частности, необходимым условием стационарности ВР $\{X_t\}$ является $d_t = 0$.

Определение 17.4. Пусть $D_{\varepsilon}(t) = \sigma^2 > 0$, тогда функция

$$r_{\varepsilon}(k) = \frac{R_{\varepsilon}(k)}{\sigma^2} \tag{17.16}$$

называется *корреляционной функцией* ССП $\{\varepsilon_t\}$.

Определение 17.5. Если СП $\{\varepsilon_t, t = 0, \pm 1, \pm 2, \ldots\}$ такова, что СВ ε_t независимы в совокупности и имеют одинаковую функцию распределения $F(x)$, то СП называется *белым шумом.* Если, кроме того, $R_{\varepsilon}(k) = 0$ при всех $k \neq 0$, то СП $\{\varepsilon_t\}$ называется *стационарным белым шумом.*

Если между элементами ряда X_t имеется корреляционная зависимость, то это указывает на то, что случайная компонента ε_t является СП, отличной от белого шума. Для описания ε_t в эконометрических исследованиях в этом случае широко применяются *параметрические модели*, формирующие СП ε_t с необходимыми корреляционными характеристиками из белого шума. Стационарными моделями такого типа являются следующие:

а) *модель авторегрессии* порядка $p \geqslant 1$ (АР(p)-модель):

$$\varepsilon_t = a_1\varepsilon_{t-1} + \ldots + a_p\varepsilon_{t-p} + e_t, \tag{17.17}$$

где $\theta = \{a_1, \ldots, a_p\}^\top$ — вектор параметров модели, а $\{e_t\}$ — стационарный белый шум с параметрами $\mathbf{M}\{e_t\} = 0$, $\mathbf{D}\{e_t\} = \sigma_e^2 > 0$. СП $\{\varepsilon_t\}$ будет стационарной, если θ удовлетворяет *условию асимптотической устойчивости*: все корни алгебраического уравнения

$$x^p - a_1x^{p-1} - \ldots - a_p = 0 \tag{17.18}$$

лежат строго внутри круга единичного радиуса с центром в нуле (единичного круга);

б) *модель скользящего среднего* порядка $q \geqslant 1$ (СС(q)-модель)

$$\varepsilon_t = e_t + b_1e_{t-1} + \ldots + b_qe_{t-q}, \tag{17.19}$$

где $\theta = \{b_1, \ldots, b_q\}^\top$ — вектор параметров модели. Если θ удовлетворяет условию: все корни уравнения

$$x^q + b_1x^{q-1} + \ldots + b_q = 0 \tag{17.20}$$

лежат внутри единичного круга, то СП ε_t является стационарной. Данное условие известно как *условие обратимости* модели (17.19);

в) *модель авторегрессии и скользящего среднего* порядка (p, q) (АРСС (p, q)-модель) объединяет модели (17.17) и (17.19) в одну более сложную модель:

$$\varepsilon_t = a_1\varepsilon_{t-1} + \ldots + a_p\varepsilon_{t-p} + e_t + b_1e_{t-1} + \ldots + b_qe_{t-q}. \tag{17.21}$$

СП (17.21) является стационарной, если вектор параметров $\theta = \left[a_1, \ldots, a_p, b_1, \ldots, b_q\right]^\top$ удовлетворяет условиям асимптотической устойчивости (17.19) и обратимости (17.20).

Можно показать, что СП $\{\varepsilon_t\}$, удовлетворяющая уравнению (17.21), может быть представлена в виде бесконечного скользящего среднего:

$$\varepsilon_t = e_t + \sum_{m=1}^{\infty} \beta_m e_{t-m}, \tag{17.22}$$

где коэффициенты $\{\beta_m\}$ выражаются через коэффициенты θ модели (17.21) и удовлетворяют условию $\sum\limits_{k=1}^{\infty} |\beta_m| < \infty$.

Заметим также, что ковариационная функция $R_\varepsilon(k)$ процесса (17.22) имеет вид

$$R_\varepsilon(k) = \sigma_e^2 \sum_{m=0}^{\infty} \beta_m \beta_{m+|k|}, \tag{17.23}$$

где $\beta_0 = 1$. В частности, из того, что $\mathbf{D}\{\varepsilon_t\} = R_\varepsilon(0)$, следует, что

$$\sigma_\varepsilon^2 = \mathbf{D}\{\varepsilon_t\} = \sigma_e^2 \sum_{m=0}^{\infty} \beta_m^2. \tag{17.24}$$

Из (17.24) для модели СС(q) вида (17.20) следует, что $\sigma_\varepsilon^2 = \left(1 + b_1^2 + \cdots + b_q^2\right)\sigma_e^2$, причем $R_\varepsilon(k) = 0$ для всех k таких, что $|k| > q$. Таким образом, коррелированными оказываются только элементы ряда ε_t, отстоящие друг от друга не более, чем на q шагов.

Более подробно с моделями авторегрессии и скользящего среднего можно познакомиться в монографиях [3, 41].

18. Анализ и прогнозирование нестационарных временны́х рядов

В данном разделе мы рассмотрим прикладные методы анализа временны́х рядов (ВР), предназначенные для решения задачи оценивания и прогнозирования детерминированной компоненты d_t временно́го ряда $\{X_t, t = 1, \ldots, n\}$.

18.1. Выделение тренда методом наименьших квадратов

Рассмотрим аддитивную модель ВР

$$X_t = d_t + \varepsilon_t, \quad t = 1, \ldots, n, \tag{18.1}$$

где d_t — детерминированная компонента ВР, а $\{\varepsilon_t\}$ — случайная компонента, представляющая стационарный белый шум, $\varepsilon_t \sim \mathcal{N}(0; \sigma^2)$.

Предположим, что d_t не содержит сезонной s_t и циклической c_t компонент, а тренд φ_t является достаточно гладкой функцией времени t:

$$\varphi_t = \varphi_0(t)\theta_0 + \varphi_1(t)\theta_1 + \cdots + \varphi_m(t)\theta_m, \quad t = 1, \ldots, n, \tag{18.2}$$

где $\{\varphi_k(t), k = 0, \ldots, m\}$ — известные функции времени t, $\theta = \{\theta_0, \ldots, \theta_m\}^\top$ — вектор неизвестных параметров, $p = m + 1$ — заданный порядок модели.

Обозначим $Z_n = [X_1, \ldots, X_n]^\top$, $h_t = [\varphi_0(t), \ldots, \varphi_m(t)]^\top$, H_n — матрица размера $n \times (m+1)$, строками которой являются $h_1^\top, \ldots, h_n^\top$, $E_n = [\varepsilon_1, \ldots, \varepsilon_n]^\top$. Тогда из (18.1) и (18.2) в предположении, что $d_t = \varphi_t$, следует

$$Z_n = H_n\theta + E_n. \tag{18.3}$$

Модель (18.3) является линейной регрессионной моделью, изученной ранее в разд. 10.1.

Применяя для оценивания θ в (18.3) метод наименьших квадратов, получаем оценку

$$\widehat{\theta}_n = W_n^{-1} H_n^\top Z_n, \tag{18.4}$$

где $W_n = H_n^\top H_n$.

Пусть $1 \leqslant t \leqslant n$, тогда *оценка сглаживания* детерминированной компоненты d_t имеет вид

$$\widehat{d}_t = \widehat{\varphi}_t = h_t^\top \widehat{\theta}_n. \tag{18.5}$$

В рамках сделанных выше предположений оценка $\widehat{d}_t$ обладает следующими статистическими свойствами:

$$1) \quad \mathbf{M}\left\{\widehat{d}_t\right\} = d_t, \quad t = 1, \ldots, n; \tag{18.6}$$

$$2) \quad \mathbf{M}\left\{\left(\widehat{d}_t - d_t\right)^2\right\} = \sigma^2 h_t^\top W_n^{-1} h_t. \tag{18.7}$$

Обычно на практике дисперсия σ^2 случайной компоненты $\{\varepsilon_t\}$ неизвестна. Это нисколько не затрудняет вычисление $\widehat{d}_t$, так как формулы (18.4) и (18.5) не содержат σ^2. Однако выражение (18.7) для дисперсии ошибки оценки $\widehat{d}_t$ содержит σ^2 в явном виде. Поэтому оцениv σ^2 по наблюдениям Z_n следующим образом:

$$\widehat{\sigma}_n^2 = \frac{1}{n-(m+1)} \sum_{t=1}^{n} \left(X_t - h_t^\top \widehat{\theta}_n \right)^2 . \tag{18.8}$$

Заметим, что $\widehat{\varepsilon}_t = X_t - h_t^\top \widehat{\theta}_n$ является оценкой для случайной компоненты ε_t, а оценка $\widehat{\sigma}_n^2$ является несмещенной и состоятельной.

После того как оценка $\widehat{\theta}_n$ построена, можно найти прогноз $\widetilde{d}_t$ для компоненты d_t на момент $t = n+k$, где $k \geqslant 1$ по формуле (18.5):

$$\widetilde{d}_t = h_t^\top \widehat{\theta}_n, \tag{18.9}$$

где $h_t^\top = [\varphi_0(n+k), \ldots, \varphi_m(n+k)]$, $t = n+k$. При этом точность прогноза $\widetilde{d}_t$ вычисляется по формуле (18.7).

При практическом анализе ВР в качестве функций $\varphi_k(t)$ обычно выбираются алгебраические полиномы: $\varphi_k(t) = t^k$, $k = 0, 1, \ldots, m$.

Если априори порядок $p = m+1$ модели тренда не задан, то его можно подобрать, используя следующий несложный метод.

Пусть $\widehat{\theta}_n(m)$ — МНК-оценка вектора параметров θ, вычисленная в предположении, что $\theta = [\theta_0, \ldots, \theta_m]^\top$, т.е. $m+1$ — правильный порядок модели тренда φ_t. Пусть также

$$\widehat{\sigma}_n^2(m) = \frac{1}{n-(m+1)} \sum_{t=1}^{n} \left(X_t - h_t^\top \widehat{\theta}_n(m) \right)^2, \quad m = 0, 1, \ldots . \tag{18.10}$$

Вычислим $\widehat{\sigma}_n^2(0)$ и $\widehat{\sigma}_n^2(1)$ по формуле (18.10). Если $\widehat{\sigma}_n^2(0) > \widehat{\sigma}_n^2(1)$, то вычисляем $\widehat{\sigma}_n^2(2)$, сравниваем $\widehat{\sigma}_n^2(2)$ с $\widehat{\sigma}_n^2(1)$ и т.д. Пусть для некоторого m_0 реализовалась следующая ситуация: $\widehat{\sigma}_n^2(m_0 - 1) > \widehat{\sigma}_n^2(m_0)$, но $\widehat{\sigma}_n^2(m_0) \leqslant \widehat{\sigma}_n^2(m_0 + 1)$. Тогда полагаем, что оценка $\widehat{p}$ порядка p модели тренда φ_t равна $m_0 + 1$. Можно показать, что если истинный порядок модели p^* намного меньше числа n элементов ВР, то с высокой вероятностью $m_0 + 1 \geqslant p^*$, т.е. истинный порядок модели будет найден (точнее, $\mathbf{P}(m_0 + 1 < p^*) \to 0$ при $n \to \infty$). Описанный алгоритм известен как «упрощенный вариант критерия Фишера».

В заключение заметим, что при нелинейной параметризации модели тренда (см. (17.7) — (17.9)), оценки вектора θ можно вычислить с помощью нелинейного МНК, рассмотренного в разд. 15.

18.2. Оценивание сезонной компоненты

Пусть ВР описывается аддитивной моделью

$$X_t = \varphi_t + s_t + \varepsilon_t, \quad t = 1, \ldots, n,$$

где сезонная компонента имеет период $r > 0$, т.е.

$$s_t = s_{t+r} \text{ для всех } 1 \leqslant t, t+r \leqslant n.$$

Будем полагать, что величина r известна, а $n = (m+1)r$, где m — целое число (на интервале наблюдения укладывается целое число периодов).

Пусть построена предварительная оценка $\widehat{\varphi}_t$ тренда φ_t. Рассмотрим для каждого сезона i, где $1 \leqslant i \leqslant r$ разности

$$X_i - \widehat{\varphi}_i, \quad X_{i+r} - \widehat{\varphi}_{i+r}, \ldots, X_{i+mr} - \widehat{\varphi}_{i+mr}.$$

Каждое из этих отклонений можно рассматривать как результат наличия сезонной компоненты, причем в силу периодичности компоненты s_t величины

$$\delta_{i,l} = X_{i+lr} - \widehat{\varphi}_{i+lr}, \quad l = 0, \ldots, m, \tag{18.11}$$

можно трактовать как измерения значения сезонной компоненты s_i для $i = 1, \ldots, r$. Поэтому для получения оценки $\widehat{s}_i$ величины s_i можно вычислить выборочное среднее величин $\{\delta_{i,l}, \ l = 0, \ldots, m\}$:

$$\widehat{s}_i = \frac{1}{m+1} \sum_{l=0}^{m} \delta_{i,l}, \quad i = 1, \ldots, r. \tag{18.12}$$

Используя периодичность s_t, получаем для каждого $t = 1, \ldots, n$ соотношение

$$\widehat{s}_t = \widehat{s}_i,$$

где i — целый остаток от деления t на период r.

Если модель ВР имеет мультипликативно-аддитивный вид

$$X_t = \varphi_t s_t + \varepsilon_t, \quad t = 1, \ldots, n,$$

то формула (18.12) остается в силе, а величины $\{\delta_{i,l}\}$ вычисляются по формуле

$$\delta_{i,l} = \frac{X_{i+lr}}{\widehat{\varphi}_{i+lr}}, \quad l = 0, \ldots, m; \ i = 1, \ldots, r. \tag{18.13}$$

Заметим, что вместо выборочного среднего (18.12) для оценивания $\{s_i\}$ можно использовать выборочную медиану, рассмотренную в разд. 14.2:

$$\widehat{s}_i = \operatorname{med}\{\delta_{i,0}; \delta_{i,1}; \ldots; \delta_{i,m}\}. \tag{18.14}$$

Оценка (18.14) обладает существенно большей устойчивостью к аномально большим значениям $\delta_{i,k}$ по сравнению с оценкой (18.12).

Получив оценку $\{\widehat{s}_t, \ t = 1, \ldots, n\}$ сезонной компоненты, мы можем провести *сезонное выравнивание* временного ряда, удалив из него сезонную компоненту. Для аддитивного ВР

$$\widehat{X}_t = X_t - \widehat{s}_t, \quad t = 1, \ldots, n. \tag{18.15}$$

Теперь ряд $\{\widehat{X}_t,\ t = 1, \ldots, n\}$ содержит только тренд φ_t, случайную компоненту ε_t и остатки $s_t - \widehat{s}_t = \Delta\widehat{s}_t$ от сезонной компоненты, которые могут быть отнесены к случайной компоненте. Процедуру оценивания тренда можно повторить, используя в МНК наблюдения $\{\widehat{X}_t\}$ вместо $\{X_t\}$. Таким образом, мы получим уточненную оценку тренда $\widetilde{\varphi}_t$, $t = 1, \ldots, n$. Теперь можно построить прогноз детерминированной компоненты d_t временного ряда для любого $t = n + k$, $k \geqslant 0$:

$$\widetilde{d}_t = \widetilde{\varphi}_t + \widetilde{s}_t, \quad t = n + k,\ k \geqslant 1,$$

$\widetilde{s}_t = \widehat{s}_i$, $n + k = lr + i$, а $l \geqslant m$ — целое число.

Для сезонного выравнивания мультипликативного ряда вместо (18.15) следует использовать

$$\widehat{X}_t = \frac{X_t}{\widehat{s}_t}, \quad t = 1, \ldots, n,$$

а прогноз компоненты d_t вычисляется по формуле

$$\widetilde{d}_t = \widetilde{\varphi}_t \cdot \widetilde{s}_t, \quad t = n + k,\ k \geqslant 1.$$

Для построения предварительной оценки $\widehat{\varphi}_t$ тренда обычно пользуются одним из следующих способов:

1) $\widehat{\varphi}_t^{(1)}$ — *МНК-оценка* тренда φ_t, построенная по формуле (18.5) без учета наличия в модели сезонной компоненты s_t;

2) $\widehat{\varphi}_t^{(2)}$ — результат *скользящего осреднения* ряда $\{X_t\}$:

$$\widehat{\varphi}_t^{(2)} = \begin{cases} \dfrac{1}{2m+1} \displaystyle\sum_{k=-m}^{m} X_{t-k}, & \text{если } r = 2m+1, \\ \dfrac{1}{2m} \left[\displaystyle\sum_{k=-m+1}^{m-1} X_{t-k} + \dfrac{1}{2}\bigl(X_{t-m} + X_{t+m}\bigr) \right], & \text{если } r = 2m, \end{cases} \tag{18.16}$$

где $r \geqslant 2$ — период сезонной компоненты, причем t таковы, что $t - m \geqslant 1$, а $t + m \leqslant n$; m — параметр, определяемый заранее;

3) $\widehat{\varphi}_t^{(3)}$ — результат *экспоненциального сглаживания* ряда по формуле

$$\begin{cases} \widehat{\varphi}_t^{(3)} = \alpha\widehat{\varphi}_{t-1}^{(3)} + (1 - \alpha)X_t, \\ \widehat{\varphi}_0^{(3)} = 0, \end{cases} \tag{18.17}$$

где $\alpha \in (0, 1)$ — параметр сглаживания, выбираемый заранее исходя из имеющихся представлений о скорости изменения тренда φ_t.

Сделаем некоторые замечания относительно предложенных методов нахождения $\widehat{\varphi}_t^{(i)}$, $i = 1, 2, 3$.

1. Оценка $\widehat{\varphi}_t^{(1)}$ может быть весьма неточной, если число периодов r мало по сравнению с n. В этом случае компонента s_t скорее напоминает тренд, а не случайную компоненту $\{\varepsilon_t\}$, на исключение которой нацелен МНК.

2. Оценка $\widehat{\varphi}_t^{(2)}$ используется в случаях, когда модель тренда не ясна, а также в составе d_t присутствует циклическая компонента c_t. Возможность избежать четкого описания моделей компонент φ_t и c_t является достоинством метода. В то же время из (18.16) следует, что φ_t будет хорошо оцениваться, если параметр m, определяющий длину интервала осреднения, достаточно велик (так как удается исключить случайную компоненту). С другой стороны, использование в (18.16) выборочного среднего означает, что систематическая погрешность оценки $\widehat{\varphi}_t^{(2)}$ будет маленькой, если значения $\varphi_{t-m}, \ldots, \varphi_{t+m}$ тренда на интервале осреднения мало отличаются друг от друга. Последнее означает, что параметр m не может быть слишком большим. Кроме того, из (18.16) следует, что мы получаем оценки тренда не для всех $t = 1, \ldots, n$, а только для $t = m+1, \ldots, n-m$. Таким образом, для оценивания φ_t на «концах» интервала наблюдений следует применить некоторые дополнительные меры (например, построить МНК-оценки φ_t, используя линейную или параболическую локальную модель тренда).

3. При использовании алгоритма (18.17) экспоненциального сглаживания важно правильно выбрать величину параметра α. Действительно, из (18.17) следует, что

$$\widehat{\varphi}_t = (1-\alpha)\sum_{k=0}^{t-1} \alpha^k X_{t-k}. \tag{18.18}$$

Поэтому, если α близко к 1, то $\widehat{\varphi}_t$ есть усреднение с приблизительно постоянными весами всех наблюдений, предшествующих текущему моменту t. Это означает, что оценка $\widehat{\varphi}_t$ будет удовлетворительной, если динамика тренда φ_t на интервале наблюдения невелика (заметим, что из (18.18) следует, что если $\varphi_t = \text{const}$, а $\varepsilon_t = 0$, то $\widehat{\varphi}_t = \varphi_t$). Если же тренд φ_t на интервале наблюдения меняется существенно, то α следует выбирать близким к нулю (в этом случае $\widehat{\varphi}_t$ является усреднением лишь последних элементов временно́го ряда, так как α^k быстро стремится к нулю при увеличении k).

Модель (18.17) чрезвычайно удобна для прогнозирования, так как для $t > n$ прогноз $\widetilde{\varphi}_t$ для φ_t строится рекуррентно по формуле

$$\begin{cases} \widetilde{\varphi}_t^{(3)} = \alpha\widetilde{\varphi}_{t-1}^{(3)}, \quad t \geqslant n+1, \\ \widetilde{\varphi}_n^{(3)} = \widehat{\varphi}_n^{(3)}. \end{cases} \tag{18.19}$$

Естественно, точность прогноза (18.19) будет быстро падать при увеличении t, так как $\widetilde{\varphi}_t^{(3)} \to 0$ при $t \to \infty$ в силу того, что $\alpha \in (0,1)$.

18.3. Выделение тренда в модели с коррелированными ошибками

Выше предполагалось, что случайная компонента $\{\varepsilon_t\}$ является стационарным белым шумом. Кратко рассмотрим процедуру проверки гипотезы о «белошумности» СП $\{\varepsilon_t\}$.

