МАТЕМАТИКА
В ТЕХНИЧЕСКОМ УНИВЕРСИТЕТЕ

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

XVII

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Издательство МГТУ имени Н.Э.Баумана

Комплекс учебников из 20 выпусков

Под редакцией В. С. Зарубина и А. П. Крищенко

I. Введение в анализ
II. Дифференциальное исчисление функций одного переменного
III. Аналитическая геометрия
IV. Линейная алгебра
V. Дифференциальное исчисление функций многих переменных
VI. Интегральное исчисление функций одного переменного
VII. Кратные и криволинейные интегралы. Элементы теории поля
VIII. Дифференциальные уравнения
IX. Ряды
X. Теория функций комплексного переменного
XI. Интегральные преобразования и операционное исчисление
XII. Дифференциальные уравнения математической физики
XIII. Приближенные методы математической физики
XIV. Методы оптимизации
XV. Вариационное исчисление и оптимальное управление
XVI. Теория вероятностей
XVII. Математическая статистика
XVIII. Случайные процессы
XIX. Дискретная математика
XX. Исследование операций

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Под редакцией
д-ра техн. наук, профессора В. С. Зарубина
и д-ра физ.-мат. наук, профессора А. П. Крищенко

Допущено
Министерством образования
Российской Федерации
в качестве учебника для студентов
высших технических учебных заведений

Москва
Издательство МГТУ им. Н. Э. Баумана
2001

УДК 519.22(075.8)
ББК 22.172
М34

Рецензенты: проф. Ю.Н. Тюрин, проф. Э.К. Лецкий

М34 **Математическая статистика**: Учеб. для вузов / В.Б. Горяинов, И.В. Павлов, Г.М. Цветкова и др.; Под ред. В.С. Зарубина, А.П. Крищенко. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2001. –424 с. (Сер. Математика в техническом университете; Вып. XVII).

ISBN 5-7038-1730-7 (Вып. XVII)
ISBN 5-7038-1270-4

Предлагаемая книга, выпущенная в серии „Математика в техническом университете", знакомит читателя с основными понятиями математической статистики и некоторыми из ее приложений. Ее отличительной особенностью является взвешенное сочетание математической строгости с прикладной направленностью задач. Каждую главу книги завершает большой набор типовых примеров, контрольных вопросов и задач для самостоятельного решения.

Содержание учебника соответствует курсу лекций, который авторы читают в МГТУ им. Н.Э. Баумана.

Для студентов технических университетов. Может быть полезен преподавателям, аспирантам и инженерам.

Ил. 28. Табл. 63. Библиогр. 35 назв.

Выпуск книги финансировал
Московский государственный технический
университет им. Н.Э. Баумана

УДК 519.22(075.8)
ББК 22.172

ISBN 5-7038-1730-7 (Вып. XVII)
ISBN 5-7038-1270-4

ПРЕДИСЛОВИЕ

Предлагаемая книга является выпуском XVII комплекса учебников „Математика в техническом университете“. Хочется надеяться, что она будет полезна при овладении прикладными методами теории математической статистики.

Математическая статистика — раздел математики, который занимается разработкой методов получения научно обоснованных выводов о массовых явлениях и процессах по данным наблюдений или экспериментов. Например, по имеющейся информации о числе бракованных изделий в партии готовой продукции надо сделать вывод о качестве используемого технологического процесса.

Математическая статистика предполагает вероятностную природу данных наблюдений, поэтому она основана на понятиях и методах теории вероятностей. Задачи математической статистики в известной мере являются обратными к задачам теории вероятностей. Если в теории вероятностей мы считаем заданной вероятностную модель случайного явления и делаем расчет вероятностей интересующих нас событий, то в математической статистике исходим из того, что вероятностная модель не задана (или задана не полностью), а в результате эксперимента стали известны реализации каких-либо случайных событий. На основе статистических данных математическая статистика подбирает подходящую вероятностную модель для получения вывода о рассматриваемом явлении или процессе.

Проиллюстрируем сказанное на примере. Постановка задачи теории вероятностей. Вероятность выпадения „герба“ при подбрасывании монеты известна и равна p. Какова вероятность того, что при n подбрасываниях монеты герб выпадет k раз, где $0 \leqslant k \leqslant n$?

