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 $(x_1, \dots, x_n)$, т.е. $x_{(1)} \leqslant x_{(2)} \leqslant \leqslant \dots \leqslant x_{(n)}$.

Определение 1.6. СВ $X_{(k)}$, реализацией которой для каждой z_n является число $x_{(k)}$, называется *k-й порядковой статистикой*, $k = 1, \dots, n$. Случайный вектор $Z_{(n)} = \{X_{(1)}, \dots, X_{(n)}\}^\top$ 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) $\overline{S}_n^2 = \dfrac{1}{n}\sum\limits_{k=1}^{n}(X_k - \overline{X}_n)^2$ называется *выборочной дисперсией.*

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

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

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

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: $m_X = \mathbf{M}\{X_k\}$, $D_X = \mathbf{D}\{X_k\}$, $\nu_r = \mathbf{M}\{(X_k)^r\}$, $\mu_r = \mathbf{M}\{(X_k - m_X)^r\}$, $r = 2, 3, \dots$ существуют и конечны. Тогда справедливо следующее утверждение.

Теорема 1.1. *При неограниченном увеличении объема выборки n выборочные моменты $\overline{\nu}_r(n)$ и $\overline{\mu}_r(n)$, $r = 1, 2, \dots$ 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 $\overline{S}_n^2 \xrightarrow{\text{п.н.}} D_X$, $n \to \infty$.*

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 $A_X = \{X \leqslant x\}$, 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* $\mathbf{M}\left\{|X|^4 + |Y|^4\right\} < \infty$, *то*

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

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

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:

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

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 $X \sim \mathcal{N}(67{,}46; 6, 60)$:

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

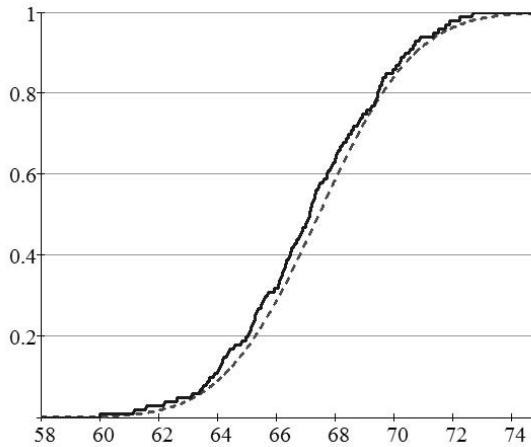

Рис. 1.1. — $\widehat{F}_{100}(x)$ и --- $\widehat{F}_n(x)$

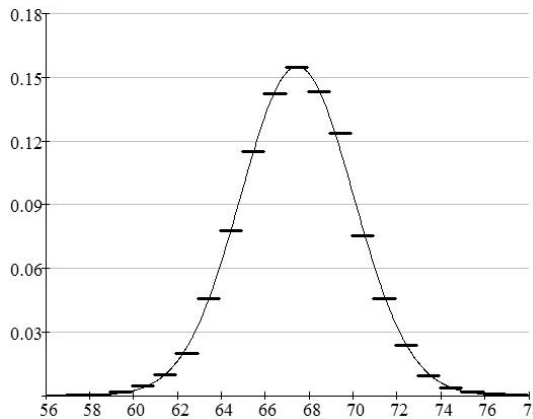

Рис. 1.2. Гистограмма $\widehat{p}_n(x)$ и функция $\widetilde{p}(x)$

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

Пример 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 $F_{(n)}(x) = F^n(x)$.

Р е ш е н и е. 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

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

П р и м е р 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 $n \to \infty$.

Р е ш е н и е. 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. $\blacksquare$

П р и м е р 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.