Предположим, что мы оценили тренд и по формуле (18.5) построили оценки остатков $\{\widehat{\varepsilon}_t\}$. Для каждого $k \geqslant 1$ рассмотрим оценку $\widehat{r}_{\widehat{\varepsilon}}(k)$ корреляции $r_{\widehat{\varepsilon}}(k) = \dfrac{\mathbf{cov}\,(\widehat{\varepsilon}_t, \widehat{\varepsilon}_{t+k})}{\mathbf{D}\{\widehat{\varepsilon}_t\}}$ (см. разд. 17.7) следующего вида:

$$\widehat{r}_{\widehat{\varepsilon}}(k) = \frac{\sum\limits_{t=1}^{n-k} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t+k}}{\sum\limits_{t=1}^{n} (\widehat{\varepsilon}_t)^2}, \quad k = 1, 2, \ldots. \tag{18.20}$$

Оценка $\widehat{r}_{\widehat{\varepsilon}}(k)$ называется k-й *выборочной автокорреляцией* СП $\{\widehat{\varepsilon}_t\}$. Если $\{\widehat{\varepsilon}_t\}$ — стационарный центрированный белый шум, а длина ВР $n \gg 1$, то можно показать [22], что с высокой степенью точности

$$\widehat{r}_{\widehat{\varepsilon}}(k) \sim \mathcal{N}\left(-\frac{1}{n}; \frac{1}{n}\right), \quad k = 1, 2, \ldots. \tag{18.21}$$

Таким образом, проверку гипотезы H_0: $r_{\widehat{\varepsilon}}(k) = 0$ о том, что $\{\widehat{\varepsilon}_t\}$ является стационарным белым шумом, можно осуществить следующим образом:

1) выбрать уровень значимости $\alpha \in [0{,}001; 0{,}1]$;

2) выбрать целое число K из интервала $\left[1; \dfrac{n}{2}\right]$;

3) вычислить выборочные автокорреляции $\widehat{r}_{\widehat{\varepsilon}}(k)$ для всех $k = 1, \ldots, K$ по формуле (18.20);

4) если для всех $1 \leqslant k \leqslant K$ справедливы неравенства

$$-\frac{1}{n} - \frac{u_\gamma}{\sqrt{n}} \leqslant \widehat{r}_{\widehat{\varepsilon}}(k) \leqslant -\frac{1}{n} + \frac{u_\gamma}{\sqrt{n}}, \tag{18.22}$$

где u_γ — квантиль уровня $\gamma = 1 - \dfrac{\alpha}{2}$ распределения $\mathcal{N}(0; 1)$, то гипотеза H_0 принимается на уровне значимости α.

На практике обычно число K выбирают не слишком большим ($K \leqslant 10$).

Если для $k = 1$ неравенство (18.22) нарушается, то это указывает на сильную корреляцию соседних элементов случайной компоненты $\{\varepsilon_t\}$. В этом случае для более адекватного описания случайной компоненты можно использовать АР(1)-модель:

$$\varepsilon_t = a\varepsilon_{t-1} + e_t, \tag{18.23}$$

где $|a| < 1$, а $\{e_t\}$ — центрированный стационарный белый шум, $\mathbf{D}\{e_t\} = \sigma_e^2 > 0$.

Пусть параметр a известен, а модель ВР имеет вид

$$X_t = h_t^\top \theta + \varepsilon_t, \quad t = 1, \ldots, n. \tag{18.24}$$

Покажем, что в этом случае (18.24) можно преобразовать так, чтобы случайная компонента стала белошумной. Действительно,

$$aX_{t-1} = \left(ah_{t-1}\right)^\top \theta + a\varepsilon_{t-1}. \tag{18.25}$$

Вычитая (18.25) из (18.24), получаем

$$X_t - aX_{t-1} = \left(h_t - ah_{t-1}\right)^\top \theta + \varepsilon_t - a\varepsilon_{t-1}. \tag{18.26}$$

Обозначим $\widetilde{X}_t = X_t - aX_{t-1}$, $\widetilde{h}_t = h_t - ah_{t-1}$ и заметим, что $\varepsilon_t - a\varepsilon_{t-1} = e_t$ в силу (18.23). Теперь

$$\widetilde{X}_t = \widetilde{h}_t^\top \theta + e_t, \quad t = 1, \ldots, n. \tag{18.27}$$

Видно, что (18.27) полностью совпадает с (18.3), поэтому

$$\widehat{\theta}_n = \widetilde{W}_n^{-1} \sum_{t=1}^{n} \widetilde{h}_t \widetilde{X}_t, \tag{18.28}$$

где $\widetilde{W}_n = \sum\limits_{t=1}^{n} \widetilde{h}_t \widetilde{h}_t^\top$. Теперь из (18.5) — (18.7) следует, что

$$\widehat{d}_t = h_t^\top \widehat{\theta}_n, \quad \mathbf{M}\left\{(\widehat{d}_t - d_t)^2\right\} = \sigma_e^2 h_t^\top \widetilde{W}_n^{-1} h_t.$$

Если параметр a в модели случайной компоненты (18.23) неизвестен, то в (18.26) вместо a следует использовать его оценку, построенную по остаткам $\{\widehat{\varepsilon}_t\}$. Соответствующий алгоритм содержит следующие шаги:

1) положите $a = 0$;

2) вычислите $\{\widetilde{X}_t, \widetilde{h}_t\}$ и найдите $\widehat{\theta}_n$ по формуле (18.28);

3) вычислите остатки $\left\{\widehat{\varepsilon}_t = X_t - h_t^\top \widehat{\theta}_n,\ t = 1, \ldots, n\right\}$;

4) найдите МНК-оценку $\widehat{a}_n$ параметра a, используя модель (18.23), в которой ε_t заменены на $\widehat{\varepsilon}_t$:

$$\widehat{a}_n = \frac{\sum\limits_{t=1}^{n} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-1}}{\sum\limits_{t=1}^{n} \widehat{\varepsilon}_{t-1}^2}; \tag{18.29}$$

5) найдите выборочную оценку $\widehat{\sigma}_e^2$ дисперсии σ_e^2 СП $\{e_t\}$:

$$\widehat{\sigma}_e^2 = \frac{1}{n} \sum_{t=1}^{n} \left(\widehat{\varepsilon}_t - \widehat{a}_n \widehat{\varepsilon}_{t-1}\right)^2; \tag{18.30}$$

6) повторно выполните п. 2), заменив a на $\widehat{a}_n$.

Сделаем несколько замечаний по поводу практического использования описанного алгоритма:

а) шаги 2) — 5) можно итерационно повторять до достижения сходимости процесса итераций (например, перестает изменяться от итерации к итерации оценка $\widehat{\theta}_n$);

б) оценка $\widehat{a}_n$ в (18.29) действительно является МНК-оценкой, так как

$$\widehat{a}_n = \arg\min_{a} \sum_{t=1}^{n} \left(\widehat{\varepsilon}_t - a\widehat{\varepsilon}_{t-1}\right)^2;$$

в) во всех расчетах полагаем, что X_t, $\widehat{\varepsilon}_t$, h_t равны нулю, если t не принадлежит множеству индексов $\{1, 2, \ldots, n\}$ (например, если $t = 0$);

г) с.к.-погрешность оценки $\widehat{\varphi}_t = h_t^\top \widehat{\theta}_n$ тренда φ_t вычисляется по формуле

$$\mathbf{M}\Big\{(\widehat{\varphi}_t - \varphi_t)^2\Big\} = \widehat{\sigma}_e^2 h_t^\top \widetilde{W}_t^{-1} h_t;$$

д) алгоритм очень просто обобщается на случай, когда ε_t описывается АР-моделью произвольного порядка $p \geqslant 1$. В этом случае

$$\varepsilon_t = \sum_{k=1}^{p} a_k \varepsilon_{t-k} + e_t,$$

поэтому

$$\widetilde{X}_t = X_t - \sum_{k=1}^{p} a_k X_{t-k}, \quad \widetilde{h}_t = h_t - \sum_{k=1}^{p} a_k h_{t-k}.$$

Оценки $\{\widehat{a}_k,\ k = 1, \ldots, p\}$ вычисляется методом Юла—Уолкера (см. разд. 19.1). Заметим, что для случая $p = 1$ метод Юла—Уолкера приводит к оценке (18.29).

19. Стационарные временны́е ряды

Предположим, что детерминированная компонента d_t временно́го ряда X_t отсутствует (или уже удалена одним из описанных ранее способов). В этом случае

$$X_t = \varepsilon_t, \quad t = 1, 2, \ldots,$$

где $\{\varepsilon_t\}$ — случайная компонента временно́го ряда. Мы будем далее предполагать, что случайная последовательность $\{\varepsilon_t,\ t = 1, 2, \ldots\}$ является стационарной.

19.1. Оценивание параметров АР-модели

Предположим, что $\{\varepsilon_t\}$ описывается АР(p)-моделью

$$\varepsilon_t = a_1\varepsilon_{t-1} + \cdots + a_p\varepsilon_{t-p} + e_t, \quad t = 1, 2, \ldots, n, \tag{19.1}$$

где $\{a_1, \ldots, a_p\}$ — неизвестные параметры, $\{e_t\}$ — стационарный белый шум, $\mathbf{M}\{e_t\} = 0$, $\mathbf{D}\{e_t\} = \sigma_e^2 > 0$, $t = 1, 2, \ldots, n$. Предположим, что порядок модели $p \geqslant 1$ нам известен. В этом случае рассматривается задача оценивания вектора неизвестных параметров $\theta = \left[a_1, \ldots, a_p\right]^\top$ модели (19.1) и дисперсии σ_e^2 белого шума $\{e_t\}$ по наблюдениям $\{\varepsilon_t,\ t = 1, \ldots, n\}$.

Для нахождения оценок вектора параметров θ в модели (19.1) найдем сначала их связь с корреляционной функцией $r_\varepsilon(k)$ СП $\{\varepsilon_t\}$. Напомним, что

$$r_\varepsilon(k) = \frac{\mathbf{M}\{\varepsilon_t\varepsilon_{t+k}\}}{\sigma_\varepsilon^2}, \tag{19.2}$$

где $\sigma_\varepsilon^2 = \mathbf{D}\{\varepsilon_t\}$, причем σ_ε^2 не зависит от t в силу стационарности СП $\{\varepsilon_t\}$. Умножим обе части уравнения (19.1) на ε_{t-1}:

$$\varepsilon_t\varepsilon_{t-1} = a_1\varepsilon_{t-1}^2 + a_2\varepsilon_{t-2}\varepsilon_{t-1} + \cdots + a_p\varepsilon_{t-p}\,\varepsilon_{t-1}. \tag{19.3}$$

Применяя к обеим частям (19.3) операцию вычисления математического ожидания и деля обе части на σ_ε^2, получаем следующее уравнение:

$$r_1 = a_1 + a_2 r_1 + \cdots + a_p r_{p-1},$$

где для краткости обозначено $r_k = r_\varepsilon(k)$.

Умножая (19.1) последовательно на $\varepsilon_{t-2}, \varepsilon_{t-3}, \ldots, \varepsilon_{t-p}$ и проделывая операции, описанные выше, получаем p уравнений, связывающих $\{a_1, \ldots, a_p\}$ с $\{r_1, \ldots, r_p\}$:

$$\begin{cases} a_1 + r_1 a_2 + \cdots + r_{p-1}a_p & = & r_1 \\ r_1 a_1 + a_2 + \cdots + r_{p-2}a_p & = & r_2 \\ \vdots & & \vdots \\ r_{p-1}a_1 + r_{p-2}a_2 + \cdots + a_p & = & r_p. \end{cases} \tag{19.4}$$

Полученная система уравнений называется *системой Юла—Уолкера* [3, 22].

Неизвестные значения $\{r_1, \ldots, r_p\}$ корреляционной функции $r_\varepsilon(k)$ в (19.4) заменим их выборочными оценками, построенными по наблюдениям $\{\varepsilon_t, t = 1, \ldots, n\}$:

$$\widehat{r}_k = \frac{\sum\limits_{t=1}^{n-k} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t+k}}{\sum\limits_{t=1}^{n} (\widehat{\varepsilon}_t)^2}, \quad k = 1, 2, \ldots, p. \tag{19.5}$$

Получим систему уравнений

$$\begin{cases} a_1 + \widehat{r}_1 a_2 + \cdots + \widehat{r}_{p-1} a_p & = & \widehat{r}_1 \\ \vdots & & \vdots \\ \widehat{r}_{p-1} a_1 + \widehat{r}_{p-2} a_2 + \cdots + a_p & = & \widehat{r}_p. \end{cases} \tag{19.6}$$

Решая (19.6) относительно неизвестных $\{a_1, \ldots, a_p\}$, находим оценку $\widehat{\theta}_n = \left[\widehat{a}_1, \ldots, \widehat{a}_p\right]^\top$ вектора θ неизвестных параметров АР(p)-модели (19.1).

В практических задачах наиболее часто используются модели АР(1) и АР(2). Из общих соотношений (19.6) для случая АР(1) имеем

$$\widehat{a}_1 = \widehat{r}_1 = \frac{\sum\limits_{t=1}^{n-1} \varepsilon_t \varepsilon_{t+1}}{\sum\limits_{t=1}^{n} (\varepsilon_t)^2}. \tag{19.7}$$

Для случая АР(2) из (19.6) следует, что

$$\widehat{a}_1 = \frac{\widehat{r}_1 - \widehat{r}_1 \widehat{r}_2}{1 - \widehat{r}_1^2}, \quad \widehat{a}_2 = \frac{\widehat{r}_2 - \widehat{r}_1^2}{1 - \widehat{r}_1^2}. \tag{19.8}$$

Для случая $p \geqslant 3$ аналитические соотношения, аналогичные (19.7), (19.8), для оценок $\{\widehat{a}_1, \ldots, \widehat{a}_p\}$ выписать достаточно сложно, поэтому их определяют, решая численно систему уравнений (19.6) соответствующего порядка. Так как система уравнений (19.6) линейна по параметрам $\{a_1, \ldots, a_p\}$, то ее численное решение не вызывает никаких затруднений.

19.2. Оценивание параметров АР(p)-модели методом наименьших квадратов

Введем обозначение $h_t = \left[\varepsilon_{t-1}, \ldots, \varepsilon_{t-p}\right]^\top$. Тогда АР($p$)-модель (19.1) принимает вид, схожий с моделью линейной регрессии порядка p:

$$\varepsilon_t = h_t^\top \theta + e_t, \quad t = 1, 2, \ldots, n. \tag{19.9}$$

В соответствии с общими принципами МНК приходим к следующей задаче оптимизации:

$$L(\theta) = \sum_{t=1}^{n} \left(\varepsilon_t - h_t^\top \theta\right)^2 \to \min_{\theta}. \tag{19.10}$$

Решение задачи (19.10) нам уже известно:

$$\widehat{\theta}_n = \left(\sum_{t=1}^{n} h_t h_t^\top\right)^{-1} \sum_{t=1}^{n} h_t \varepsilon_t. \tag{19.11}$$

МНК-оценка $\widehat{\theta}_n$ не отличается от оценки, полученной методом Юла—Уолкера.

Замечание. Модель (19.9) только по структуре схожа с моделью линейной регрессии, так как вектор h_t является случайным. Поэтому свойства МНК-оценки, изучавшиеся в разделе 10, нельзя перенести автоматически на МНК-оценку (19.11) параметров АР(p)-модели.

Замечание. Оценка (19.11) выражается не только через $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\}$, но также через $\{\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-p+1}\}$, которые, вообще говоря, неизвестны. В практических расчетах всегда полагают $\varepsilon_k = 0$ для всех $k \leqslant 0$. Это предположение не оказывает какого-либо серьезного влияния на оценки, если АР(p)-модель (19.1) асимптотически устойчива (см. (17.18)).

Важным свойством МНК-оценок (19.11) является то, что их можно вычислить *рекуррентно*, т.е. по мере поступления новых результатов наблюдений.

Пусть $\widehat{\theta}_{n-1}$ — МНК-оценка вектора θ, построенная по наблюдениям $\{\varepsilon_1, \ldots, \varepsilon_{n-1}\}$. Тогда МНК-оценку $\widehat{\theta}_n$, построенную с учетом наблюдения ε_n, можно вычислить, «подправив» уже известную оценку $\widehat{\theta}_{n-1}$. Соответствующие уравнения называются уравнениями *рекуррентного МНК* и имеют следующий вид:

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} + K_n \left(\varepsilon_n - h_t^\top \widehat{\theta}_{n-1}\right), \quad n = 1, 2, \ldots, \tag{19.12}$$

$$\begin{cases} K_n = \dfrac{P_{n-1} h_n}{1 + h_n^\top P_{n-1} h_n}, \\ P_n = \left(I - K_n h_n^\top\right) P_{n-1}. \end{cases} \tag{19.13}$$

Формула (19.12) описывает процедуру корректировки оценки $\widehat{\theta}_{n-1}$, а формулы (19.13) описывают рекуррентную процедуру вычисления вспомогательного вектора K_n, называемого векторным *коэффициентом усиления*, и вспомогательной матрицы P_n.

Для того чтобы начать процедуру рекуррентного вычисления оценок по формулам (19.12) и (19.13), следует задать начальные значения для $n = 0$:

$$\widehat{\theta}_0 = 0, \quad P_0 = \alpha I, \tag{19.14}$$

где $\alpha = 10^5 \div 10^8$ (достаточно большое число), а I — единичная матрица размера $(p \times p)$.

После того, как найдены оценки $\widehat{\theta}_n$ параметров АР(p)-модели, можно найти оценку σ_e^2 дисперсии возмущающего процесса $\{e_t\}$. Оценим e_t по формуле

$$\widehat{e}_t = \varepsilon_t - h_t^\top \widehat{\theta}_n, \quad t = 1, 2, \ldots, n, \tag{19.15}$$

а дисперсию σ_e^2 по формуле

$$\widehat{\sigma}_e^2 = \frac{1}{n} \sum_{t=1}^{n} (\widehat{e}_t)^2 . \tag{19.16}$$

З а м е ч а н и е. Матрица P_n, вычисляемая в (19.13), характеризует точность оценки $\widehat{\theta}_n$. А именно, для достаточно большого объема выборки n

$$\sqrt{n} \left(\widehat{\theta}_n - \theta \right) \sim \mathcal{N} \left(0; \sigma_e^2 P_n \right) . \tag{19.17}$$

19.3. Подбор порядка АР-модели

Пусть $\{\varepsilon_t\}$ описывается некоторой АР-моделью, истинный порядок p_0 которой неизвестен. Для определения p_0 нам понадобится понятие *частной автокорреляционной функции* (*ЧАКФ*).

Рассмотрим последовательность решений

$$\theta(p) = \{a_{1p}, a_{2p}, \ldots, a_{pp}\}^\top$$

систем уравнений (19.4) нарастающего порядка $p = 1, 2, \ldots$, где $\{r_1, r_2, \ldots\}$ — значения корреляционной функции СП $\{\varepsilon_t\}$, т.е. $r_k = r_\varepsilon(k)$, $k = 1, 2, \ldots$.

О п р е д е л е н и е 19.1. *Частной автокорреляционной функцией* (*ЧАКФ*) СП $\{\varepsilon_t\}$ называется последовательность $\{\psi(p), p = 1, 2, \ldots\}$, определенная соотношениями

$$\psi(p) = a_{pp}, \quad p = 1, 2, \ldots . \tag{19.18}$$

ЧАКФ $\psi(p)$ обладает следующими свойствами:

1) $|\psi(p)| < 1$ для любого $p = 1, 2, \ldots$;

2) $\psi(p) \to 0$, если $p \to \infty$;

3) если $\{\varepsilon_t\}$ описывается АР(p_0)-моделью, то $\psi(p) = 0$ для всех $p \geqslant p_0 + 1$.

Если во всех системах (19.4) для $p = 1, 2, \ldots$ истинные корреляции $\{r_k\}$ заменить их выборочными оценками $\{\widehat{r}_k\}$, построенными по (19.5), то мы получаем последовательность $\widehat{\theta}(p) = \left[\widehat{a}_{1p}, \ldots, \widehat{a}_{pp}\right]^\top$ соответствующих решений, $p = 1, 2, \ldots$. Из (19.18) тогда следует, что

$$\widehat{\psi}(p) = \widehat{a}_{pp}, \quad p = 1, 2, \ldots \tag{19.19}$$

будет оценкой ЧАКФ, которая называется *выборочной ЧАКФ*. Можно показать, что $\widehat{\psi}(p)$ обладает следующим важным статистическим свойством: если $n \gg 1$, то $\widehat{\psi}(p) \sim \mathcal{N}\left(0; \dfrac{1}{n}\right)$ для любого $p \geqslant p_0 + 1$. Указанное свойство позволяет сформулировать простое *правило подбора порядка модели АР(p_0)*: в качестве оценки для p_0 выбираем такое целое $\widehat{p}_0$, что

$$|\widehat{\varphi}(p)| \leqslant \frac{u_{0,975}}{\sqrt{n}} \approx \frac{2}{\sqrt{n}} \text{ для всех } p \geqslant \widehat{p}_0 + 1. \tag{19.20}$$

Заметим, что (19.20) выполняется с вероятностью 0,95 для любого $p \geqslant p_0 + 1$.

19.4. Модель авторегрессии и скользящего среднего

Если при построении модели временного ряда в виде АР(p) выясняется, что порядок p_0 достаточно велик (например, $p_0 > 4$), то это может указывать на то, что возмущающий процесс $\{e_t\}$ не является белым шумом. В этом случае обычно делается попытка описать возмущение в модели (19.1) в виде процесса СС(q), введенного в разд. 17.7. Результирующая модель называется АРСС(p, q)-моделью и имеет вид

$$\varepsilon_t = a_1\varepsilon_{t-1} + \cdots + a_p\varepsilon_{t-p} + e_t + b_1 e_{t-1} + \cdots + b_q e_{t-q}, \tag{19.21}$$

где $\{e_t\}$ — стационарный белый шум, $\mathbf{M}\{e_t\} = 0$, $\mathbf{D}\{e_t\} = \sigma_e^2 > 0$.

Сразу заметим, что для оценивания параметров $\{b_1, \ldots, b_q\}$ СС(q)-модели нет общего алгоритма, аналогичного методу Юла—Уолкера. Даже для случая $q = 2$ оценивание параметров $\{b_1, b_2\}$ требует решения весьма сложной системы нелинейных алгебраических уравнений. Последней проблемы, тем не менее, можно избежать, если воспользоваться методом наименьших квадратов в его рекуррентном варианте (19.12) и (19.13). Действительно, если ввести обозначения

$$\begin{gathered} \theta = \left[a_1, \ldots, a_p, b_1, \ldots, b_q\right]^\top, \\ h_t = \left[\varepsilon_{t-1}, \ldots, \varepsilon_{t-p}, e_{t-1}, \ldots, e_{t-q}\right]^\top, \end{gathered} \tag{19.22}$$

то АРСС(p,q)-модель (19.21) принимает вид, сходный с линейной регрессионной моделью порядка $(p+q)$:

$$\varepsilon_t = h_t^\top \theta + e_t, \quad t = 1, 2, \ldots, n. \tag{19.23}$$

Отличие модели (19.23) от (19.9), построенной для АР(p)-модели, состоит в том, что вектор h_t нам известен лишь частично. Действительно, первые p компонент (см. (19.22)) наблюдаются, а компоненты с $(p+1)$-й по $(p+q)$-ю — неизвестны, так как СП $\{e_t\}$ ненаблюдаема. Поэтому $\{e_{t-1}, \ldots, e_{t-q}\}$ в (19.22) следует заменить на некоторые их оценки $\{\widehat{e}_{t-1}, \ldots, \widehat{e}_{t-q}\}$, а вектор h_t заменить на $\widehat{h}_t = \left[\varepsilon_{t-1}, \ldots, \varepsilon_{t-p}, \widehat{e}_{t-1}, \ldots, \widehat{e}_{t-q}\right]$. Оценки $\widehat{e}_t$ вычисляются последовательно по формуле

$$\widehat{e}_t = \varepsilon_t - \widehat{h}_t^\top \widehat{\theta}_t, \quad t = 1, 2, \ldots, n, \tag{19.24}$$

где $\widehat{\theta}_t$ — текущая оценка вектора θ, построенная по наблюдениям $[\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_t]^\top$. Таким образом, окончательный вид алгоритма оценивания описывают следующие соотношения (см. (19.12) и (19.13)):

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} + K_n \left(\varepsilon_n - \widehat{h}_n^\top \widehat{\theta}_{n-1}\right), \quad n = 1, 2, \ldots, \tag{19.25}$$

$$\begin{cases} K_n & = & \dfrac{P_{n-1}\widehat{h}_n}{1 + \widehat{h}_n^\top P_{n-1}\widehat{h}_n}, \\ P_n & = & \left(I - K_n \widehat{h}_n^\top\right) P_{n-1}, \end{cases} \tag{19.26}$$

$$\widehat{e}_n = \varepsilon_n - \widehat{h}_n^\top \widehat{\theta}_n. \tag{19.27}$$

Формула (19.25) описывает рекуррентную процедуру оценивания вектора θ, формулы (19.26) — вычисление вспомогательных параметров алгоритма, а (19.27) — оценку новой компоненты e_n, которая будет использована на следующем шаге алгоритма:

$$\widehat{h}_{n+1} = \left[\varepsilon_n, \varepsilon_{n-1}, \ldots, \varepsilon_{n-p+1}, \widehat{e}_n, \ldots, \widehat{e}_{t-q+1}\right]^\top.$$

Заметим, что оценки $\{\widehat{e}_{n-1}, \ldots, \widehat{e}_{t-q+1}\}$ были компонентами вектора $\widehat{h}_n$ и, следовательно, уже известны.

Начальные условия $\widehat{\theta}_0$ и P_0 для алгоритма (19.25) — (19.27) определяются соотношениями (19.14).

В заключение отметим, что при моделировании реальных эконометрических рядов данных обычно удается ограничиться моделями АРСС(4,2) и более простыми для вполне приемлемого по точности описания стационарных рядов данных и их прогнозирования.

19.5. Примеры

Пример 19.1. Числовые данные о $n = 50$ значениях случайной компоненты временного ряда $\{\varepsilon_t,\ t = 1, \ldots, 50\}$ приведены в табл. 19.1.

Таблица 19.1

t	ε_t	t	ε_t	t	ε_t	t	ε_t	t	ε_t
1	0,103	11	0,977	21	1,324	31	−0,589	41	−0,271
2	−0,207	12	0,868	22	2,454	32	−1,048	42	−0,908
3	−0,383	13	0,636	23	2,568	33	−1,129	43	−1,703
4	0,100	14	0,253	24	2,705	34	−0,726	44	−2,269
5	−0,332	15	0,411	25	2,097	35	−0,149	45	−2,642
6	−0,687	16	0,240	26	1,689	36	0,638	46	−3,313
7	−0,537	17	0,236	27	0,905	37	0,306	47	−3,620
8	0,148	18	0,904	28	0,462	38	0,070	48	−3,698
9	0,329	19	0,860	29	0,004	39	0,289	49	−3,508
10	0,550	20	0,917	30	−0,355	40	0,048	50	−3,357

Требуется подобрать модель компоненты ε_t в виде уравнения авторегрессии подходящего порядка.

Решение. По определению модель авторегрессии порядка $p \geqslant 1$ (АР(p)-модель) имеет вид

$$\varepsilon_t = a_1\varepsilon_{t-1} + \ldots + a_p\varepsilon_{t-p} + e_t, \quad t = 1, 2, \ldots, n, \tag{19.28}$$

где $\{e_t\}$ — последовательность центрированных некоррелированных СВ с одинаковой дисперсией $D\{\varepsilon_t\} = \sigma_e^2 > 0$, а $\theta = \left[a_1, \ldots, a_p\right]^\top$ — вектор неизвестных (оцениваемых) параметров модели. Обозначая $h_t = \left[\varepsilon_{t-1}, \ldots, \varepsilon_{t-p}\right]^\top$, представим (19.28) в виде линейной регрессии:

$$\varepsilon_t = h_t^\top\theta + e_t, \quad t = 1, 2, \ldots, n. \tag{19.29}$$

Применяя к (19.29) метод наименьших квадратов, находим оценку $\widehat{\theta}$ вектора θ, решая систему нормальных уравнений:

$$W_n\theta = w_n, \tag{19.30}$$

где $W_n = \{w_n(i,j)\}$, $i,j = 1, \ldots, n$, причем $w_n(i,j) = \sum\limits_{t=1}^{n} \varepsilon_{t-i}\varepsilon_{t-j}$; вектор $w_n = \{w_n(1), \ldots, w_n(p)\}^\top$ имеет компоненты $w_n(i) = \sum\limits_{t=1}^{n} \varepsilon_t\varepsilon_{t-i}$, $i = 1, \ldots, n$. Если объем выборки достаточно велик, то можно считать, что

$$\mathbf{M}\left\{\widehat{\theta}\right\} = \theta, \text{ а } \widehat{K}_\theta = \mathbf{cov}\{\widehat{\theta} - \theta, \widehat{\theta} - \theta\} = \sigma_\varepsilon^2 W_n^{-1}. \tag{19.31}$$

1. Предположение о том, что $p = 0$, т.е. $\varepsilon_t = e_t$ — белый шум совершенно не согласуется с имеющимися данными, приведенными на рис. 19.1.

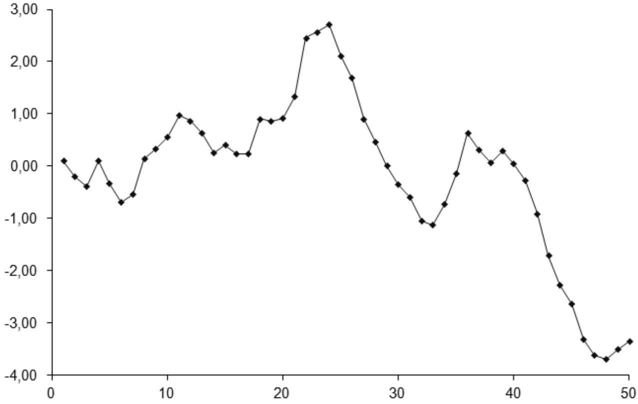


Рис. 19.1. График случайной компоненты ВР

Визуальный анализ показывает наличие существенной положительной корреляции между различными компонентами последовательности $\{\varepsilon_t\}$. Поэтому предположение $p = 0$ следует сразу отвергнуть.

2. Положим $p = 1$ и рассмотрим АР(1)-модель

$$\varepsilon_t = a\varepsilon_{t-1} + e_t, \quad t = 1, \ldots, n.$$

Из (19.30) следует, что

$$\widehat{a} = \frac{\sum\limits_{t=1}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum\limits_{t=1}^{n} \varepsilon_t^2} = \frac{107{,}36}{117{,}61} = 0{,}91.$$

Заметим, что здесь и далее мы всегда будем полагать, что $\varepsilon_{t-j} = 0$, если $j \geqslant t$, $t = 1, \ldots, n$.

Подобранная АР(1)-модель имеет вид

$$\varepsilon_t = 0{,}91\varepsilon_{t-1} + \widetilde{e}_t, \quad t = 1, \ldots, n,$$

где $\{\widetilde{e}_t\}$ — остатки АР(1)-модели. Если предположение $p = 1$ верно, то $\{\widetilde{e}_t\}$ — белый шум.

Для проверки последнего утверждения вычислим реализацию первой выборочной автокорреляции:

$$\widetilde{r}_1 = \frac{\sum\limits_{t=1}^{n} \widetilde{e}_t \widetilde{e}_{t-1}}{\sum\limits_{t=1}^{n} \widetilde{e}_t^2} = 0{,}583. \tag{19.32}$$

Если предположение о том, что $p = 1$ верно, то

$$\widetilde{r}_1 \in I = \left[-\frac{1}{n} - \frac{2}{\sqrt{n}}; -\frac{1}{n} + \frac{2}{\sqrt{n}}\right] = [-0{,}303; 0{,}263]$$

с вероятностью 0,95. Но из (19.32) следует, что $\widetilde{r}_1 \notin I$, поэтому гипотезу о том, что данные описываются АР(1)-моделью, следует отвергнуть.

Визуальный анализ графика остатков $\{\widetilde{\varepsilon}_t\}$, приведенный на рис. 19.2, также указывает на существенную коррелированность.

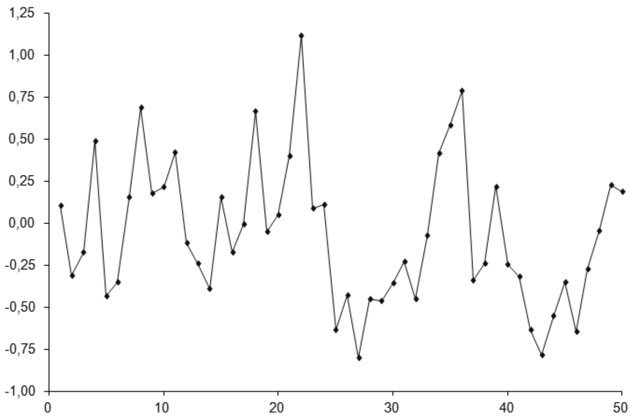


Рис. 19.2. Остатки в модели АР(1) случайной компоненты ВР

3. Рассмотрим теперь АР(2)-модель

$$\varepsilon_t = a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2} + e_t, \quad t = 1, \ldots, n.$$

В этом случае

$$W_n = \begin{bmatrix} \sum\limits_{t=1}^{n} e_{t-1}^2 & \sum\limits_{t=1}^{n} e_{t-1} e_{t-2} \\ \sum\limits_{t=1}^{n} e_{t-1} e_{t-2} & \sum\limits_{t=1}^{n} e_{t-2}^2 \end{bmatrix} = \begin{bmatrix} 106{,}32 & 95{,}58 \\ 95{,}58 & 94{,}01 \end{bmatrix},$$

$$w_n = \left[\sum_{t=1}^{n} e_t e_{t-1};\ \sum_{t=1}^{n} e_t e_{t-2}\right]^\top = [107{,}36; 92{,}128]^\top.$$

Теперь из (19.30) следует

$$\begin{cases} 106{,}32a_1 + 95{,}58a_2 = 107{,}36 \\ 95{,}58a_1 + 94{,}01a_2 = 92{,}128. \end{cases} \tag{19.33}$$

Решая систему (19.33), находим реализации оценок параметров a_1 и a_2:

$$\widehat{a}_1 = 1{,}499; \quad \widehat{a}_2 = -0{,}544.$$

Остатки $\{\widehat{e}_t\}$ в подобранной АР(2)-модели вычисляются по формуле

$$\widehat{e}_t = \varepsilon_t - 1{,}499\varepsilon_{t-1} + 0{,}544\varepsilon_{t-2}, \quad t = 1, \ldots, n. \tag{19.34}$$

График остатков $\{\widehat{e}_t\}$ представлен на рис. 19.3.

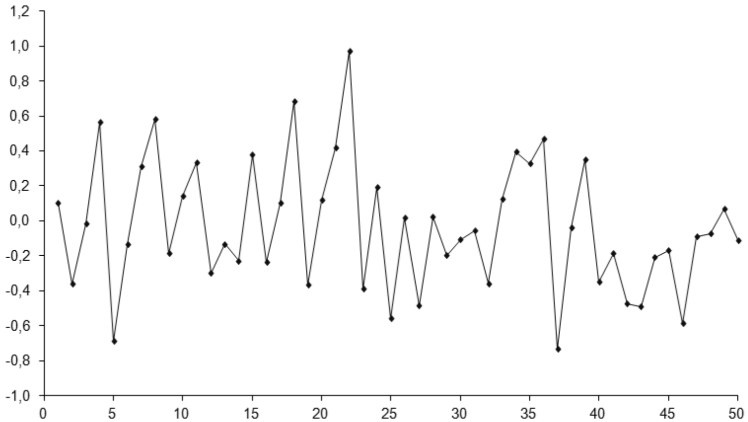


Рис. 19.3. Остатки в модели АР(2) случайной компоненты ВР

Из рис. 19.3 видно, что последовательность $\{\widehat{e}_t\}$ похожа на реализацию центрированного белого шума.

Для проверки последнего предположения вычислим $\widehat{r}_1$:

$$\widehat{r}_1 = \frac{\sum\limits_{t=1}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum\limits_{t=1}^{n} \widehat{e}_t^2} = -0{,}048.$$

Очевидно, что $\widehat{r}_1 \in I$. Оказывается также, что

$$\widehat{r}_k = \frac{\sum\limits_{t=1}^{n} \widehat{e}_t \widehat{e}_{t-k}}{\sum\limits_{t=1}^{n} \widehat{e}_t^2} \in I \text{ для } k = 2, \ldots, 10.$$

Все это позволяет сделать вывод о том, что предположение $p = 2$ согласуется с имеющимися данными. Теперь можно найти оценку σ_e — с.к.о. возмущающего шума $\{e_t\}$ в модели АР(2):

$$\widehat{\sigma}_e = \left[\frac{1}{n}\sum_{t=1}^{n} \widehat{e}_t^2\right]^{\frac{1}{2}} = 0{,}37.$$

Теперь в дальнейших вычислениях можно использовать подобранную АР(2)-модель случайной компоненты

$$\varepsilon_t = 1{,}499\varepsilon_{t-1} - 0{,}544\varepsilon_{t-2} + \widehat{e}_t, \quad t = 1, \ldots, n,$$

где $\{\widehat{e}_t\}$ — центрированный белый шум со средним квадратическим отклонением $\widehat{\sigma}_e = 0{,}37$. ■

Пример 19.2. В табл. 19.2 приведены статистические данные о месячных производствах молока в РФ с января 1992 г. по декабрь 1995 г., где t — номер месяца, X_t — соответствующий объем производства.

Таблица 19.2

t	X_t	t	X_t	t	X_t	t	X_t
1	2015	13	1759	25	1510	37	1172
2	2123	14	1773	26	1484	38	1226
3	2624	15	2361	27	1988	39	1651
4	2891	16	2649	28	2211	40	1859
5	3335	17	3203	29	2559	41	2392
6	4071	18	3936	30	3209	42	2864
7	4040	19	3861	31	3204	43	2714
8	3392	20	3321	32	2687	44	2420
9	2467	21	2438	33	2031	45	1925
10	2092	22	1760	34	1506	46	1338
11	1494	23	1299	35	1050	47	984
12	1562	24	1345	36	1054	48	1020

Требуется построить математическую модель временно́го ряда $\{X_t, t = 1, \ldots, n\}$, где $n = 48$, и построить с ее помощью прогноз производства молока на первые 10 месяцев 1996 г.

Р е ш е н и е. Представим данные, приведенные в табл. 19.2, в графическом виде (рис. 19.4).

Визуальный анализ статистических данных показывает, что переменная X_t имеет тенденцию к уменьшению, причем присутствуют как линейный тренд, так и значительная сезонная компонента.

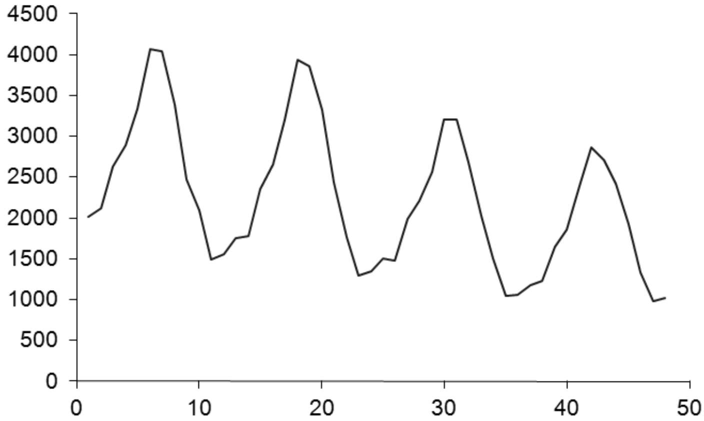


Рис. 19.4. Ежемесячное производство молока в РФ с января 1992 г. по декабрь 1995 г. (в тыс. т)

Видно также, что амплитуда сезонных колебаний имеет тенденцию к уменьшению. Все это позволяет предложить следующую модель временного ряда $\{X_t\}$:

$$X_t = \varphi_t s_t + \varepsilon_t, \quad t = 1, \ldots, n, \tag{19.35}$$

где

$$\varphi_t = a_1 + a_2 t \tag{19.36}$$

является линейным трендом.

Найдем оценки $\widehat{a}_1$, $\widehat{a}_2$ параметров модели (19.36) методом наименьших квадратов, используя данные табл. 19.2. Система нормальных уравнений имеет вид

$$\begin{cases} na_1 + \sum\limits_{t=1}^{n} ta_2 = \sum\limits_{t=1}^{n} X_t \\ \sum\limits_{t=1}^{n} ta_1 + \sum\limits_{t=1}^{n} t^2 a_2 = \sum\limits_{t=1}^{n} tX_t. \end{cases} \tag{19.37}$$

Подставляя в (19.37) числовые данные, находим

$$\begin{cases} 48a_1 + 1176a_2 = 107\,869 \\ 1176a_1 + 38024a_2 = 2\,397\,408. \end{cases} \tag{19.38}$$

Решая (19.38), находим

$$\widehat{a}_1 = 2899{,}88; \quad \widehat{a}_2 = -26{,}64.$$

Отсюда следует, что оценка тренда φ_t принимает вид

$$\widehat{\varphi}_t = 2899{,}88 - 26{,}64t, \quad t = 1, \ldots, n.$$

Перейдем теперь к оценке сезонной компоненты. Вычислим «измерения» $\widetilde{s}_t$, $t = 1, \ldots, n$ сезонной компоненты, используя принятую модель (19.35) и оценку:

$$\widetilde{s}_t = \frac{X_t}{\widehat{\varphi}_t}. \quad t = 1, \ldots, n. \tag{19.39}$$

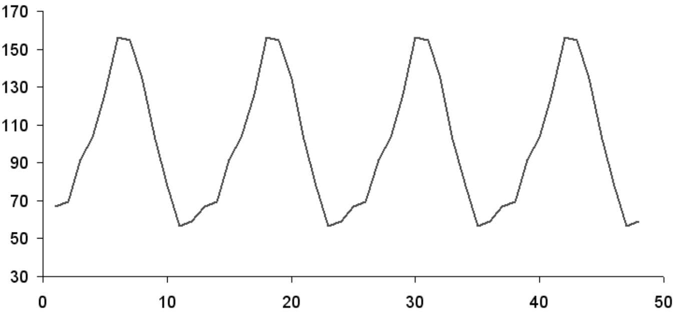


Рис. 19.5. Оценка сезонной компоненты временно́го ряда

Полученные результаты теперь можно использовать для вычисления сезонной компоненты s_t с периодом $r = 12$:

$$\widehat{s}_t = \frac{1}{4}\left(\widetilde{s}_t + \widetilde{s}_{t+12} + \widetilde{s}_{t+24} + \widetilde{s}_{t+36}\right), \; t = 1, \ldots, 12. \tag{19.40}$$

Результаты вычислений по формуле (19.40) приведены в табл 19.3.