Р е ш е н и е. 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 $\mathbf{M}\{\xi_k\} = = \mathbf{M}\left\{\left((X_k - m_X) - (\overline{X}_n - m_X)\right)^2\right\} = \mathbf{M}\left\{\left(\overset{\circ}{X}_k - \left(\frac{1}{n}\sum\limits_{i=1}^{n}\overset{\circ}{X}_i\right)\right)^2\right\} = = \mathbf{M}\left\{\overset{\circ}{X}_k^2\right\} - \frac{2}{n}\sum\limits_{i=1}^{n}\mathbf{M}\left\{\overset{\circ}{X}_k\overset{\circ}{X}_i\right\} + \frac{1}{n^2}\sum\limits_{i,j=1}^{n}\mathbf{M}\left\{\overset{\circ}{X}_i\overset{\circ}{X}_j\right\}$. 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 $\mathbf{M}\{\overline{S}_n^2\} = \frac{1}{n}\sum\limits_{k=1}^{n}\mathbf{M}\{\xi_k\} = \frac{n-1}{n}D_X$. 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 $d_n = \operatorname{tr}[K_n]$, где $K_n = \mathbf{cov}(\Delta\widehat{\theta}_n, \Delta\widehat{\theta}_n)$ — ковариационная матрица ошибки $\Delta\widehat{\theta}_n$, $\operatorname{tr}[A]$ — 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 $n \geqslant 1$ $\mathbf{M}\Big\{\Delta\widehat{\theta}_n\Big\} = \mathbf{M}\{\Delta\overline{X}_n\} = \mathbf{M}\{\overline{\varepsilon}_n\} = \dfrac{1}{n}\sum\limits_{k=1}^{n}\mathbf{M}\{\varepsilon_k\} = 0$, т.е. оценка $\overline{X}_n$ — несмещенная. ■

Пример 2.2. Пусть в условиях примера 2.1 СВ $\{\varepsilon_k, k = 1, 2, \dots\}$ 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 $\dfrac{\sqrt{n}}{\sigma}\Delta\overline{X}_n = \dfrac{\sqrt{n}}{\sigma}\overline{\varepsilon}_n \xrightarrow{d} \xi \sim \mathcal{N}(0;1)$, $n \to \infty$. Таким образом, $C_n\Delta\overline{X}_n \xrightarrow{d} \xi \sim \mathcal{N}(0;1)$, $n \to \infty$, если $\left\{C_n = \dfrac{\sqrt{n}}{\sigma},\ n = 1, 2, \dots\right\}$. ■

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

Решение. По условию $F(x) = \mathbf{P}(X \leqslant x) = \dfrac{x}{\theta}$, $x \in [0;\theta]$. Из примера 1.2 следует, что $F_{(n)}(x) = \mathbf{P}(X_{(n)} \leqslant x) = F^n(x) = \dfrac{x^n}{\theta^n}$, $x \in [0;\theta]$. Тогда $\mathbf{M}\Big\{X_{(n)}\Big\} = \int\limits_0^\theta x dF_{(n)}(x) = \int\limits_0^\theta x\dfrac{nx^{n-1}}{\theta^n}dx = \dfrac{n}{\theta^n}\int\limits_0^\theta x^n dx = \dfrac{n}{n+1}\theta$. Отсюда $\mathbf{M}\Big\{\Delta\widehat{\theta}_n\Big\} = \mathbf{M}\Big\{X_{(n)} - \theta\Big\} = \mathbf{M}\Big\{X_{(n)}\Big\} - \theta = \dfrac{n}{n+1}\theta - \theta = -\dfrac{\theta}{n+1} \to 0$, $n \to \infty$. Итак, при любом $\theta > 0$ $\mathbf{M}\Big\{\Delta\widehat{\theta}_n\Big\} < 0$, поэтому смещение $l_n \neq 0$, но $\mathbf{M}\Big\{\Delta\widehat{\theta}_n\Big\} \to 0$, $n \to \infty$, т.е. $\widehat{\theta}_n = X_{(n)}$ асимптотически несмещенная. ■

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

Р е ш е н и е. 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 $\mathbf{M}\{\varphi(Z_n) - \theta\} = \frac{C}{n}\sum\limits_{k=1}^{n} \mathbf{M}\{|X_k - m|\} - \theta = \left(C\sqrt{\frac{2}{\pi}} - 1\right)\theta$. 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: $\nu_n = \frac{1}{n}\sum\limits_{k=1}^{n} |X_k - m| \xrightarrow{\text{п.н.}} \mathbf{M}\{\xi\} = \sqrt{\frac{2}{\pi}}\theta$, $n \to \infty$. Поэтому $\widehat{\theta}_n = C\nu_n \xrightarrow{\text{п.н.}} C\mathbf{M}\{\xi\} = \theta$, если $C = \sqrt{\frac{\pi}{2}}$. Итак, оценка $\widehat{\theta}_n = \frac{\sqrt{\pi}}{n\sqrt{2}}\sum\limits_{k=1}^{n} |X_k - m|$ — несмещенная и сильно состоятельная оценка среднего квадратического отклонения θ. ■

П р и м е р 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 $\widetilde{\theta}_n = \sum\limits_{k=1}^{n} \alpha_k X_k$, где $\{\alpha_k\}$ — некоторые числовые коэффициенты.