Таблица 19.3

t	1	2	3	4	5	6
$\widehat{s}_t$	67,0	69,2	91,8	103,6	126,0	156,2
t	7	8	9	10	11	12
$\widehat{s}_t$	154,8	134,5	102,6	77,7	56,8	59,3

Всего у нас четыре сезона.

На рис. 19.5 приведен график сезонного индекса, равного $\widehat{S}_t = \widehat{s}_t \cdot 100\%$, $t = 1, \ldots, 12$ и $\widehat{S}_{[t]} = \widehat{s}_{[t]}$, где $[t]$ — целый остаток от деления t на 12 для $t = 13, \ldots, n$.

Теперь можно перейти к анализу случайной компоненты $\{\varepsilon_t\}$ модели (19.35). Для этого вычислим и проанализируем остатки

$$\widehat{\varepsilon}_t = X_t - \widehat{\varphi}_t \cdot \frac{\widehat{S}_t}{100}, \quad t = 1, \ldots, n. \tag{19.41}$$

Выберем в качестве модели $\{\widehat{\varepsilon}_t\}$ АР(1)-модель:

$$\widehat{\varepsilon}_t = a\widehat{\varepsilon}_{t-1} + e_t, \quad t = 1, \ldots, n. \tag{19.42}$$

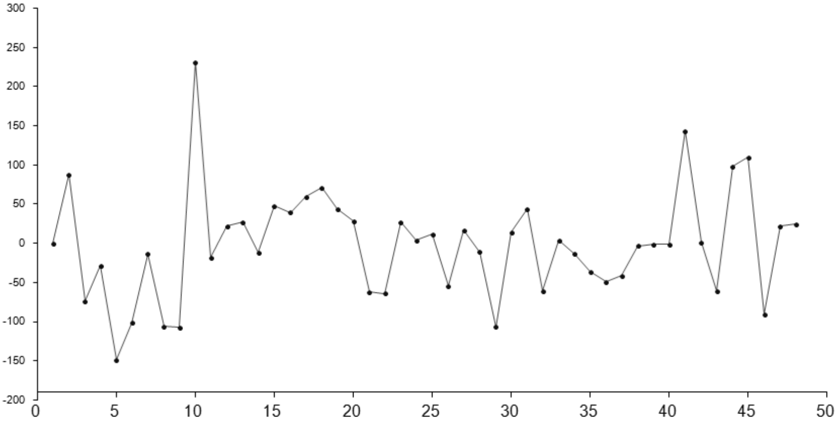


Рис. 19.6. Остатки модели АР(1) случайной компоненты ряда

Повторяя вычисления, описанные в примере 19.1 для АР(1)-модели, находим:

$$\widehat{a} = 0{,}708; \quad \widehat{\sigma}_e = 69{,}24. \tag{19.43}$$

Для проверки правильности выбора порядка $p = 1$ вычислим реализации первой выборочной автокорреляции $\widehat{r}_1$ и выборочной ЧАКФ $\widehat{\psi}(2)$:

$$\widehat{r}_1 = \widehat{\psi}(1) = 0{,}027; \quad \widehat{\psi}(2) = -0{,}009. \tag{19.44}$$

Из (19.44) следует, что

$$\widehat{r}_1 \in I = \left[-\frac{1}{n} - \frac{2}{\sqrt{n}}; -\frac{1}{n} + \frac{2}{\sqrt{n}}\right] = [-0{,}310; 0{,}268]\,,$$

$$\widehat{\psi}(2) \in \left[-\frac{2}{\sqrt{n}}; \frac{2}{\sqrt{n}}\right] = [-0{,}289; 0{,}289]\,.$$

Таким образом, остатки $\{\widehat{e}_t\}$ в модели (19.42), вычисляемые по формуле $\widehat{e}_t = \widehat{\varepsilon}_t - 0{,}708\widehat{\varepsilon}_{t-1}$, $t = 1, \ldots, n$, можно считать центрированным белым шумом со средним квадратическим отклонением $\widehat{\sigma}_e = 69{,}24$. График остатков $\{\widehat{\varepsilon}_t,\ t = 1, \ldots, n\}$ приведен на рис. 19.6.

Теперь перейдем к прогнозированию ряда X_t на 10 месяцев вперед, т.е. $t = n+1, \ldots, n+10$. Во-первых, для любого t прогноз $\widehat{d}_t$ детерминированной компоненты ряда $d_t = \varphi_t s_t$ имеет вид

$$\widehat{d}_t = (2899{,}88 - 26{,}64t) \cdot \frac{\widehat{S}_{[t]}}{100}, \quad t = n+1, \ldots, n+10. \tag{19.45}$$

Прогноз случайной компоненты вычисляется по рекуррентной формуле, следующей из (19.42) и (19.43):

$$\widehat{\varepsilon}_t = 0{,}708\widehat{\varepsilon}_{t-1}, \quad t \geqslant n+1, \tag{19.46}$$

где начальное значение $\widehat{\varepsilon}_{48} = 58{,}37$ было вычислено по формуле (19.41) для $t = n$.

Теперь прогноз $\widehat{X}_t$ производства молока можно вычислить по формуле

$$\widehat{X}_t = \widehat{d}_t + \widehat{\varepsilon}_t, \quad t = n+1, \ldots, n+10. \tag{19.47}$$

В табл. 19.4 приведены результаты прогнозирования $\widehat{X}_t$ истинного объема производства молока X_t и величины соответствующих относительных погрешностей прогнозирования, т.е.

$$\Delta_t = \frac{|X_t - \widehat{X}_t|}{X_t} \cdot 100\%, \quad t = n+1, \ldots, n+10. \tag{19.48}$$

Таблица 19.4

t	X_t	$\widehat{d}_t$	$\widehat{\varepsilon}_t$	$\widehat{X}_t$	Δ_t
49	1038	1068	41	1109	6,8
50	1104	1086	29	1115	1,0
51	1439	1416	21	1436	0,2
52	1521	1570	15	1585	4,2
53	1827	1875	10	1886	3,2
54	2446	2283	7	2290	6,4
55	2369	2222	5	2227	6,0
56	2081	1893	4	1897	8,8
57	1577	1418	3	1421	9,9
58	1081	1053	2	1055	2,4

Последняя колонка табл. 19.4 показывает, что модель позволяет построить достаточно точный прогноз объема производства молока практически на целый год вперед. ∎

Пример 19.3. В табл. 19.5 приведены поквартальные статистические данные японской экономики.

Таблица 19.5

t	$C(t)$	$YP(t)$	$YD(t)$	$YP(t)/YD(t)$
1	14566,9	8913,5	58,2	153,15
2	15071,4	10594,2	59,1	179,26
3	15970,3	11506,7	59,4	193,72
4	18029,9	15247,2	60,7	251,19
5	15472,5	10511,8	61,9	169,82
6	16038,4	11949,2	62,9	189,97
7	16714,4	13000,2	63,5	204,73
8	19137,5	16592,9	64,3	258,05
9	16667,9	11538,5	65,0	177,52
10	17435,5	13880,4	66,4	209,04
11	18482,0	14860,0	67,0	221,79
12	21212,6	20049,1	67,7	296,15
13	18667,3	13856,8	69,5	199,38
14	18938,4	16695,2	72,4	230,60
15	20126,4	19202,7	74,2	258,80
16	22972,5	25737,8	77,6	331,67
17	18478,2	16741,1	83,7	200,01
18	19091,9	22435,9	87,7	255,83
19	20311,7	24276,7	90,5	268,25
20	22321,4	31200,6	95,3	327,39
21	19530,6	20484,4	96,9	211,40
22	19919,6	25619,3	99,3	258,00
23	20889,1	26791,2	100,4	266,84
24	23286,3	34874,8	102,7	339,58
25	20117,4	23380,8	104,8	223,10
26	20488,2	29391,9	108,4	271,14
27	21703,5	29601,6	109,7	269,84
28	24151,4	39377,8	111,7	352,53
29	20970,9	25299,2	114,2	221,53
30	21451,3	32249,8	116,6	276,58
31	22389,5	32554,8	117,4	277,30
32	24756,0	42802,6	118,0	362,73
33	21715,4	28351,1	120,5	235,28
34	22111,2	36115,0	122,4	295,06
35	23369,1	34526,1	123,1	280,47
36	26323,4	45865,4	123,0	372,89
37	23203,5	29607,3	123,8	239,15
38	23793,1	39212,7	126,1	310,97
39	24746,9	37610,6	127,1	295,91
40	27273,3	48643,7	128,4	378,85

$C(t)$ — реальные потребительские расходы домашних хозяйств в t-м квартале (в млрд иен), $YP(t)$ — располагаемый доход домашних хозяйств, $YD(t)$ — дефлятор потребительских расходов.

Математическая модель динамики ряда $\{C(t)\}$ имеет вид

$$C(t) = \theta_1 + \theta_2 \frac{YP(t)}{YD(t)} + \theta_3 C(t-1) + E(t), \quad t = 2, 3, \ldots, 40. \tag{19.49}$$

Используя модель (19.49), необходимо построить прогноз динамики $C(t)$ для $t = 41, \ldots, 44$ и сравнить его с реальными расходами в указанный период времени.

Решение. Оценим параметры θ_1, θ_2 и θ_3 по наблюдениям, приведенным в табл. 19.5, методом наименьших квадратов. Для этого введем обозначения

$$h_{1,t} = 1, \quad h_{2,t} = \frac{YP(t)}{YD(t)}, \quad h_{3,t} = C(t-1), \quad t = 2, \ldots, 40,$$

(значения $h_{2,t}$ приведены в пятой колонке табл. 19.5).

Теперь из (19.49) следует, что

$$C(t) = h_{1,t}\theta_1 + h_{2,t}\theta_2 + h_{3,t}\theta_3 + E(t), \quad t = 2, \ldots, 40. \tag{19.50}$$

Применяя МНК к модели (19.50), находим

$$\left[\sum_{k=2}^{40} h_t (h_t)^\top\right] \theta = \sum_{t=2}^{40} h_t C(t), \tag{19.51}$$

где $h_t = \left[h_{1,t}; h_{2,t}; h_{3,t}\right]^\top$, $\theta = [\theta_1, \theta_2, \theta_3]^\top$.

Решая систему уравнений (19.51), находим реализации МНК-оценок параметров θ_1, θ_2, θ_3:

$$\widehat{\theta}_1 = 7730{,}77; \;\; \widehat{\theta}_2 = 36{,}88; \;\; \widehat{\theta}_3 = 0{,}135.$$

Итак, подобранная модель (19.50), описывающая динамику ряда $C(t)$ для $t = 2, \ldots, 40$, принимает вид:

$$\widehat{C}(t) = 7730{,}77 + 36{,}88 \frac{YP(t)}{YD(t)} + 0{,}135 \widehat{C}(t-1).$$

На рис. 19.7 приведены графики изменения и оценки $\widehat{C}(t)$ для $t = 2, 3, \ldots, 40$.

Из рис. 19.7 видно, что оценка $\widehat{C}(t)$ весьма точно отслеживает как общую тенденцию к нарастанию $C(t)$ (тренд), так и существенные сезонные колебания.

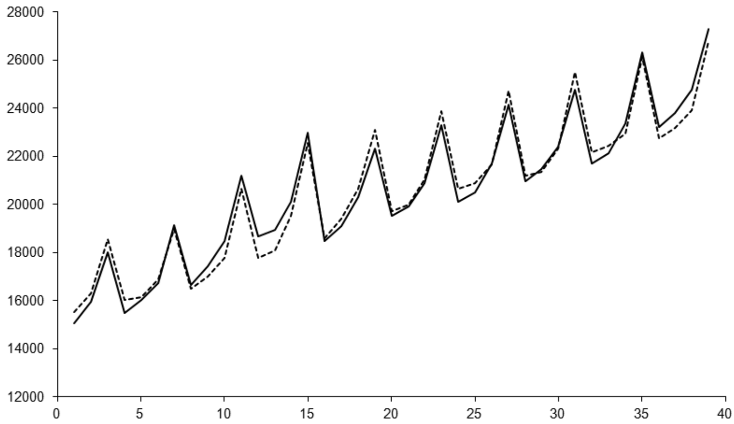


Рис. 19.7. — объем потребления C и его оценка - - -

Перейдем теперь к анализу случайной компоненты $E(t)$ ряда (19.49). Используя данные $C(t)$ и оценки $\widehat{C}(t)$, находим оценки $\widehat{E}(t)$ случайной компоненты $E(t)$:

$$\widehat{E}(t) = C(t) - \widehat{C}(t), \quad t = 2, \ldots, 40.$$

Для подбора модели, описывающей динамику $E(t)$, представим ряд $\{\widehat{E}(t),\ t = 2, \ldots, 40\}$ в графическом виде на рис.19.8.

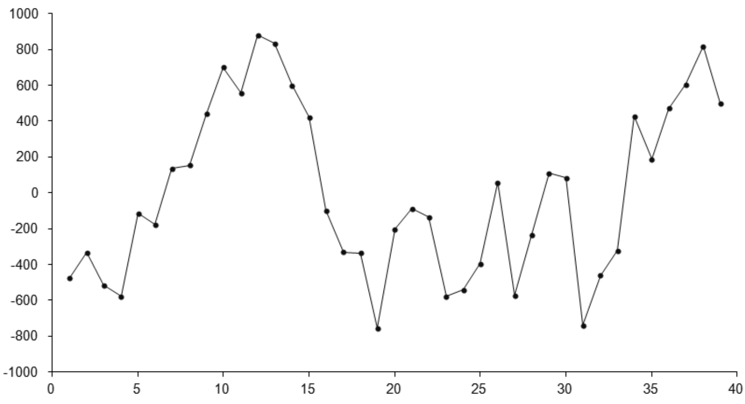


Рис. 19.8. Остатки в исходной модели модели динамики объема потребления C

Визуальный анализ графика остатков $\widehat{E}(t)$ на рис. 19.8 позволяет сделать предположение о наличии значительной положительной корреляции между сечениями $\widehat{E}(t)$ и $\widehat{E}(t+1)$.

Для проверки этого предположения вычислим реализацию первой выборочной автокорреляции ряда $\widehat{E}(t)$:

$$\widehat{r}_1^E = \frac{\sum\limits_{t=3}^{40} \widehat{E}(t)\widehat{E}(t-1)}{\sum\limits_{t=3}^{40} \widehat{E}^2(t-1)} = 0{,}749.$$

Величина $\widehat{r}_1^E$ вычислена по выборке объема $n-1=39$, поэтому

$$I = \left[-\frac{1}{n-1} - \frac{2}{\sqrt{n-1}}; -\frac{1}{n-1} + \frac{2}{\sqrt{n-1}}\right] = [-0{,}346; 0{,}295]\,. \quad (19.52)$$

Так как $\widehat{r}_1^E \notin I$, то следует принять гипотезу о коррелированности ряда $\{\widehat{E}(t)\}$.

Предположим, что $\widehat{E}(t)$ можно описать АР(1)-моделью

$$\widehat{E}(t) = a\widehat{E}(t-1) + e(t), \quad t = 2.\ldots, 40.$$

Оценивая a методом наименьших квадратов, находим

$$\widehat{a} = \widehat{r}_1^E = 0{,}749,$$

откуда получаем модель вида

$$\widehat{E}(t) = 0{,}749\widehat{E}(t-1) + \widehat{e}(t), \quad t = 2, \ldots, 40.$$

Для проверки адекватности модели найдем остатки

$$\widehat{e}(t) = \widehat{E}(t) - 0{,}749\widehat{E}(t-1), \quad t = 2, \ldots, 40.$$

График временного ряда остатков $\{\widehat{e}(t)\}$ в тестируемой АР(1)-модели приведен на рис. 19.9.

Сравнение рис. 19.8 и 19.9 показывает, что реализация ряда $\{\widehat{e}(t)\}$ существенно больше похожа на реализацию белого шума, чем $\{\widehat{E}(t)\}$. Вычислим реализацию первой выборочной автокорреляции $\widehat{r}_1^e$ ряда $\{\widehat{e}(t)\}$

$$\widehat{r}_1^e = \frac{\sum\limits_{t=3}^{40} \widehat{e}(t)\widehat{e}(t-1)}{\sum\limits_{t=3}^{40} \widehat{e}^2(t-1)} = -0{,}011. \quad (19.53)$$

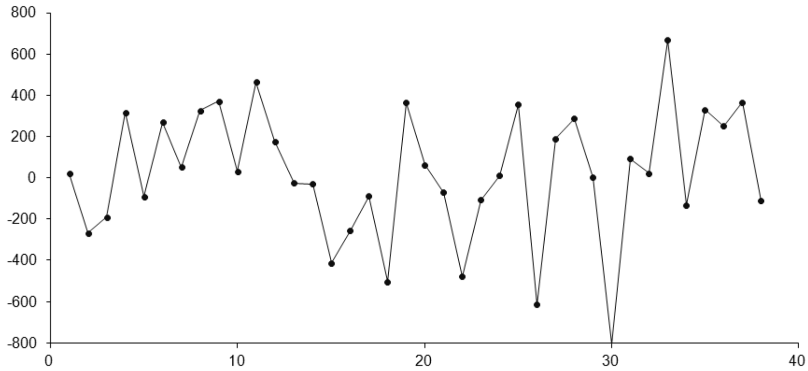


Рис. 19.9. Остатки в модели АР(1) случайной компоненты модели динамики C

Из (19.52) и (19.53) следует, что $\widehat{r}_1^e \in I$, что согласуется с гипотезой о некоррелированности остатков $\{\widehat{e}(t)\}$. Дальнейшие вычисления показывают, что k-е выборочные автокорреляции $\{\widehat{r}_k^e,\ k = 2, \ldots, 10\}$ также принадлежат I. Последнее позволяет сделать окончательный вывод о модели, описывающей динамику случайной компоненты:

$$E(t) = 0{,}749E(t-1) + e(t), \quad t = 2, \ldots, 40. \tag{19.54}$$

При этом динамика исходного ряда $\{C(t)\}$ описывается моделью

$$\begin{gathered} C(t) = 7730{,}77 + 36{,}88\frac{YP(t)}{YD(t)} + 0{,}135C(t-1) + E(t), \\ t = 2, 3, \ldots, 40. \end{gathered} \tag{19.55}$$

Для прогнозирования $C(t)$, т.е. оценивания $C(t)$ для $t > 40$ с использованием (19.54) и (19.55), воспользуемся оценкой

$$\widetilde{C}(t) = \widehat{C}(t) + \widehat{E}(t), \tag{19.56}$$

где

$$\widehat{C}(t) = 7730{,}77 + 36{,}88\frac{YP(t)}{YD(t)} + 0{,}135\widehat{C}(t-1), \tag{19.57}$$

$$\widehat{E}(t) = 0{,}749\widehat{E}(t-1). \tag{19.58}$$

Формулы (19.56) — (19.58) используются для $t \geqslant 41$, причем начальные условия для $t = 40$ уже вычислены:

$$\begin{cases} \widehat{C}(40) = C(40) = 27\,273{,}3, \\ \widehat{E}(40) = 501{,}135. \end{cases} \tag{19.59}$$

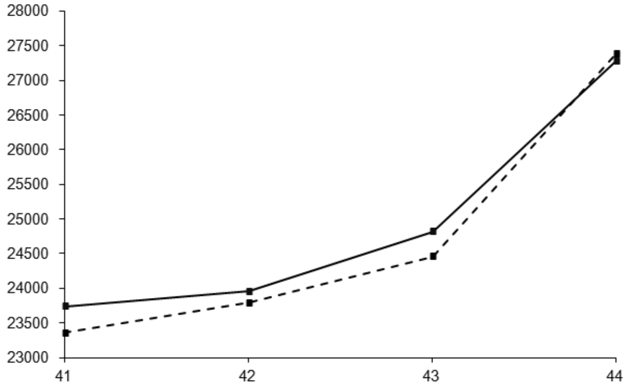


Рис. 19.10. — истинный объем потребления $C(t)$ в млрд иен;
--- прогнозные значения $\widetilde{C}(t)$

Результаты вычислений прогнозов для $t = 41, \ldots, 44$ по формулам (19.56) — (19.59) приведены в табл. 19.6.

Таблица 19.6

t	$C(t)$	$\widehat{C}(t)$	$\widehat{E}(t)$	$\widetilde{C}(t)$	$\Delta(t)$	$\delta(t)$
41	23741,9	23354,8	375,3	23730,1	11,78	0,0496
42	23954,7	23794,4	281,1	24075,5	−120,79	0,5043
43	24816,6	24459,1	210,5	24669,6	146,98	0,5923
44	27288,3	27387,9	157,7	27545,6	−257,27	0,9428

Во второй колонке табл. 19.6 приведены реальные значения $C(t)$, в шестой — реализация абсолютной ошибки $\Delta(t) = C(t) - \widetilde{C}(t)$ прогноза, а в последней — относительная (в %): $\delta(t) = \dfrac{|\Delta(t)|}{C(t)} \cdot 100\%$. Полученные данные показывают, что прогноз $\widetilde{C}(t)$ оказался весьма точным, так как $\delta(t) < 1\%$ для всех $t = 41, \ldots, 44$.

Графики $C(t)$ и $\widetilde{C}(t)$ приведены на рис. 19.10. ■

19.6. Задачи для самостоятельного решения

1. Для модели АР(1) вида (17.17) получите представление (17.22) и найдите ковариационную функцию $R_\varepsilon(k)$ вида (17.23).

Ответ: $\varepsilon_t = e_t + \sum\limits_{m=1}^{\infty} a^m e_{t-m}$; $R_\varepsilon(k) = a^{|k|}\sigma_\varepsilon^2$, где $\sigma_\varepsilon^2 = \dfrac{\sigma_e^2}{1 - a^2}$.

2. В табл. 19.7 приведены статистические данные по отлову рыси в Канаде за период 1821—1922 гг.

Таблица 19.7

№ года	Число голов	№ года	Число голов	№ года	Число голов
1	269	35	1638	69	39
2	321	36	2725	70	49
3	585	37	2871	71	59
4	871	38	2119	72	188
5	1475	39	684	73	377
6	2821	40	299	74	1292
7	3928	41	236	75	4031
8	5943	42	245	76	3495
9	4950	43	552	77	587
10	2577	44	1623	78	105
11	523	45	3311	79	153
12	98	46	6721	81	758
14	279	48	687	82	1307
15	409	49	255	83	3465
16	2285	50	473	84	6991
17	2685	51	358	85	6313
18	3409	52	784	86	3794
19	1824	53	1594	87	1836
20	409	54	1676	88	345
21	151	55	2251	89	382
22	45	56	1426	90	808
23	68	57	756	91	1388
24	213	58	299	92	2713
25	546	59	201	93	3800
26	1033	60	229	94	3091
27	2129	61	469	95	2985
28	2536	62	736	96	3790
29	957	63	2042	97	674
30	361	64	2811	98	81
31	377	65	4431	99	80
32	225	66	2511	100	108
33	360	67	389	101	229
34	731	68	73	102	399

Покажите, что соответствующий временно́й ряд описывается следующей моделью:

$$X_t = \theta_1 + \theta_2 t + s_t + \varepsilon_t, \quad t = 1, \ldots, 102,$$

где s_t — сезонная компонента, ε_t — случайная компонента, описываемая АР(2)-моделью.

3. В условиях примера 19.3 покажите, что объем потребления C_t достаточно точно описывается аддитивной линейной моделью с сезонной и случайной компонентами:

$$C_t = \theta_1 + \theta_2 t + s_t + \varepsilon_t, \quad t = 1, \ldots, 40,$$

где s_t — сезонная компонента, ε_t — АР(1)-модель. Постройте оценки параметров и переменных в модели для C_t по наблюдениям $t = 1, \ldots, 40$ и постройте прогноз C_t для $t = 41, \ldots, 44$. Сравните построенный прогноз с реальными значениями C_t и с прогнозом $\widetilde{C}_t$, построенным в примере 19.3.

4. В табл. 19.8 приведены данные о годовом количестве забастовок за период 1951—1980 гг. в США.

Таблица 19.8

№ года	Число забастовок	№ года	Число забастовок	№ года	Число забастовок
1	4737	11	3367	21	5138
2	5117	12	3614	22	5010
3	5091	13	3362	23	5353
4	3468	14	3655	24	6074
5	4320	15	3963	25	5031
6	3825	16	4405	26	5648
7	3673	17	4595	27	5506
8	3694	18	5045	28	4230
9	3708	19	5700	29	4827
10	3333	20	5716	30	3885

Требуется построить математическую модель соответствующего временного ряда и построить прогноз количества забастовок на три года 1981—1983 гг.

Указание.

1. В качестве модели тренда используйте многочлен третьего порядка.

2. Покажите, что случайная компонента ряда может быть описана АР(1)-моделью.

3. Сезонная компонента отсутствует.

5. В табл.19.9 приведены статистические данные о количестве несчастных случаев за месяц с 1973 по 1978 гг. в США.

Таблица 19.9

Месяц	1973	1974	1975	1976	1977	1978
Январь	9 007	7 750	8 162	7 717	7 792	7 836
Февраль	8 106	6 981	7 306	7 461	6 957	6 892
Март	8 928	8 038	8 124	7 776	7 726	7 791
Апрель	9 137	8 422	7 870	7 925	8 106	8 129
Май	10 017	8 714	9 387	8 634	8 890	9 115
Июнь	10 826	9 512	9 556	8 945	9 299	9 434
Июль	11 317	10 120	10 093	10 078	10 625	10 484
Август	10 744	9 823	9 620	9 179	9 302	9 827
Сентябрь	9 713	8 743	8 285	8 037	8 314	9 110
Октябрь	9 938	9 129	8 433	8 488	8 850	9 070
Ноябрь	9 161	8 710	8 160	7 874	8 265	8 633
Декабрь	8 927	8 680	8 034	8 647	8 796	9 240

Постройте математическую модель временного ряда вида

$$X_t = \varphi_t + s_t + \varepsilon_t, \quad t = 1, \ldots,$$

где $\varphi_t = \theta_1 + \theta_2 t + \theta_3 t^2$, s_t — сезонная компонента, ε_t — случайная компонента, описываемая АР(1)-моделью.

Г Л А В А V

МАТЕМАТИЧЕСКОЕ ПРИЛОЖЕНИЕ

В данной главе приводятся необходимые сведения из курсов функционального анализа и теории вероятностей, а также статистические таблицы, используемые для вычисления квантилей СВ и значений функции Лапласа.

20. Необходимые сведения из функционального анализа

20.1. Алгебры и σ-алгебры множеств

Определение 20.1. Система $\mathcal{A}$ подмножеств некоторого множества X называется *алгеброй*, если:

1) $\varnothing, X \in \mathcal{A}$;

2) $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A},\ A \cap B \in \mathcal{A}$;

3) $A \in \mathcal{A} \implies \overline{A} \in \mathcal{A}$.

Свойство 2) выполнено для любого конечного набора подмножеств, т.е. если $A_k \in \mathcal{A}$, $k = 1, \ldots, n$, то

$$\bigcup_{k=1}^{n} A_k \in \mathcal{A}, \qquad \bigcap_{k=1}^{n} A_k \in \mathcal{A}.$$

Определение 20.2. Алгебра $\mathcal{A}$ называется σ-*алгеброй*, если справедливо следующее усиление свойства 2):

$$A_n \in \mathcal{A},\ n = 1, 2, \ldots \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{A},\ \bigcap_{n=1}^{\infty} A_n \in \mathcal{A}.$$

Определение 20.3. Множество X вместе с некоторой σ-алгеброй его подмножеств $\mathcal{A}$ называется *измеримым пространством* и обозначается $\{X, \mathcal{A}\}$.

На одном и том же множестве X могут быть заданы различные σ-алгебры его подмножеств, при этом, соответственно, возникают различные измеримые пространства. Примерами σ-алгебр являются системы множеств

$$\mathcal{A}_0 = \{\varnothing, X\}, \qquad \mathcal{A}^0 = \{A : A \subseteq X\}.$$

При этом $\mathcal{A}_0$ — самая «бедная» σ-алгебра, называемая *тривиальной*, а $\mathcal{A}^0$ — самая «богатая» σ-алгебра, состоящая *из всех подмножеств* X.

Теорема 20.1. *Пусть на множестве X задана некоторая система его подмножеств $\mathcal{D}$. Тогда существует наименьшая σ-алгебра, обозначаемая $\sigma(\mathcal{D})$, содержащая все множества из $\mathcal{D}$.*

Система $\sigma(\mathcal{D})$ является наименьшей в том смысле, что если $\mathcal{A}$ — любая σ-алгебра подмножеств X, содержащая систему $\mathcal{D}$, то $\sigma(\mathcal{D}) \subseteq \mathcal{A}$.

Замечания. 1. Систему множеств $\sigma(\mathcal{D})$ называют *σ-алгеброй, порожденной системой множеств* $\mathcal{D}$.

2. Если $\{\mathcal{A}_\alpha\}$ — произвольное семейство σ-алгебр на X, то $\mathcal{A} = \bigcap\limits_\alpha \mathcal{A}_\alpha$ также является σ-алгеброй. Очевидно, что $\sigma(\mathcal{D}) = \bigcap\limits_\alpha \mathcal{A}_\alpha$, где $\{\mathcal{A}_\alpha\}$ — семейство всех σ-алгебр, содержащих систему множеств $\mathcal{D}$.

20.2. Меры (определения и свойства)

Определение 20.4. Пусть на множестве X задана некоторая алгебра его подмножеств $\mathcal{A}$. Функция $\mu(A)$, определенная на множествах $A \in \mathcal{A}$, называется *мерой, заданной на* $\mathcal{A}$, если:

а) $\mu(A) \geqslant 0$ для всех $A \in \mathcal{A}$;

б) для любого счетного набора попарно непересекающихся множеств $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$, (т.е. $A_i \bigcap A_j$, $i \neq j$) таких, что $\bigcup\limits_{n=1}^{\infty} A_n \in \mathcal{A}$ выполнено свойство *счетной аддитивности* (*σ-аддитивности*):

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Мера μ называется конечной, если дополнительно выполнено:

в) $\mu(X) < \infty$.

Мера μ, заданная на алгебре $\mathcal{A}$, обладает следующими свойствами:

1) $\mu(\varnothing) = 0$;

2) $\mu(A) \leqslant \mu(B)$ для $A, B \in \mathcal{A}$ таких, что $A \subseteq B$;

3) $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$ для всех $A, B \in \mathcal{A}$;

4) если $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$ — убывающая последовательность множеств, т.е. $A_1 \supseteq A_2 \supseteq \ldots$, такая, что $\mu(A_1) < \infty$ и $\bigcap\limits_{n=1}^{\infty} A_n \in \mathcal{A}$, то

$$\mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n);$$

5) если $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$ — возрастающая последовательность множеств, т.е. $A_1 \subseteq A_2 \subseteq \ldots$, такая, что $\bigcup\limits_{n=1}^{\infty} A_n \in \mathcal{A}$, то

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n).$$

Замечание. Если на алгебре $\mathcal{A}$ задана функция множества μ, обладающая свойством *аддитивности*:

$$A, B \in \mathcal{A},\ A \cap B = \varnothing \quad \Longrightarrow \quad \mu(A \cup B) = \mu(A) + \mu(B),$$

то для доказательства того, что μ является мерой на $\mathcal{A}$, достаточно проверить свойство 4) для случая, когда $\bigcap\limits_{n=1}^{\infty} A_n = \varnothing$, т.е. $\mu(A_n) \to 0$ при $n \to \infty$.

20.3. Способы задания мер

Следующие примеры измеримых пространств являются наиболее важными для теории вероятностей и случайных процессов.

Дискретное измеримое пространство. Пусть множество $X = \{x_1, x_2, \ldots\}$ не более чем счетно, а $\mathcal{A}$ — σ-алгебра всех подмножеств X. Всякая мера μ на *дискретном измеримом пространстве* $\{X, \mathcal{A}\}$ задается числами $\mu_n = \mu(\{x_n\}) \geqslant 0$:

$$\mu(A) = \sum_{n: x_n \in A} \mu_n,$$

где $A \in \mathcal{A}$ — любое подмножество X. Мера μ конечна, если

$$\mu(X) = \mu\left(\bigcup_{n=1}^{\infty} \{x_n\}\right) = \sum_{n=1}^{\infty} \mu(\{x_n\}) = \sum_{n=1}^{\infty} \mu_n < \infty.$$

Измеримое пространство $\{\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1)\}$. Пусть $X = \mathbb{R}^1$ — действительная прямая, а $\langle a, b\rangle$ — промежуток, т.е. одно из множеств вида

$$(a, b], \quad [a, b), \quad (a, b), \quad [a, b],$$

где $-\infty \leqslant a \leqslant b \leqslant \infty$. Обозначим через $\mathcal{A}$ систему подмножеств $A \subseteq \mathbb{R}^1$, состоящих из конечных объединений непересекающихся промежутков:

$$A = \bigcup_{i=1}^{n} \langle a_i, b_i\rangle, \quad n < \infty. \tag{20.1}$$

Очевидно, система $\mathcal{A}$ образует алгебру, но не является σ-алгеброй.

О п р е д е л е н и е 20.5. σ-алгебра, порожденная системой $\mathcal{A}$, обозначается $\mathcal{B}(\mathbb{R}^1)$ и называется *борелевской σ-алгеброй* множеств действительной прямой, а ее элементы — *борелевскими множествами.*

Аналогично вводится измеримое пространство $\{[a, b], \mathcal{B}([a, b])\}$, где $\mathcal{B}([a, b])$ состоит из множеств $B \subseteq [a, b]$ таких, что $B \in \mathcal{B}(\mathbb{R}^1)$. Система $\mathcal{B}([a, b])$ называется *борелевской σ-алгеброй* отрезка $[a, b]$.

О п р е д е л е н и е 20.6. *Мерой Лебега на* $\mathbb{R}^1$ называется мера λ, определенная на $\{\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1)\}$, такая, что

$$\lambda(\langle a, b\rangle) = b - a.$$

для всех $-\infty \leqslant a \leqslant b \leqslant \infty$.

Отметим, что мера Лебега на $\mathbb{R}^1$ не является конечной, т.е. $\lambda(\mathbb{R}^1) = \infty$.

Конечная мера на борелевской σ-алгебре прямой или отрезка может быть задана с помощью *функции распределения*, которая определяется следующим образом.

О п р е д е л е н и е 20.7. Функция $F(x)$, $x \in \mathbb{R}^1$ обладающая свойствами:

1) $F(x)$ — неубывающая функция;

2) $F(-\infty) = \lim\limits_{x\to-\infty} F(x) = 0$, $F(+\infty) = \lim\limits_{x\to+\infty} F(x) < \infty$;

3) $F(x)$ непрерывна справа, т.е. $\lim\limits_{h\to+0} F(x + h) = F(x)$ для всех $x \in \mathbb{R}^1$,

называется *функцией распределения* на $\mathbb{R}^1$.

З а м е ч а н и е. Для функция распределения на отрезке $[a, b]$ свойство 2) принимает вид

$$F(a) = \lim_{x\to a} F(x) = 0, \qquad F(b) = \lim_{x\to b} F(x) < \infty.$$

З а м е ч а н и я. 1. С каждой конечной мерой μ, заданной на $\{\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1)\}$, можно связать функцию

$$F_{\mu}(x) = \mu((-\infty, x]), \quad x \in \mathbb{R}^1.$$

Из свойств меры (см. разд. 20.2) вытекает, что F_μ является функцией распределения, причем $F_\mu(+\infty) = \lim\limits_{x\to+\infty} F_\mu(x) = \mu(\mathbb{R}^1)$.

2. Между функциями распределения и конечными мерами на $\{\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1)\}$ существует взаимно однозначное соответствие, т.е. всякой конечной мере μ соответствует функция распределения $F_\mu(x)$, и наоборот, для всякой функции распределения $F(x)$ на $\mathbb{R}^1$ существует конечная мера μ такая, что $F_\mu \equiv F$.

Измеримое пространство $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\}$. Пространство $\mathbb{R}^n$ есть прямое произведение n экземпляров прямых $\mathbb{R}^1$, т.е. $\mathbb{R}^n = \mathbb{R}^1 \times \ldots \times \mathbb{R}^1$ — множество упорядоченных наборов $x = (x_1, \ldots, x_n)^\top$. Определим на этом пространстве систему подмножеств $\mathcal{A}^n$, образованную множествами

$$A = A_1 \times \ldots \times A_n = \prod_{k=1}^{n} A_k, \quad A_k \in \mathcal{A},$$

где $\mathcal{A}$ — алгебра подмножеств прямой вида (20.1). Нетрудно показать, что $\mathcal{A}^n$ образует алгебру.

Определение 20.8. σ-алгебра, порожденная системой $\mathcal{A}^n$, обозначается $\mathcal{B}(\mathbb{R}^n)$ и называется *борелевской σ-алгеброй* множеств $\mathbb{R}^n$, а ее элементы — *борелевскими множествами*.

Определение 20.9. *Мерой Лебега на* $\mathbb{R}^n$ называется мера λ^n, определенная на $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\}$, такая, что

$$\lambda^n\left(\prod_{i=1}^{n} \langle a_i, b_i\rangle\right) = \prod_{i=1}^{n} (b_i - a_i),$$

для всех $-\infty \leqslant a_i \leqslant b_i \leqslant \infty$.

Отметим, что мера Лебега на $\mathbb{R}^n$ не является конечной, т.е. $\lambda^n(\mathbb{R}^n) = \infty$.

Конечная мера на борелевской σ-алгебре $\mathbb{R}^n$ может быть задана с помощью *n-мерной функции распределения*, которая определяется следующим образом.

Определение 20.10. Функция $F(x_1, \ldots, x_n)$, $x_1, \ldots, x_n \in \mathbb{R}^1$, называется *$n$-мерной функцией распределения*, если она обладает следующими свойствами:

1) $F(x_1, \ldots, x_n)$ монотонна в следующем смысле: определим оператор Δ_i конечной разности по переменной x_i как

$$\Delta_i F = F(x_1, \ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots, x_n) - \\ - F(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n),$$

где $h_i \geqslant 0$, тогда для любого набора $\{h_i \geqslant 0,\ i = 1, \ldots, n\}$,

$$\Delta_1 \ldots \Delta_n F(x_1, \ldots, x_n) \geqslant 0;$$

2) если хотя бы одна из переменных $x_i \to -\infty$, то

$$F(x_1, \ldots, x_n) \to 0;$$

если все переменные $x_i \to +\infty$, то

$$F(x_1, \ldots, x_n) \to F(+\infty, \ldots, +\infty) < \infty;$$

3) $F(x_1, \ldots, x_n)$ непрерывна справа по переменным x_i.

На измеримом пространстве $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\}$ существует однозначно определенная конечная мера μ такая, что $F_\mu \equiv F$, где

$$F_\mu(x_1, \ldots, x_n) = \mu\left(\prod_{i=1}^{n}(-\infty, x_i]\right), \quad x_1, \ldots, x_n \in \mathbb{R}^1. \tag{20.2}$$

Верно и обратное, если μ — конечная мера на $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\}$, то функция $F_\mu(x_1, \ldots, x_n)$, определяемая соотношением (20.2), является n-мерной функцией распределения. Тем самым между n-мерными функциями распределения и конечными мерами на $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\}$ существует взаимно однозначное соответствие. В этом случае $F_\mu(+\infty, \ldots, +\infty) = \mu(\mathbb{R}^n)$.

20.4. Измеримые функции

Пусть $\{X, \mathcal{A}\}$ — некоторое измеримое пространство.

Определение 20.11. Вещественная функция $f(x)$, $x \in X$ называется *$\mathcal{A}$-измеримой*, если

$$f^{-1}(B) \in \mathcal{A} \quad \forall B \in \mathcal{B}(\mathbb{R}^1), \tag{20.3}$$

где $f^{-1}(B) = \{x \in X : f(x) \in B\}$ — *прообраз* множества B.

Тем самым измеримость функции означает то, что прообраз любого борелевского подмножества $\mathbb{R}^1$ является измеримым множеством в X.

Функция $\varphi(y)$, $y \in \mathbb{R}^1$, заданная на действительной прямой, называется *борелевской функцией*, если она $\mathcal{B}(\mathbb{R}^1)$-измерима. Примерами борелевских функций являются все кусочно-непрерывные функции.

Теорема 20.2. *Пусть $\{X, \mathcal{A}\}$ — измеримое пространство. Сложная функция $h(x) = \varphi(f(x))$, $x \in X$, является $\mathcal{A}$-измеримой, если функция $f(x)$, $x \in X$, $\mathcal{A}$-измерима, а функция $\varphi(y)$, $y \in \mathbb{R}^1$ является борелевской.*

Простые арифметические операции над конечным или счетным набором измеримых функций не выводят за рамки множества измеримых функций.