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

Найдем с.к.-погрешность произвольной оценки $\widetilde{\theta}_n$ из Φ_n. Так как $\mathbf{M}\Big\{\widetilde{\theta}_n - \theta\Big\} = 0$ по доказанному выше, то $\Delta_n = \mathbf{M}\Big\{|\widetilde{\theta}_n - \theta|^2\Big\} = \mathbf{D}\Big\{\widetilde{\theta}_n - \theta\Big\} = \mathbf{D}\Big\{\widetilde{\theta}_n\Big\} = \sigma^2 \sum\limits_{k=1}^{n} \alpha_k^2$. 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 $|\alpha|^2 = \sum\limits_{k=1}^{n} \alpha_k^2 \geqslant \dfrac{1}{|e|^2} = \dfrac{1}{n}$. Если теперь положить $\widehat{\alpha}_k = \dfrac{1}{n}$, $k = 1, \ldots, n$, то $\sum\limits_{k=1}^{n} \widehat{\alpha}_k^2 = \dfrac{1}{n} \leqslant \sum\limits_{k=1}^{n} \alpha_k^2$. Итак, $\widehat{\theta}_n = \sum\limits_{k=1}^{n} \widehat{\alpha}_k X_k = \dfrac{1}{n} \sum\limits_{k=1}^{n} X_k = \overline{X}_n$ — 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 $\{\widehat{\alpha}_k = \dfrac{1}{n},\ k = 1, \ldots, n\}$). ■

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 $\Delta_n = \mathbf{D}\Big\{\widehat{\theta}_n\Big\}$.

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 $S_n^2 = \dfrac{1}{n} \sum\limits_{k=1}^{n} (X_k - m)^2$ 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 $\mathbf{M}\Big\{(X_{(n)})^2\Big\}$ и учтите результат примера 2.3.

5. Выборка $\{X_k,\ k=1,\dots,n\}$ 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 $\{X_k,\ k=1,\dots,n\}$ 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 $\{X_k,\ k=1,\dots,n\}$ 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 $\{X_k,\ k=1,\dots,n\}$ 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 $\mathbf{M}\{\overline{X}_n^2\} = \mathbf{D}\{\overline{X}_n\} + \big(\mathbf{M}\{\overline{X}_n\}\big)^2$ 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):

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

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

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

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

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 $\widehat{\theta}_n = \dfrac{\sum\limits_{k=1}^{n} X_k}{Nn} = \dfrac{\overline{X}_n}{N}$. 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 $\theta = \exp\Big\{\theta_1 + \dfrac{\theta_2}{2}\Big\}$; воспользуйтесь принципом инвариантности.

Ответ: $\widehat{\theta}_n = \exp\Big\{\overline{Y}_n + \dfrac{\overline{S}_n^2}{2}\Big\}$, 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

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

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 $I_n(\theta) = n\,i(\theta) = \frac{n}{\theta(1-\theta)}$. Поэто-

му $\mathbf{D}\left\{\widehat{\theta}_n\right\} = \mathbf{D}\{P_n^*(A)\} = \dfrac{1}{I_n(\theta)}$, 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 $\theta \in \Theta$

$$\mathbf{M}\left\{\widehat{\theta}_n\right\} = \frac{n}{\theta^n}\int\limits_0^\theta x^n dx = \frac{n}{n+1}\theta.$$

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

$$\mathbf{M}\left\{(\widetilde{\theta}_n)^2\right\} = \left(\frac{n+1}{n}\right)^2 \mathbf{M}\left\{\left(\widehat{\theta}_n\right)^2\right\} = \left(\frac{n+1}{n}\right)^2 \int\limits_0^\theta x^2 \frac{nx^{n-1}}{\theta^n} dx =$$

$$= \frac{(n+1)^2}{n\theta^n}\int\limits_0^\theta x^{n+1} dx = \frac{(n+1)^2}{n(n+2)}\theta^2.$$

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 $F_G(x) = \mathbf{P}\,(G(Z_n;\theta) \leqslant x)$ 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 $\theta \in \Theta$

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

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

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

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 $\alpha = 1 - \dfrac{p}{2}$ распределения $\mathcal{N}(0;1)$. Так как функция Лапласа строго монотонна, $\Phi(u_\alpha) = 1 - \dfrac{p}{2}$. Кроме того, если $\beta = \dfrac{p}{2}$, 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 $d(\theta) = \dfrac{1}{i(\theta)}$, где $i(\theta)$ — информация Фишера одного наблюдения. Пусть распределение, определяющее выборку, регулярно (см. определение 4.1), тогда $i(\theta) > 0$, $d(\theta) = \dfrac{1}{i(\theta)}$ 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

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

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: $\mathbf{M}\left\{\chi_{n,\delta}^2\right\} = n + \delta$, $\mathbf{D}\left\{\chi_{n,\delta}^2\right\} = 2(n + 2\delta)$.