Теорема 20.3. *Пусть функции* f_n, $n = 1, 2, \ldots$ *определены на измеримом пространстве* $\{X, \mathcal{A}\}$ *и* $\mathcal{A}$*-измеримы. Тогда функции*

$$f_1(x) + f_2(x), \quad f_1(x) f_2(x), \quad 1/f_1(x) \ (\text{при условии } f_1(x) \neq 0),$$
$$|f_1(x)|, \ \max\{f_1(x), f_2(x)\}, \ \min\{f_1(x), f_2(x)\}, \ \sup_n f_n(x), \ \inf_n f_n(x)$$

также являются измеримыми.

Из приведенного результата следует, что множество

$$A = \{x \in X : \exists \lim_n f_n(x)\},$$

на котором существует предел последовательности измеримых функций $f_n(x)$, является измеримым, т.е. $A \in \mathcal{A}$.

21. Необходимые сведения из теории вероятностей

21.1. Случайные события и их вероятности

Определение 21.1. Совокупность объектов $\{\Omega, \mathcal{F}, \mathbf{P}\}$, где

Ω — пространство *элементарных событий* ω;

$\mathcal{F}$ — σ*-алгебра подмножеств* пространства Ω, образующих систему *случайных событий*;

$\mathbf{P}$ — *нормированная* (т.е. $\mathbf{P}\{\Omega\} = 1$) *мера* на $\mathcal{F}$,

называется *вероятностным пространством*, а мера $\mathbf{P}$ — *вероятностной мерой* или просто *вероятностью.*

Замечания. 1. Предполагается, что $\{\Omega, \mathcal{F}, \mathbf{P}\}$ — полное вероятностное пространство.

2. Над случайными событиями из $\mathcal{F}$ можно совершать действия, аналогичные действиям над множествами, для которых мы будем использовать следующие обозначения:

$$A + B = A \cup B, \quad AB = A \cap B, \quad A \setminus B, \quad \overline{A} = \Omega \setminus A,$$

$$\sum_k A_k = \bigcup_k A_k, \quad \prod_k A_k = \bigcap_k A_k,$$

где событие $\overline{A}$ называется *противоположным* A.

Вероятность $\mathbf{P}\{\cdot\}$ обладает следующими свойствами:

1) $0 \leqslant \mathbf{P}\{A\} \leqslant 1$ для любого события $A \in \mathcal{F}$;

2) $\mathbf{P}\{\Omega\} = 1$, где Ω — *достоверное событие*;

3) $\mathbf{P}\{\varnothing\} = 0$, где $\varnothing = \overline{\Omega}$ — *невозможное событие*;

4) $\mathbf{P}\{A + B\} = \mathbf{P}\{A\} + \mathbf{P}\{B\}$, если $AB = \varnothing$, т.е. события A и B *несовместны*;

5) $\mathbf{P}\{A+B\} = \mathbf{P}\{A\} + \mathbf{P}\{B\} - \mathbf{P}\{AB\}$ для любых событий $A, B \in \in \mathcal{F}$;

6) $\mathbf{P}\{A\} \leqslant \mathbf{P}\{B\}$, если $A \subseteq B$ (т.е. A — *частный случай* события B).

Замечание. Указанные свойства следуют из общих свойств меры. Свойство 4 распространяется очевидным образом на любое конечное или счетное множество несовместных событий:

$$\mathbf{P}\left\{\sum_{k=1}^{\infty} A_k\right\} = \sum_{k=1}^{\infty} \mathbf{P}\{A_k\}, \quad A_k \in \mathcal{F}, \quad A_m A_n = \varnothing \quad \text{при} \quad m \neq n.$$

Определение 21.2. События A и B называются *независимыми*, если $\mathbf{P}\{AB\} = \mathbf{P}\{A\}\,\mathbf{P}\{B\}$. События $\{A_n\}$ *независимы в совокупности*, если для любого конечного набора событий A_{n_k}, $k = 1, \ldots, m$

$$\mathbf{P}\left\{\prod_{k=1}^{m} A_{n_k}\right\} = \prod_{k=1}^{m} \mathbf{P}\{A_{n_k}\},$$

где m может равняться ∞.

Определение 21.3. *Условной вероятностью* события A относительно события B такого, что $\mathbf{P}\{B\} > 0$, называется величина

$$\mathbf{P}\{A \mid B\} = \frac{\mathbf{P}\{AB\}}{\mathbf{P}\{B\}}.$$

Если события A, B независимы и имеют положительные вероятности, то $\mathbf{P}\{A \mid B\} = \mathbf{P}\{A\}$ и $\mathbf{P}\{B \mid A\} = \mathbf{P}\{B\}$.

Пусть события $H_1, \ldots, H_N \in \mathcal{F}$ удовлетворяют условиям:

а) $\mathbf{P}\{H_k\} > 0$ при всех k;

б) $H_m H_n = \varnothing$, если $m \neq n$;

в) $\sum_{k=1}^{N} H_k = \Omega$,

тогда для любого $A \in \mathcal{F}$ справедлива *формула полной вероятности*:

$$\mathbf{P}\{A\} = \sum_{k=1}^{N} \mathbf{P}\{H_k\}\mathbf{P}\{A \mid H_k\}.$$

События $\{H_k\}$ обычно называют *вероятностными гипотезами.*

21.2. Случайные величины и векторы

Определение 21.4. *Случайной величиной* (СВ), определенной на $\{\Omega, \mathcal{F}, \mathbf{P}\}$, называется числовая функция $X(\omega)$, $\omega \in \Omega$, измеримая относительно $\mathcal{F}$.

Определение 21.4 означает, что для всякого борелевского подмножества $B \subseteq \mathbb{R}^1$ множество

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \tag{21.1}$$

является случайным событием.

Далее для краткости будем опускать аргумент ω: X, $\{X \in B\}$ и т. п.

Сумма, разность, произведение и частное двух случайных величин (при условии, что знаменатель не обращается в нуль) также являются случайными величинами.

Из теоремы 20.2 следует, что если X — случайная величина, а $g(x)$, $x \in \mathbb{R}^1$ — борелевская функция, то $g(X)$ также является случайной величиной, так как функция $Y(\omega) = g(X(\omega))$, $\omega \in \Omega$ является $\mathcal{F}$-измеримой. В частности, любое непрерывное преобразование СВ X приводит к СВ Y (т.е. свойство $\mathcal{F}$-измеримости сохраняется).

Определение 21.5. Функция $F_X(x) = \mathbf{P}\{X \leqslant x\}$, $x \in \mathbb{R}^1$ называется *функцией распределения* СВ X.

Функция распределения $F_X(x)$ обладает всеми свойствами функции распределения меры на измеримом пространстве $\{\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1)\}$ (см. определение 20.7, которое полностью применимо с учетом соотношения $\mu(\mathbb{R}^1) = 1$).

Для любой функции распределения $F(x)$ существует полное вероятностное пространство $\{\Omega, \mathcal{F}, \mathbf{P}\}$ и заданная на нем СВ X такая, что $F_X(x) = F(x)$.

Рассмотрим конкретные типы функций распределения.

Определение 21.6. Если СВ X принимает значения из конечного или счетного множества $\{a_1, \ldots, a_n, \ldots\}$ с вероятностями соответственно $\{p_1, \ldots, p_n, \ldots\}$, где $p_n > 0$, $\sum\limits_n p_n = 1$, то говорят, что случайная величина является *дискретной* (или имеет *дискретное распределение*). Ее функция распределения имеет вид

$$F_X(x) = \sum_{k:\, a_k \leqslant x} p_k, \quad x \in \mathbb{R}^1.$$

Таким образом, функция распределения дискретной СВ имеет разрывы первого рода в точках a_k, а величины скачков равны $F_X(a_k) - F_X(a_k-) = p_k$. При этом для любого множества $B \in \mathcal{B}(\mathbb{R}^1)$

$$\mathbf{P}\{X \in B\} = \sum_{k:\, a_k \in B} p_k.$$

Если множество значений, которые принимает дискретная СВ, конечно, то ее функция распределения кусочно-постоянна.

Функция множества $\mathbf{P}_X(B) = \mathbf{P}\{X \in B\}$ является *мерой* на $\mathcal{B}(\mathbb{R}^1)$ и называется *законом распределения* СВ X. При этом $F_X(x)$ является функцией распределения этой меры. При определенных условиях мера $\mathbf{P}_X(B)$ абсолютно непрерывна относительно меры Лебега на $\mathcal{B}(\mathbb{R}^1)$, т.е. $F_X(x)$ имеет плотность распределения $p_X(x)$.

Определение 21.7. Если функция распределения $F_X(x)$ случайной величины X допускает представление

$$F_X(x) = \int\limits_{-\infty}^{x} p_X(y)\,dy, \quad \text{где} \quad p_X(y) \geqslant 0, \quad \int\limits_{-\infty}^{\infty} p_X(y)\,dy = 1,$$

а интеграл понимается как интеграл Лебега, то СВ X называется *абсолютно непрерывной* (имеет *непрерывное распределение*). Для любого множества $B \in \mathcal{B}(\mathbb{R}^1)$

$$\mathbf{P}_X(B) = \int\limits_B p_X(y)dy.$$

Функция $p_X(y)$ называется *плотностью вероятности* СВ X.

Заметим, что функция $F_X(x)$ в данном случае не имеет разрывов и почти всюду на $\mathbb{R}^1$ дифференцируема: $\dfrac{dF_X(x)}{dx} = p_X(x)$.

В общем случае функция распределения $F_X(x)$ непрерывна справа в каждой точке разрыва $x \in \mathbb{R}^1$. Для вычисления вероятности события $\{X \in B\}$, где $B \in \mathcal{B}(\mathbb{R}^1)$, следует вычислять интеграл Лебега–Стилтьеса:

$$\mathbf{P}\{X \in B\} = \int\limits_B dF_X(y) = \mathbf{P}_X(B), \quad B \in \mathcal{B}(\mathbb{R}^1).$$

Предположим, что СВ $X_1, \ldots, X_n$ определены на одном вероятностном пространстве $\{\Omega, \mathcal{F}, \mathbf{P}\}$. Тогда упорядоченный набор n случайных величин $X = \{X_1, \ldots, X_n\}^\top$ будем называть *n-мерным случайным вектором.*

Определение 21.8. Наименьшую σ-алгебру, содержащую все события вида

$$X^{-1}(B) = \{X \in B\}, \quad B \in \mathcal{B}(\mathbb{R}^n), \tag{21.2}$$

будем называть *σ-алгеброй, порожденной случайным вектором* X. Эта σ-алгебра обозначается $\sigma\{X\}$ или $\mathcal{F}^X$.

Определение 21.9. Функция

$$F_X(x_1, \ldots, x_n) = \mathbf{P}\{(X_1 \leqslant x_1) \cdot \ldots \cdot (X_n \leqslant x_n)\}, \quad x_1, \ldots, x_n \in \mathbb{R}^1$$

называется *функцией распределения n-мерного случайного вектора* X.

Функция распределения случайного вектора обладает всеми свойствами функции распределения меры на измеримом пространстве $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\}$ (см. определение 20.10, которое полностью применимо с учетом соотношения $\mu(\mathbb{R}^n) = 1$).

Для любого борелевского множества $B \in \mathcal{B}(\mathbb{R}^n)$ вероятность случайного события $\{X \in B\}$ вычисляется как интеграл Лебега:

$$\mathbf{P}_X(B) = \mathbf{P}\{X \in B\} = \int_B dF_X(x_1, \ldots, x_n).$$

Мера $\mathbf{P}_X(\cdot)$ на $\mathcal{B}(\mathbb{R}^n)$ — *закон распределения случайного вектора* X, заданный с помощью $F_X(x_1, \ldots, x_n)$. Если мера $\mathbf{P}_X(\cdot)$ имеет плотность, т.е.

$$\mathbf{P}\{X \in B\} = \int_B p_X(y_1, \ldots, y_n) dy_1 \ldots dy_n \quad \forall B \in \mathcal{B}(\mathbb{R}^n),$$

то

$$F_X(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} p_X(y_1, \ldots, y_n) dy_1 \ldots dy_n.$$

Определение 21.10. Случайные величины $\{X_1, \ldots, X_n\}$ *независимы в совокупности*, если для произвольного набора множеств $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R}^1)$ события $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ независимы в совокупности.

Теорема 21.1. *Для того чтобы случайные величины $\{X_1, \ldots, X_n\}$ были независимы в совокупности, необходимо и достаточно, чтобы для всех $x_1, \ldots, x_n \in \mathbb{R}^1$*

$$F_X(x_1, \ldots, x_n) = \prod_{k=1}^{n} F_{X_k}(x_k).$$

Общий способ определения независимости случайных событий и величин состоит в следующем. Пусть $\mathcal{F}_1 \subseteq \mathcal{F}$ и $\mathcal{F}_2 \subseteq \mathcal{F}$ — некоторые σ-алгебры случайных событий. Системы $\mathcal{F}_1$ и $\mathcal{F}_2$ называются *независимыми*, если независимы любые два события $A \in \mathcal{F}_1$ и $B \in \mathcal{F}_2$, т.е. $\mathbf{P}\{AB\} = \mathbf{P}\{A\}\mathbf{P}\{B\}$.

Теперь нетрудно определить понятие независимости СВ X и Y: X и Y независимы тогда и только тогда, когда независимы σ-алгебры $\mathcal{F}^X$ и $\mathcal{F}^Y$. В частности, СВ X не зависит от случайного события A, если A и $\mathcal{F}^X$ независимы (т.е. независимы A и любое $B \in \mathcal{F}^X$).

Приведем примеры некоторых наиболее важных законов распределения.

1. Дискретная СВ X имеет *биномиальное распределение* с параметрами $(N; p)$, где $0 < p < 1$, и обозначается $Bi(N; p)$, если

$$\mathbf{P}\{X = m\} = C_N^m p^m q^{N-m}, \quad m = 0, 1, \ldots, N,$$

где $C_N^m = \dfrac{N!}{m!(N-m)!}$ — число сочетаний из N по m, $q = 1 - p$. Распределение $Bi(1; p)$ называется *распределением Бернулли.*

2. Дискретный k-мерный случайный вектор $X = (X_1, \ldots, X_k)^\top$ имеет полиномиальное распределение с параметрами $(N, p_1, \ldots, p_k)$, где $0 < p_i < 1$, $i = 1, \ldots, k$ и $\sum\limits_{i=1}^{k} p_i = 1$, если

$$\mathbf{P}\{X = m\} = \mathbf{P}\{X_1 = m_1, X_2 = m_2, \ldots, X_k = m_k\} = \\ = \frac{N!}{m_1! m_2! \cdots m_k!} p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$$

для $m = (m_1, \ldots m_k)^\top$ таких, что $\sum\limits_{i=1}^{k} m_i = N$.

3. Дискретная СВ X имеет *распределение Пуассона* с параметром $\lambda > 0$, и обозначается $\Pi(\lambda)$, если

$$\mathbf{P}\{X = m\} = \frac{\lambda^m}{m!} e^{-\lambda}, \quad m = 0, 1, \ldots$$

4. Непрерывная СВ X имеет *равномерное распределение* на отрезке $[a, b]$, и обозначается $R[a, b]$, если ее плотность распределения имеет вид

$$p_X(x) = \begin{cases} 1/(b-a), & \text{если} \quad x \in [a, b], \\ 0, & \text{если} \quad x \notin [a, b]. \end{cases}$$

5. Непрерывная СВ X имеет *экспоненциальное* (или *показательное*) *распределение* с параметром $\lambda > 0$, и обозначается $E(\lambda)$, если ее плотность распределения имеет вид

$$p_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0, \\ 0, & x < 0. \end{cases}$$

6. Непрерывная СВ X имеет распределение *Лапласа* (или *двойное экспоненциальное*) с параметром $\lambda > 0$, и обозначается $\mathcal{L}(\lambda)$, если ее плотность распределения имеет вид

$$p_X(x) = \frac{\lambda}{2} \exp -\lambda |x|.$$

7. Непрерывная СВ X имеет *логистическое распределение* с параметрами $(m; \sigma^2)$, и обозначается $Lg(m; \sigma^2)$, если ее плотность распределения имеет вид

$$p_X(x) = \frac{\pi \exp - \dfrac{\pi}{\sqrt{3}} \left(\dfrac{x-m}{\sigma} \right)}{\sigma\sqrt{3} \left(1 + \exp - \dfrac{\pi}{\sqrt{3}} \dfrac{(x-m)}{\sigma} \right)^2}.$$

Числовые характеристики этих распределений приведены в следующем пункте.

21.3. Характеристическая функция

Определение 21.11. Комплексная функция $\Psi_X(\lambda)$, $\lambda \in \mathbb{R}^n$,

$$\Psi_X(\lambda) = \mathbf{M}\left\{\exp^{i\lambda^\top x}\right\} = \int\limits_{\mathbb{R}^n} \exp^{i\lambda^\top x} \, dF_X(x),$$

называется *характеристической функцией* распределения $F_X(x)$.

Теорема 21.2. *Характеристическая функция однозначно определяет функцию распределения, т.е. если СВ X и СВ Y имеют одну характеристическую функцию $\Psi_X(\lambda) = \Psi_Y(\lambda)$, $\lambda \in \mathbb{R}^n$, то $F_X(x) = F_Y(x)$, $x \in \mathbb{R}^n$.*

Использование аппарата характеристических функций позволяет исследовать зависимость случайных величин в силу следующего утверждения.

Теорема 21.3. *Случайные величины $\{X_i,\ i = 1, \dots, n\}$ независимы тогда и только тогда, когда*

$$\Psi_X(\lambda_1, \dots, \lambda_n) = \prod_{i=1}^{n} \Psi_{X_i}(\lambda_i),$$

где $\Psi_X(\lambda_1, \dots, \lambda_n)$ — характеристическая функция случайного вектора $X = \{X_1, \dots, X_n\}^\top$, а $\Psi_{X_i}(\lambda_i)$ — характеристическая функция СВ X_i, $i = 1, \dots, n$.

21.4. Числовые характеристики случайных величин

Определение 21.12. *Математическим ожиданием* (или *средним*) СВ X, определенной на $\{\Omega, \mathcal{F}, \mathbf{P}\}$, называется число

$$\mathbf{M}\{X\} = m_X = \int\limits_{\Omega} X(\omega)\mathbf{P}\{d\omega\}\,. \qquad (21.3)$$

Математическое ожидание определено, если интеграл Лебега в правой части равенства (21.3) существует.

Определение и свойства интеграла Лебега можно найти, например в [25].

Таким образом, математическое ожидание СВ X есть интеграл Лебега от функции $X(\omega)$ на Ω по вероятностной мере $\mathbf{P}$. Если $\mathbf{P}_X(\cdot)$ — закон распределения СВ X на $\mathcal{B}(\mathbb{R}^1)$, а $F_X(y)$ — соответствующая функция распределения, то $\mathbf{M}\{X\}$ можно вычислить следующим образом:

$$\mathbf{M}\{X\} = \int\limits_{\mathbb{R}^1} y\,\mathbf{P}_X(dy) = \int\limits_{-\infty}^{\infty} y\,dF_X(y), \tag{21.4}$$

причем первый интеграл понимается как интеграл Лебега по мере $\mathbf{P}_X(\cdot)$, а второй — как интеграл Лебега—Стилтьеса. Существование интегралов в правой части (21.4) вытекает из существования математического ожидания, и, наоборот, из существования интегралов следует существование математического ожидания.

Если функция распределения $F_X(x)$ является комбинацией абсолютно-непрерывной и дискретной составляющих, т.е. допускает представление вида $F_X(x) = \int\limits_{-\infty}^{x} p_X(y)dy + \sum\limits_{k:\,a_k \leqslant x} p_k$, где $p_X(y) \geqslant 0$, а $p_k > 0$ — величина скачка в точке разрыва a_k, то (21.4) принимает вид

$$\mathbf{M}\{X\} = \int\limits_{-\infty}^{\infty} yp_X(y)dy + \sum_k p_k a_k.$$

СВ X называется *центрированной*, если $\mathbf{M}\{X\} = 0$.

Основные свойства математического ожидания вытекают из свойств интеграла Лебега.

1. Математические ожидания $\mathbf{M}\{X\}$ и $\mathbf{M}\{|X|\}$ существуют или не существуют одновременно, причем $|\mathbf{M}\{X\}| \leqslant \mathbf{M}\{|X|\}$.

2. $\mathbf{M}\{I_A\} = \int\limits_{\Omega} I_A(\omega)\mathbf{P}\{d\omega\} = \mathbf{P}\{A\}$, где I_A — индикатор события $A \in \mathcal{F}$.

3. Если $\mathbf{M}\{X\}$ существует, то для любой константы λ

$$\mathbf{M}\{\lambda X\} = \lambda \mathbf{M}\{X\}.$$

4. Если $\mathbf{M}\{X\}$ и $\mathbf{M}\{Y\}$ существуют, то

$$\mathbf{M}\{X+Y\} = \mathbf{M}\{X\} + \mathbf{M}\{Y\}.$$

5. Если $X \leqslant Y$, то $\mathbf{M}\{X\} \leqslant \mathbf{M}\{Y\}$.

6. Если $\varphi(x)$ — борелевская функция, то

$$\mathbf{M}\{\varphi(X)\} = \int\limits_{-\infty}^{\infty} \varphi(x) dF_X(x),$$

где математическое ожидание и интеграл существуют или не существуют одновременно.

7. Если $\mathbf{M}\{X\}$ определено, то для каждого $A \in \mathcal{F}$ существует

$$\mathbf{M}\{X I_A\} = \int\limits_A X(\omega)\mathbf{P}(d\omega).$$

8. Если $X \geqslant 0$, $\mathbf{M}\{X\} < \infty$ и $\varepsilon > 0$, то

$$\mathbf{P}\{X \geqslant \varepsilon\} \leqslant \frac{\mathbf{M}\{X\}}{\varepsilon}.$$

9. Если $\mathbf{M}\{|X|^p\} < \infty$, $p > 0$, то выполняется *неравенство Маркова*

$$\mathbf{P}\{|X| \geqslant \varepsilon\} \leqslant \frac{\mathbf{M}\{|X|^p\}}{\varepsilon^p}.$$

Если $X = \{X_1, \ldots, X_n\}^\top$ — n-мерный случайный вектор, то его *математическим ожиданием* называется вектор $\mathbf{M}\{X\} = \{\mathbf{M}\{X_1\}, \ldots, \mathbf{M}\{X_n\}\}^\top$.

Определение 21.13. *Дисперсией* $\mathbf{D}\{X\}$ СВ X называется число

$$\mathbf{D}\{X\} = D_X = \mathbf{M}\{|X - m_X|^2\} = \int\limits_{-\infty}^{\infty} |y - m_X|^2 dF_X(y).$$

Из определения и свойств интеграла Лебега следует:

1) $\mathbf{D}\{X\} \geqslant 0$;

2) $\mathbf{D}\{aX + b\} = |a|^2 \mathbf{D}\{X\}$, если $a, b = \mathrm{const}$;

3) $\mathbf{D}\left\{\sum\limits_{k=1}^{n} X_k\right\} = \sum\limits_{k=1}^{n} \mathbf{D}\{X_k\}$, если $\mathbf{D}\{X_k\} < \infty$, а СВ $\{X_k\}$ — независимы в совокупности;

4) $\mathbf{D}\{X\} = \mathbf{M}\{|X|^2\} - |m_X|^2$.

Кроме того верна следующая теорема.

Теорема 21.4. *Если* $\mathbf{M}\{|X|^2\} < \infty$*, то для любых* $\varepsilon > 0$ *выполнено неравенство Чебышева*

$$\mathbf{P}(|X_n| > \varepsilon) \leqslant \frac{\mathbf{M}\{X_n^2\}}{\varepsilon^2}. \tag{21.5}$$

Определение 21.14. Медианой СВ X называется значение x', для которого

$$\mathbf{P}\{X \leqslant x'\} \geqslant \frac{1}{2}, \quad \mathbf{P}\{X \geqslant x'\} \geqslant \frac{1}{2}.$$

Определение 21.15. *Ковариацией* СВ X и Y называется величина

$$\mathbf{cov}\{X, Y\} = \mathbf{M}\{(X - m_X)(Y - m_Y)\}.$$

Если $X = \{X_1, \ldots, X_n\}^\top$ — n-мерный случайный вектор, то его *ковариационной матрицей* называется матрица $K_X = \mathbf{M}\{XX^\top\} - m_X m_X{}^\top$.

Перечислим важнейшие свойства ковариации:

1) если $\mathbf{M}\{|X|^2\} < \infty$, $\mathbf{M}\{|Y|^2\} < \infty$, то ковариация СВ X, Y существует и удовлетворяет *неравенству Коши–Буняковского*:

$$|\mathbf{cov}\{X, Y\}|^2 \leqslant \mathbf{D}\{X\}\,\mathbf{D}\{Y\};$$

2) если $\mathbf{M}\{X\} = \mathbf{M}\{Y\} = 0$, то $\mathbf{cov}\{X, Y\} = (X, Y)$ — скалярное произведение случайных величин X, Y;

3) если X и Y независимы, то $\mathbf{cov}\{X, Y\} = 0$;

4) $\mathbf{D}\{X\} = \mathbf{cov}\{X, X\}$;

5) $\mathbf{D}\{X + Y\} = \mathbf{D}\{X\} + \mathbf{D}\{Y\} + 2\,\mathbf{cov}\{X, Y\}$.

Если $\mathbf{cov}\{X, Y\} = 0$, то СВ X и Y называются *некоррелированными*.

Кроме ковариации СВ X и Y также рассматривают *коэффициент корреляции* СВ:

$$r_{XY} = \frac{\mathbf{cov}\{X, Y\}}{\sqrt{\mathbf{D}\{X\}\,\mathbf{D}\{Y\}}}. \tag{21.6}$$

Приведем числовые характеристики случайных величин, рассмотренных в конце п. 21.2:

1) биномиальное распределение с параметрами $(N; p)$:

$$m_X = Np, \quad D_X = Npq;$$

2) полиномиальное распределение с параметрами $(N, p_1, \ldots, p_k)$:

$$\mathbf{M}\{X_i\} = Np_i, \quad \mathbf{D}\{X_i\} = Np_i(1 - p_i),\ i = 1, \ldots, k,$$

$$\mathbf{cov}(X_i, X_j) = -Np_ip_j \text{ при } i \neq j,\ i, j = 1, \ldots, k;$$

3) распределение Пуассона с параметром $\lambda > 0$:

$$m_X = D_X = \lambda;$$

4) равномерное распределение на отрезке $[a, b]$:

$$m_X = \frac{a + b}{2}, \quad D_X = \frac{(b - a)^2}{12};$$

5) экспоненциальное распределение с параметром $\lambda > 0$:

$$m_X = \frac{1}{\lambda}, \quad D_X = \frac{1}{\lambda^2};$$

6) распределение Лапласа с параметром $\lambda > 0$:

$$m_X = 0, \quad D_X = \frac{2}{\lambda^2};$$

7) логистическое распределение с параметрами $(m; \sigma^2)$:

$$m_X = m, \quad D_X = \sigma^2.$$

21.5. Гауссовские (нормальные) случайные величины и векторы

Определение 21.16. Функция

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt, \quad x \in \mathbb{R}^1$$

называется *интегралом вероятностей* или *функцией Лапласа.*

Определение 21.17. Случайная величина $X \in \mathbb{R}^1$ называется *гауссовской* или *нормальной* с параметрами $(m; \sigma^2)$, где $\sigma > 0$, если

$$F_X(x) = \mathbf{P}\{X \leqslant x\} = \Phi\left(\frac{x-a}{\sigma}\right). \tag{21.7}$$

Математическое ожидание и дисперсия СВ X равны: $\mathbf{M}\{X\} = m$, $\mathbf{D}\{X\} = \sigma^2$.

Так как функция Лапласа $\Phi(x)$ всюду дифференцируема на $\mathbb{R}^1$, распределение $F_X(x)$ гауссовской СВ имеет плотность

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad \sigma > 0.$$

Для обозначения гауссовской СВ будем писать $X \sim \mathcal{N}(m_X; D_X)$. Вероятность попадания X в произвольный интервал $[\alpha, \beta] \in \mathbb{R}^1$ можно вычислить по следующей известной формуле:

$$\mathbf{P}\{\alpha \leqslant X \leqslant \beta\} = \Phi\left(\frac{\beta - m_X}{\sigma_X}\right) - \Phi\left(\frac{\alpha - m_X}{\sigma_X}\right),$$

где $\sigma_X = \sqrt{D_X} > 0$.

Случайная величина X с распределением $\mathcal{N}(0; 1)$ называется стандартной гауссовской. Ее функция распределения совпадает с $\Phi(x)$.

Для описания *гауссовского случайного вектора* (т.е. упорядоченной системы гауссовских СВ) удобно воспользоваться аппаратом характеристических функций.

Пусть $m_X \in \mathbb{R}^n$ — произвольный вектор, а $K_X \in \mathbb{R}^{n\times n}$ — симметричная неотрицательно определенная матрица ($K_X = K_X{}^\top$, $K_X \geqslant 0$).

Определение 21.18. Случайный вектор $X \in \mathbb{R}^n$ имеет *n-мерное гауссовское распределение* с параметрами $(m_X; K_X)$, если его характеристическая функция $\Psi_X(\lambda)$, $\lambda \in \mathbb{R}^n$ имеет вид

$$\Psi_X(\lambda) = \exp\Big\{i\lambda^\top m_X - \frac{1}{2}\lambda^\top K_X\lambda\Big\}, \tag{21.8}$$

где i — мнимая единица ($i^2 = -1$).

Обозначение: $X \sim \mathcal{N}(m_X; K_X)$.

Параметры m_X и K_X являются соответственно математическим ожиданием и ковариационной матрицей вектора X.

Определение 21.19. Гауссовский вектор X называется *невырожденным*, если матрица K_X — положительно определенная ($K_X > 0$).

Если $K_X > 0$, а $\Delta_X = \det[K_X]$ — определитель матрицы K_X, то X имеет в каждой точке $x \in \mathbb{R}^n$ плотность вероятности следующего вида:

$$p_X(x) = [(2\pi)^n\Delta_X]^{-\frac{1}{2}}\exp\Big\{-\frac{1}{2}(x - m_X)^\top K_X^{-1}(x - m_X)\Big\}. \tag{21.9}$$

Гауссовский вектор X имеет следующие основные свойства.

1. Если $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ — неслучайные матричные параметры, $X \sim \mathcal{N}(m_X; K_X)$, а $Y = AX + b$, то $Y \sim \mathcal{N}(m_Y; K_Y)$, где

$$m_Y = Am_X + b; \quad K_Y = AK_XA^\top. \tag{21.10}$$

Из (21.10) следует, в частности, что любой подвектор гауссовского вектора также является гауссовским. Например, если X_i — i-я компонента вектора X, то $X_i \sim \mathcal{N}(m_i; \sigma_i^2)$, где m_i — i-я компонента m_X, а σ_i^2 — i-й диагональный элемент матрицы K_X.

2. Если $X \sim \mathcal{N}(m_X; K_X)$, причем компоненты $\{X_1, \ldots, X_n\}$ вектора X некоррелированы (т.е. K_X — диагональная матрица), то случайные величины $\{X_1, \ldots, X_n\}$ независимы в совокупности. Наоборот, произвольная совокупность $\{X_1, \ldots, X_n\}$ независимых гауссовских случайных величин образует гауссовский случайный вектор.

21.6. Сходимость последовательностей случайных величин

Пусть $\{X_1, \ldots, X_n, \ldots\}$ — последовательность произвольных случайных величин (заданных на одном вероятностном пространстве $\{\Omega, \mathcal{F}, \mathbf{P}\}$).

Определение 21.20. $X_n \to X$ *по вероятности*, если для любого $\varepsilon > 0$ $\lim\limits_{n\to\infty} \mathbf{P}\left(|X_n - X| > \varepsilon\right) = 0$.

Определение 21.21. $X_n \to X$ *в среднем квадратическом*, если $\lim\limits_{n\to\infty} \mathbf{M}\bigl\{|X_n - X|^2\bigr\} = 0$.

Определение 21.22. $X_n \to X$ *почти наверное (с вероятностью 1)*, если $\mathbf{P}\Bigl(\omega \in \Omega\colon \lim\limits_{n\to\infty} X_n(\omega) = X(\omega)\Bigr) = 1$.

Указанные виды сходимости будем обозначать соответственно $X_n \xrightarrow{\mathbf{P}} X$, $X_n \xrightarrow{\text{с.к.}} X$, $X_n \xrightarrow{\text{п.н.}} X$, $n \to \infty$.

Перечислим некоторые свойства сходящихся последовательностей.

1. Если $X_n \xrightarrow{\text{п.н.}} X$ или $X_n \xrightarrow{\text{с.к.}} X$, $n \to \infty$, то $X_n \xrightarrow{\mathbf{P}} X$, $n \to \infty$.

2. Если $X_n \xrightarrow{\text{п.н.}} X$, $Y_n \xrightarrow{\text{п.н.}} Y$, $n \to \infty$, $a, b = \text{const}$, тогда $aX_n + bY_n \xrightarrow{\text{п.н.}} aX + bY$, $n \to \infty$.

3. Пусть $g(x)$ — произвольная борелевская функция, заданная на прямой $\mathbb{R}^1$, а A — множество точек разрыва функции $g(x)$. Если $X_n \xrightarrow{\text{п.н.}} X$, $n \to \infty$, причем $\mathbf{P}(X \in A) = 0$, то $g(X_n) \xrightarrow{\text{п.н.}} g(X)$, $n \to \infty$.

4. Если $X_n \sim \mathcal{N}(m_n; D_n)$ и $X_n \xrightarrow{\text{с.к.}} X$, $n \to \infty$, то $X \sim \mathcal{N}(m_X; D_X)$, где $m_X = \lim\limits_{n\to\infty} m_n$, $D_X = \lim\limits_{n\to\infty} D_n$, причем пределы существуют и конечны.

5. Свойства 2 и 3 справедливы для сходимости по вероятности, а свойство 2 — для с.к.-сходимости.

Пусть также B — любое борелевское множество действительной оси $\mathbb{R}^1$, а ∂B — его граница.

Определение 21.23. Последовательность $\{X_n,\ n = 1, 2, \ldots\}$ *сходится по распределению* к случайной величине X, если для любого B такого, что $\mathbf{P}\left(X \in \partial B\right) = 0$, выполнено

$$\mathbf{P}(X_n \in B) \to \mathbf{P}(X \in B), \quad n \to \infty. \tag{21.11}$$

Сходимость по распределению, которую также называют *слабой сходимостью*, будем обозначать

$$X_n \xrightarrow{d} X, \quad n \to \infty. \tag{21.12}$$

Пусть $F_n(x)$ и $F_X(x)$ — функции распределения соответственно X_n и X.

Теорема 21.5. *Следующие три утверждения эквивалентны:*

1) $X_n \xrightarrow{d} X$, $n \to \infty$;

2) $F_n(x) \to F_X(x)$, $n \to \infty$ *для любой точки непрерывности функции* $F_X(x)$;

3) $\mathbf{M}\{g(X_n)\} \to \mathbf{M}\{g(X)\}$, $n \to \infty$ *для любой непрерывной и ограниченной на* $\mathbb{R}^1$ *функции* $g(x)$.

Теорема 21.6. *Если* $X_n \xrightarrow{\mathbf{P}} X$, $n \to \infty$, *то* $X_n \xrightarrow{d} X$, $n \to \infty$.

Таким образом, слабая сходимость случайной последовательности следует из сходимости по вероятности, а следовательно, из сходимости почти наверное и сходимости в среднем квадратическом.

В одном важном частном случае сходимость по распределению и сходимость по вероятности эквивалентны: если $X_n \xrightarrow{d} a$, $n \to \infty$, где $a = \mathrm{const}$, то $X_n \xrightarrow{\mathbf{P}} a$, $n \to \infty$.

Для установления факта слабой сходимости обычно используется следующее утверждение.

Теорема 21.7. *Пусть* $\Psi_n(\lambda)$ *и* $\Psi(\lambda)$ — *характеристические функции соответственно* X_n *и* X. *Пусть также для любого* $\lambda \in \in \mathbb{R}^1$

$$\Psi_n(\lambda) \to \Psi(\lambda), \quad n \to \infty.$$

Тогда $X_n \xrightarrow{d} X$, $n \to \infty$.

Для исследования сходимости последовательностей имеют важное значение следующие вспомогательные утверждения.

Теорема 21.8 (Борель—Кантелли). *Пусть задана бесконечная последовательность случайных событий* $\{A_1, A_2, \ldots, A_n, \ldots\}$, *а* $B = \bigcap\limits_{n\geqslant 1} \bigcup\limits_{k\geqslant n} A_k$ — *событие, состоящее в том, что произойдет бесконечно много событий* $\{A_i\}$. *Тогда:*

1) *если* $\sum\limits_{k=1}^{\infty} \mathbf{P}(A_k) < \infty$, *то* $\mathbf{P}(B) = 0$;

2) *если* $\{A_1, A_2, \ldots, A_n, \ldots\}$ *независимы и* $\sum\limits_{k=1}^{\infty} \mathbf{P}(A_k) = \infty$, *то* $\mathbf{P}(B) = 1$.

21.7. Центральная предельная теорема

Пусть $\{X_n,\ n = 1, 2, \ldots\}$ — последовательность случайных величин. Особое значение имеет случай, когда пределом последовательности СВ является гауссовская случайная величина

Определение 21.24. Последовательность $\{X_n, n = 1, 2, \ldots\}$ называется *асимптотически нормальной* с параметрами $(m; \sigma^2)$, если

$$X_n \xrightarrow{d} X, \quad n \to \infty, \text{ где } X \sim \mathcal{N}(m; \sigma^2). \tag{21.13}$$

Из определения 21.24 и теоремы 21.5 следует, что для любого $x \in \in \mathbb{R}^1$

$$F_n(x) \to \Phi\left(\frac{x-m}{\sigma}\right), \quad n \to \infty. \tag{21.14}$$

Практическое значение (21.14) состоит в том, что СВ X_n можно считать нормально распределенной с параметрами $(m;\sigma^2)$, если n достаточно велико.

Пусть теперь $\{X_k,\ k=1,2,\ldots\}$ — последовательность независимых случайных величин с параметрами

$$\mathbf{M}\{X_k\}=a;\quad \mathbf{D}\{X_k\}=\sigma^2>0.$$

Рассмотрим случайную *последовательность сумм* этих величин:

$$X_n=\sum_{k=1}^{n}X_k,\quad n=1,2,\ldots \tag{21.15}$$

Очевидно, $\mathbf{M}\{X_n\}=na$, $\mathbf{D}\{X_n\}=n\sigma^2$.

Введем *стандартизованную сумму*

$$\widetilde{X}_n=\frac{X_n-na}{\sigma\sqrt{n}},\quad n=1,2,\ldots \tag{21.16}$$

Нетрудно проверить, что $\mathbf{M}\Big\{\widetilde{X}_n\Big\}=0$, $\mathbf{D}\Big\{\widetilde{X}_n\Big\}=1$.

Следующее утверждение, называемое *центральной предельной теоремой* (ЦПТ), имеет особое значение для математической статистики.

Теорема 21.9 (ЦПТ для одинаково распределенных слагаемых). *Пусть случайные величины $\{X_k,\ k=1,2,\ldots\}$ одинаково распределены, тогда последовательность $\{\widetilde{X}_n,\ n=1,2,\ldots\}$ асимптотически нормальна с параметрами $(0;1)$.*

Следствие 21.1. *Для любых чисел $a\leqslant b$ выполнено*

$$\mathbf{P}(a\leqslant\widetilde{X}_n\leqslant b)\to\Phi(b)-\Phi(a),\quad n\to\infty. \tag{21.17}$$

Следствие 21.2 (Интегральная теорема Муавра—Лапласа). *Пусть X_n — число успехов в серии из n испытаний Бернулли, а p — вероятность успеха в одном испытании. Тогда при $n\to\infty$*

$$\frac{X_n-np}{\sqrt{np(1-p)}}\xrightarrow{d}X\sim\mathcal{N}(0;1). \tag{21.18}$$

Если слагаемые $\{X_k,\ k=1,2,\ldots\}$ распределены не одинаково, то утверждение, аналогичное ЦПТ, останется в силе при некоторых дополнительных ограничениях.

Пусть $\mathbf{M}\{X_k\}=a_k$, $\mathbf{D}\{X_k\}=\sigma_k^2$, $\mathbf{M}\Big\{|X_k-a_k|^3\Big\}=c_k^3$. Обозначим $A_n=\mathbf{M}\{X_n\}=\sum\limits_{k=1}^{n}a_k$, $D_n^2=\mathbf{D}\{X_n\}=\sum\limits_{k=1}^{n}\sigma_k^2$, $C_n^3=\sum\limits_{k=1}^{n}c_k^3$.

Теорема 21.10 (Ляпунов). *Пусть* A_n, D_n, C_n *конечны при всех* $n \geqslant 1$, *причем* $\frac{C_n}{D_n} \to 0$, $n \to \infty$. *Тогда последовательность* $\{\widetilde{X}_n, n = 1, 2, \ldots\}$, *где* $\widetilde{X}_n = \frac{X_n - A_n}{D_n}$, *асимптотически нормальна с параметрами* $(0; 1)$.

21.8. Закон больших чисел

Во всех приводимых ниже утверждениях (теоремы 21.11 —21.15) предполагается, что $\{X_k, k = 1, 2, \ldots\}$ — последовательность *независимых СВ*.