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

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

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

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

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

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

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 $G_n(Z_n; \theta) = \dfrac{\sqrt{n-1}\,(\overline{X}_n - \theta)}{\overline{S}_n}$, 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:

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

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. ■

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

Решение. 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. ■

Пример 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.

Решение. 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 $\sqrt{n}\,(\widehat{\theta}_n - \theta) \xrightarrow{d} \eta \sim \sim \mathcal{N}(0; d(\theta))$, 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.

$$\text{Указание. Покажите, что } \frac{\sum\limits_{k=1}^{n} (X_k - \mu)^2}{\theta} \text{ — центральная статистика с}$$

распределением $\mathcal{H}_n$. Воспользуйтесь примером 5.3.

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

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

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

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

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 $\left[X_{(1)} + \frac{\ln(1-q)}{n};\ X_{(1)}\right]$.

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

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

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

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 $z_n = [x_1, \ldots, x_n]^\top$ — 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*

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

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

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

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 $T(Z_n) = \overline{X}_n = \frac{1}{n}\sum\limits_{k=1}^{n} X_k$. 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

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

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

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

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

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

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

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

Найдем теперь δ_1 с учетом того, что $\overline{X}_n \sim \mathcal{N}\left(\theta_0; \dfrac{\sigma^2}{n}\right)$, 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 $\Phi\left(\dfrac{\sqrt{n}(\delta_1 - \theta_0)}{\sigma}\right) = 1 - p$, т.е. $\dfrac{\sqrt{n}(\delta_1 - \theta_0)}{\sigma} = u_\alpha$, где u_α — квантиль уровня $\alpha = 1 - p$ распределения $\mathcal{N}(0;1)$. Таким образом, $\delta_1 = \theta_0 + u_\alpha \dfrac{\sigma}{\sqrt{n}}$.

Итак, если реализация выборочного среднего $\overline{X}_n$ удовлетворяет неравенству $\overline{x}_n \geqslant \theta_0 + u_\alpha \dfrac{\sigma}{\sqrt{n}}$, 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 $T(Z_n) = \overline{X}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n} X_k$. Из примера 5.5 следует, что при $n \gg 1$, $N = 1$, $\theta = \theta_0$ (т.е. H_0 — верна) и $\alpha = 1 - \dfrac{p}{2}$

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

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 $\overline{x}_n = T(z_n) = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{68}{1000}$ — соответствующая реализация выборочного среднего $\overline{X}_n$ (т.е. статистики критерия), то гипотезу H_0 на уровне значимости $p = 0{,}05$ следует отвергнуть, если $|\overline{x}_n - \theta_0| > u_\alpha \sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}$. 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 $|\overline{X}_n - \theta_0| \leqslant t_\alpha \sqrt{\frac{\overline{S}_n^2}{n-1}}$, где t_α — квантиль уровня $\alpha = 1 - \frac{p}{2}$ распределения Стьюдента с $n-1$ степенью свободы, $\overline{S}_n^2$ — 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 $\overline{S}_n^2 \in \left[k_1 \frac{\sigma^2}{n}; k_2 \frac{\sigma^2}{n} \right]$, где k_1 и k_2 — квантили распределения $\mathcal{H}_{n-1}$ уровней $\frac{p}{2}$ и $1 - \frac{p}{2}$ 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 $|\overline{X}_n - \overline{Y}_n| \leqslant u_\alpha \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}$, u_α — квантиль распределения $\mathcal{N}(0;1)$ уровня $\alpha = 1 - \frac{p}{2}$.

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

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

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

Ответ: $n_0 = 9$.

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 $b > 0$.

Ответ: $n_0 = \left[\dfrac{\sigma^2(u_a + u_b)}{(\theta_0 - \gamma)^2}\right] + 1$, 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 $H_1 : \theta = \sigma > \theta_0$.

Ответ: 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 $\mathcal{H}_{n-1}$.

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