Определение 21.25. *Выборочным средним* называется СВ

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k, \quad n = 1, 2, \ldots .$$

Законом больших чисел называется совокупность утверждений о поведении последовательности $\{\overline{X}_n, n = 1, 2, \ldots\}$ выборочных средних при $n \to \infty$.

Будем далее обозначать $\mathbf{M}\{X_k\} = a_k$, $\mathbf{D}\{X_k\} = D_k = \sigma_k^2$.

Теорема 21.11. *Если* $\{X_k, k = 1, 2, \ldots\}$ *одинаково распределены и* $\mathbf{M}\{X_k\} = a$ *конечно, то* $\overline{X}_n \xrightarrow{\text{п.н.}} a$, $n \to \infty$.

Теорема 21.11 называется *«Усиленный закон больших чисел А.Н. Колмогорова»*.

Теорема 21.12. *Если* $\mathbf{M}\{X_k\} = a$, $\mathbf{D}\{X_k\} = D_k < \infty$, *причем*

$$\sum_{k=1}^{\infty} \frac{D_k}{k^2} < \infty, \tag{21.19}$$

то $\overline{X}_n \xrightarrow{\text{п.н.}} a$, $n \to \infty$ *и* $\overline{X}_n \xrightarrow{\text{с.к.}} a$, $n \to \infty$.

Следствие 21.3. *Если* $a_k = a$, $D_k \leqslant \overline{D} < \infty$ $\forall k = 1, 2, \ldots$, *то утверждение теоремы 21.12 справедливо.*

Пусть теперь $\{a_k\}$ не одинаковы. Обозначим $\overline{a}_n = \frac{1}{n} \sum_{k=1}^{n} a_k$.

Следствие 21.4. *Если выполнено условие* (21.19), *то*

$$\overline{X}_n - \overline{a}_n \xrightarrow{\text{п.н.}} 0,\ n \to \infty; \quad \overline{X}_n - \overline{a}_n \xrightarrow{\text{с.к.}} 0,\ n \to \infty.$$

Следующее утверждение показывает, как будет вести себя последовательность выборочных средних, если СВ $\{X_k, k = 1, 2, \ldots\}$ не имеют конечного среднего (например, X_k имеет распределение Коши).

Теорема 21.13. *Пусть* $\{X_k,\ k = 1, 2, \ldots\}$ *одинаково распределены, но* $\mathbf{M}\{X_k\}$ *не существует ни при каких* $k = 1, 2, \ldots$. *Тогда последовательность* $\{\overline{X}_n,\ n = 1, 2, \ldots\}$ *с вероятностью* 1 *неограниченна, т.е.* $\mathbf{P}\left(\sup\limits_{n\geqslant 1} |\overline{X}_n| > C\right) = 1$ *для любого* $C > 0$.

Весьма часто последовательность выборочных средних асимптотически нормальна (см. определение 21.24).

Теорема 21.14. *Пусть* $\{X_k,\ k = 1, 2, \ldots\}$ — *последовательность независимых одинаково распределенных СВ, причем* $\mathbf{M}\{X_k^2\} < \infty$, $k = 1, 2, \ldots$. *Тогда*

$$\sqrt{n}(\overline{X}_n - a) \xrightarrow{d} X \sim \mathcal{N}(0; \sigma^2), \quad n \to \infty, \tag{21.20}$$

где $\sigma = \sqrt{\mathbf{D}\{X_k\}}$, $a = \mathbf{M}\{X_k\}$, $k = 1, 2, \ldots$.

Утверждение (21.20) позволяет пользоваться при $n \gg 1$ приближенным соотношением

$$\overline{X}_n \sim \mathcal{N}\left(a; \frac{\sigma^2}{n}\right). \tag{21.21}$$

Для случая, когда СВ $\{X_k,\ k = 1, 2, \ldots\}$ распределены неодинаково, аналог теоремы 21.14 можно получить, используя теорему 21.10 Ляпунова.

Теорема 21.15. *Пусть выполнены условия теоремы 21.10, тогда*

$$\frac{n}{D_n}\left(\overline{X}_n - \overline{a}_n\right) \xrightarrow{d} X \sim \mathcal{N}(0; 1), \quad n \to \infty.$$

Аналогично (21.21) в данном случае можно считать, что

$$\overline{X}_n \sim \mathcal{N}\left(\overline{a}_n; \frac{D_n^2}{n^2}\right). \tag{21.22}$$

22. Статистические таблицы

Функция $\Phi_0(x) = \dfrac{1}{\sqrt{2\pi}} \int\limits_0^x e^{-t^2/2}\, dt$

Таблица 22.1

x	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	,0039	,0079	,0119	,0159	,0199	,0239	,0279	,0318	,0358
0,1	,0398	,0438	,0477	,0517	,0556	,0596	,0635	,0674	,0714	,0753
0,2	,0792	,0831	,0870	,0909	,0948	,0987	,1025	,1064	,1102	,1140
0,3	,1179	,1217	,1255	,1293	,1330	,1368	,1405	,1443	,1480	,1517
0,4	,1554	,1591	,1627	,1664	,1700	,1736	,1772	,1808	,1843	,1879
0,5	,1914	,1949	,1984	,2019	,2054	,2088	,2122	,2156	,2190	,2224
0,6	,2257	,2290	,2323	,2356	,2389	,2421	,2453	,2485	,2517	,2549
0,7	,2580	,2611	,2642	,2673	,2703	,2733	,2763	,2793	,2823	,2852
0,8	,2881	,2910	,2938	,2967	,2995	,3023	,3051	,3078	,3105	,3132
0,9	,3159	,3185	,3212	,3228	,3263	,3289	,3314	,3339	,3364	,3389
1,0	,3413	,3437	,3461	,3485	,3508	,3531	,3554	,3576	,3599	,3621
1,1	,3643	,3665	,3686	,3707	,3728	,3749	,3769	,3790	,3810	,3829
1,2	,3849	,3868	,3887	,3906	,3925	,3943	,3961	,3979	,3997	,4014
1,3	,4032	,4049	,4065	,4082	,4098	,4114	,4130	,4146	,4162	,4177
1,4	,4192	,4207	,4222	,4236	,4250	,4264	,4278	,4292	,4305	,4318
1,5	,4331	,4344	,4357	,4369	,4382	,4394	,4406	,4417	,4429	,4440
1,6	,4452	,4463	,4473	,4484	,4495	,4505	,4515	,4525	,4535	,4544
1,7	,4554	,4563	,4572	,4581	,4590	,4599	,4608	,4616	,4624	,4632
1,8	,4640	,4648	,4656	,4663	,4671	,4678	,4685	,4692	,4699	,4706
1,9	,4712	,4719	,4725	,4732	,4738	,4744	,4750	,4755	,4761	,4767
2,0	,4772	,4777	,4783	,4788	,4793	,4798	,4803	,4807	,4812	,4816
2,1	,4821	,4825	,4830	,4834	,4838	,4842	,4846	,4850	,4853	,4857
2,2	,4861	,4864	,4867	,4871	,4874	,4877	,4880	,4884	,4887	,4889
2,3	,4892	,4895	,4898	,4901	,4903	,4906	,4908	,4911	,4913	,4915
2,4	,4918	,4920	,4922	,4924	,4926	,4928	,4930	,4932	,4934	,4936
2,5	,4937	,4939	,4941	,4943	,4944	,4946	,4947	,4949	,4950	,4952
2,6	,4953	,4954	,4956	,4957	,4958	,4959	,4960	,4962	,4963	,4964
2,7	,4965	,4966	,4967	,4968	,4969	,4970	,4971	,4972	,4972	,4973
2,8	,4974	,4975	,4976	,4976	,4977	,4978	,4978	,4979	,4980	,4980
2,9	,4981	,4981	,4982	,4983	,4983	,4984	,4984	,4985	,4985	,4986

Столбцы табл. 22.1 организованы по сотым долям аргумента x. Для значений $x \geqslant 3$ можно считать, что функция $\Phi_0(x)$ равна примерно 0,5.

В табл. 22.2 приведены квантили u_α уровня α распределения $\mathcal{N}(0; 1)$.

Таблица 22.2

α	0,6	0,8	0,9	0,95	0,975	0,99	0,995	0,9995
u_α	0,253	0,842	1,282	1,645	1,960	2,326	2,576	3,291

Для $\alpha \in (0; 0{,}5)$ квантили u_α определяются из табл. 22.2 по формуле $u_\alpha = -u_\beta$, где $\beta = 1 - \alpha \in (0{,}5; 1)$.

В табл. 22.3 приведены квантили k_α уровня α хи-квадрат распределения $\mathcal{H}_r$

Таблица 22.3

r	α								
	0,05	0,1	0,2	0,3	0,5	0,7	0,8	0,9	0,95
1	0,00	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84
2	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99
3	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,82
4	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49
5	1,15	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07
6	1,65	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59
7	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07
8	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51
9	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92
10	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31
11	4,58	5,58	6,99	8,15	10,34	12,90	14,63	17,28	19,68
12	5,23	6,30	7,81	9,03	11,34	14,01	15,81	18,55	21,03
13	5,89	7,04	8,63	9,93	12,34	15,12	16,98	19,81	22,36
14	6,57	7,79	9,47	10,82	13,34	16,22	18,15	21,06	23,69
15	7,26	8,55	10,31	11,72	14,34	17,32	19,31	22,31	25,00
16	7,96	9,31	11,15	12,62	15,34	18,42	20,47	23,54	26,29
17	8,67	10,08	12,00	13,53	16,34	19,51	21,62	24,78	27,60
18	9,39	10,86	12,86	14,44	17,34	20,60	22,76	25,59	28,87
19	10,11	11,65	13,72	15,35	18,34	21,70	23,90	27,20	30,14
20	10,85	12,44	14,58	16,27	19,34	22,80	25,04	28,41	31,41
21	11,59	13,24	15,44	17,18	20,30	23,90	26,17	29,61	32,67
22	12,34	14,04	16,31	18,10	21,30	24,90	27,30	30,81	33,92
23	13,09	14,85	17,19	19,02	22,30	26,00	28,43	32,01	35,17
24	13,85	15,66	18,06	19,94	23,30	27,10	29,55	33,20	36,42
25	14,61	16,47	18,94	20,90	24,30	28,20	30,78	34,38	37,65
26	15,38	17,29	19,82	21,80	25,30	29,20	31,80	35,56	38,89
27	16,15	18,11	20,70	22,70	26,30	30,30	32,91	36,74	40,11
28	16,93	18,94	21,60	23,60	27,30	31,40	34,03	37,92	41,34
29	17,71	19,77	22,50	24,60	28,30	32,50	35,14	39,09	42,56
30	18,49	20,60	23,40	25,50	29,30	33,50	36,25	40,26	43,77

В табл. 22.4 приведены квантили t_α уровня α распределения Стьюдента $\mathcal{T}_r$.

Таблица 22.4

r	α							
	0,6	0,8	0,9	0,95	0,975	0,99	0,995	0,9995
1	0,325	1,376	3,078	6,314	12,706	31,821	63,657	636,619
2	0,289	1,061	1,886	2,920	4,303	6,965	9,925	31,598
3	0,277	0,978	1,638	2,353	3,182	4,541	5,841	12,941
4	0,271	0,941	1,533	2,132	2,776	3,747	4,604	8,610
5	0,267	0,920	1,476	2,015	2,571	3,365	4,032	6,859
6	0,265	0,906	1,440	1,943	2,447	3,143	3,707	5,959
7	0,263	0,896	1,415	1,895	2,365	2,998	3,499	5,405
8	0,262	0,889	1,397	1,860	2,306	2,896	3,355	5,041
9	0,261	0,883	1,383	1,833	2,262	2,821	3,250	4,781
10	0,260	0,879	1,372	1,812	2,228	2,764	3,169	4,587
11	0,260	0,876	1,363	1,796	2,201	2,718	3,106	4,437
12	0,259	0,873	1,356	1,782	2,179	2,681	3,055	4,318
13	0,259	0,870	1,350	1,771	2,160	2,650	3,012	4,221
14	0,258	0,868	1,345	1,761	2,145	2,624	2,977	4,140
15	0,258	0,866	1,341	1,753	2,131	2,602	2,947	4,073
16	0,258	0,865	1,337	1,746	2,120	2,583	2,921	4,015
17	0,257	0,863	1,333	1,740	2,110	2,567	2,898	3,965
18	0,257	0,862	1,330	1,734	2,101	2,552	2,878	3,922
19	0,257	0,861	1,328	1,729	2,093	2,539	2,861	3,883
20	0,257	0,860	1,325	1,725	2,086	2,528	2,845	3,850
21	0,257	0,859	1,323	1,721	2,080	2,518	2,831	3,819
22	0,256	0,858	1,321	1,717	2,074	2,508	2,819	3,792
23	0,256	0,858	1,319	1,714	2,069	2,500	2,807	3,767
24	0,256	0,857	1,318	1,711	2,064	2,492	2,797	3,745
25	0,256	0,856	1,316	1,708	2,060	2,485	2,787	3,725
26	0,256	0,856	1,315	1,706	2,056	2,479	2,779	3,707
27	0,256	0,855	1,314	1,703	2,052	2,473	2,771	3,690
28	0,256	0,855	1,313	1,701	2,048	2,467	2,763	3,674
29	0,256	0,854	1,311	1,699	2,045	2,462	2,756	3,659
30	0,256	0,854	1,310	1,697	2,042	2,457	2,750	3,646
40	0,255	0,851	1,303	1,684	2,021	2,423	2,704	3,551
60	0,254	0,848	1,296	1,671	2,000	2,390	2,660	3,460
120	0,254	0,845	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,253	0,842	1,282	1,645	1,960	2,326	2,576	3,291

В табл. 22.5 приведены значения d_l и d_u критерия Дарбина—Уотсона для уровня значимости $\alpha = 0{,}05$, числа оцениваемых параметров p (без учета постоянного члена), числа наблюдений n. Значения приведены, начиная с числа наблюдений $n = 15$. Это сделано, поскольку никакая последовательность, содержащая, скажем, лишь три наблюдения, не может служить достаточным основанием для того, чтобы выявить наличие или отсутствие автокорреляции.

Таблица 22.5

n	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

СПИСОК ЛИТЕРАТУРЫ

1. *Айвазян С.А.* Прикладная статистика. Основы эконометрики. М.: ЮНИТИ-ДАНА, 2001.

2. *Андерсон Т.* Введение в многомерный статистический анализ. М.: Физматлит, 1963.

3. *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976.

4. *Аптон Г.* Анализ таблиц сопряженности. М.: Финансы и статистика, 1982.

5. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ. М.: Мир, 1982.

6. *Ашманов С.А.* Математические методы в экономике. М.: Изд-во Моск. ун-та, 1980.

7. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. М.: Наука, 1983.

8. *Боровков А.А.* Математическая статистика. М.: Наука, 1984.

9. *Ван дер Варден Б.Л.* Математическая статистика. М.: Изд-во иностр. лит., 1960.

10. *Васильев Ф.П.* Численные методы решения экстремальных задач. М.: Наука, 1988.

11. *Ватутин В.А. и др.* Теория вероятностей и математическая статистика в задачах. М.: Агар, 2003.

12. *Гаек Я., Шидак З.* Теория ранговых критериев. М.: Наука, 1971.

13. *Демиденко Е.З.* Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981.

14. *Джонстон Дж.* Эконометрические методы. М.: Статистика, 1980.

15. Практикум по эконометрике / под ред. И.И. Елисеевой. М.: Финансы и статистика, 2002.

16. *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. М.: Высшая школа, 1992.

17. *Катышев П.К., Магнус Я.Р., Пересецкий А.А.* Эконометрика. Начальный курс. М.: Дело, 2004.

18. *Кендалл М.* Ранговые корреляции. М.: Статистика, 1975.

19. *Кендалл М.* Временные ряды. М.: Финансы и статистика, 1981.

20. *Кендалл М., Стюарт А.* Теория распределений. М.: Наука, 1966.

21. *Кендалл М., Стюарт А.* Статистические выводы и связи. М.: Наука, 1973.

22. *Кендалл М., Стюарт А.* Многомерный статистический анализ и временные ряды. М.: Наука, 1976.

23. *Кибзун А.И., Горяинова Е.Р., Наумов А.В.* Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами. Учебное пособие. 3-е изд., перераб. и доп. М.: Физматлит, 2007.

24. *Кобзарь А.И.* Прикладная математическая статистика. М.: Физматлит, 2006.

25. *Колмогоров А.Н., Фомин С.В* Элементы теории функция и функционального анализа. М.: Наука, 1972.

26. *Королюк В.С. и др.* Справочник по теории вероятностей и математической статистике. М.: Наука, 1985.

27. *Крамер Г.* Математические методы статистики. М.: Мир, 1976.

28. *Ликеш И., Ляга Й.* Основные таблицы математической статистики. М.: Финансы и статистика, 1985.

29. *Кушко В.Л., Мудров В.И.* Методы обработки измерений: квазиправдоподобные оценки. М.: Радио и связь, 1983.

30. *Маленво Э.* Статистические методы в эконометрии. М.: Статистика, 1976.

31. *Пугачев В.С.* Теория вероятностей и математическая статистика. М.: Наука, 2002.

32. *Севастьянов Б.А.* Курс теории вероятностей и математической статистики. М.: Наука, 1982.

33. *Рао С.Р.* Линейные статистические методы и их применения. М.: Наука, 1968.

34. *Себер Дж.* Линейный регрессионный анализ. М.: Мир, 1980.

35. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере. М.: ИНФРА-М, 2003.

36. *Фишер Р.А.* Статистические методы для исследователей. М.: Госстатгиз, 1958.

37. *Хардле В.* Прикладная непараметрическая регрессия. М.: Мир, 1993.

38. *Хеттманспергер Т.* Статистические выводы, основанные на рангах. М.: Финансы и статистика, 1987.

39. *Холлендер М., Вульф Д.* Непараметрические методы статистики. М.: Финансы и статистика, 1983.

40. *Ширяев А.Н.* Вероятность. М.: Наука, 1989.

41. *Ширяев А.Н.* Основы стохастической финансовой математики. Т. 1. М.: Фазис, 1998.

42. *Bartlett M.S.* Properties of Sufficiency and Statistical Tests. Proceedings of the Royal Society of London. Series A. 1937. Vol. 160. P. 268–282.

43. *Bassett G., Koenker R.* Regression Quantiles // Econometrica. 1978. Vol. 46. P. 33–50.

44. *Durbin J.* Estimation of Parameters in Time-Series Regression Models // Journal Royal Statistical Society. Series B. 1960. Vol. 22. P. 139–153.

45. *Durbin J., Watson G.S.* Testing for Serial Correlation in Least Squares Regression // Biometrika. 1950. Vol. 37. P. 409–428; Vol. 38. P. 159–178.

46. *Friedman M., Byers S.O., Rosenman R.H., Neuman R.* Coronary-prone Individuals (Type A Behavior Pattern) Growth Hormone Responeses // Journal American Medical Association. 1971. Vol. 217. P. 929–932.

47. *Goldfield S.M., Quandt R.E.* Some Tests for Homoscedasticity // Journal American Statistical Association. 1965. Vol. 60. P. 539–547.

48. *Goldberger A.* A Course in Econometrics. Cambridge, MA: Harvard University Press, 1990.

49. *Noether G.E.* On a Theorem of Pitman // Ann. Math. Stat. 1955. Vol. 26. P. 64–58.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Учебное издание

Горяинова Елена Рудольфовна
Панков Алексей Ростиславович
Платонов Евгений Николаевич

Прикладные методы анализа статистических данных

Зав. редакцией *Е.А. Бережнова*
Редактор *З.А. Басырова*
Художественный редактор *А.М. Павлов*
Компьютерная верстка и графика: *Е.Н. Платонов*
Корректор *С.М. Хорошкина*

Подписано в печать 12.09.2012. Формат 60×88 1/16
Усл. печ. л. 18,9. Уч.-изд. л. 15,2
Тираж 1000 экз. Изд. № 1409

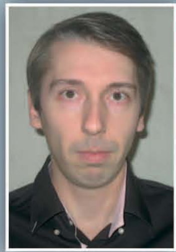
Национальный исследовательский университет
«Высшая школа экономики»
101000, Москва, ул. Мясницкая, 20
Тел./факс: (499) 611-15-52

Горяинова Елена Рудольфовна

Окончила механико-математический факультет МГУ им. М.В. Ломоносова, кандидат физико-математических наук, доцент кафедры высшей математики на факультете экономики НИУ ВШЭ. Область научных интересов: непараметрические методы статистического анализа данных.

Панков Алексей Ростиславович

Окончил факультет прикладной математики и физики МАИ, доктор физико-математических наук, профессор кафедры «Теория вероятностей» МАИ. Автор более 130 научных работ в области системного анализа, управления и обработки информации.



Платонов Евгений Николаевич

Окончил в 1999 году факультет прикладной математики и физики МАИ, кандидат физико-математических наук, доцент кафедры «Теория вероятностей» МАИ и кафедры высшей математики на факультете экономики НИУ ВШЭ. Основные научные интересы: теория статистических решений, минимаксная оптимизация, математическая экономика.

ISBN 978-5-7598-0866-4