Постановка задачи математической статистики. Монету подбрасывали n раз, и „герб“ выпал k раз. Что можно сказать о вероятности выпадения герба при одном подбрасывании?

Возможны и другие постановки задачи. Например, проверить, можно ли на основании полученных данных считать, что вероятность выпадения „герба“ при одном подбрасывании равна p_0.

В настоящее время математическая статистика — обширный раздел математики. В предлагаемой читателю книге рассмотрены основные понятия математической статистики и решаемые ею задачи: оценивание неизвестных параметров распределения вероятностей, проверка статистических гипотез, установление формы и степени связи между несколькими случайными переменными.

Отличительной особенностью прелагаемого учебника является также изложение непараметрических методов математической статистики, приобретающих все большую популярность в технических приложениях. Непараметрические методы стали доступными в инженерной практике благодаря появлению персональных компьютеров и пакетов прикладных программ по математической статистике.

Построение книги имеет блочную структуру. В каждой главе после изложения теоретического материала даны типовые примеры с решениями, а также контрольные вопросы и задачи для самостоятельного решения. Наличие большого количества примеров и задач позволяет использовать книгу не только как учебник, но и как задачник при проведении семинарских занятий.

Место этого учебника среди многих других книг и руководств по математической статистике определяется желанием дать доступное для инженеров изложение основ теории, не перегруженное строгими выводами. Более требовательный читатель может обратиться к специальной литературе, используя имеющиеся в тексте ссылки.

Содержательная сторона статистических моделей разъясняется на конкретных примерах преимущественно из инженерной практики. Таким образом, по замыслу авторов, учебник должен заполнить пробел между руководствами по математической статистике, имеющими „рецептурный" стиль изложения, и университетскими курсами, требующими математической подготовки в объеме, не предусмотренном программами технических вузов.

Следует отметить, что, поскольку книга рассчитана на читателей, впервые знакомящихся с математической статистикой, авторы старались не допускать двойственного толкования обозначений, как это принято в учебных пособиях и монографиях, предназначенных для подготовленного читателя.

Для уточнения того, что нужно знать из других разделов математики для чтения учебника, в начале книги сформулированы вопросы для самопроверки. При этом понятия и термины, которые нужно знать и которые были введены в других выпусках серии „Математика в техническом университете", в вопросах выделены **прямым полужирным шрифтом**. Далее помещен список основных обозначений, содержащий часто встречающиеся в тексте символы и их расшифровку.

В конце книги приведены таблицы некоторых распределений, список рекомендуемой литературы и предметный указатель, в который входят в алфавитном порядке (по существительному в именительном падеже) все выделенные в тексте ***полужирным курсивом*** термины с указанием страниц, на которых они строго определены или описаны. Выделение термина *светлым курсивом* означает, что в данном параграфе он является одним из ключевых слов и читателю должно быть известно значение термина. Читатель может уточнить это значение, найдя при помощи предметного указателя необходимую страницу.

Ссылки в тексте на номера формул, рисунков и таблиц набраны обычным шрифтом (например, (1.5) — пятая формула в главе 1; рис. 3.2 — второй рисунок в главе 3; табл. 1.4 —

четвертая таблица в главе 1), а на параграфы и таблицы в приложениях — полужирным (например, **1.3** — третий параграф в главе 1, табл. **П.2** — вторая таблица приложения). В квадратных скобках даны ссылки на другие выпуски серии, например [X] — ссылка на десятый выпуск.

Авторы приносят глубокую благодарность И.К. Волкову, который оказал серьезную помощь при написании книги.

Задания для самопроверки

1. Что такое **множество, подмножество**? Какие множества называют **конечными, счетными**? Какие операции над множествами (подмножествами) Вы знаете? Какими свойствами обладают эти операции? Что такое **отрезок, интервал, полуинтервал**? [I]

2. Дайте определение **отображения**. Дайте определение **действительной функции действительного переменного**. Какую **функцию** называют **монотонной, возрастающей, убывающей, неубывающей, четной, нечетной**? Какую **функцию** называют **обратной** к данной? Какие функции называют **полиномами**? [I]

3. Что называют **пределом функции** $f(x)$ при $x \to x_0$, при $x \to +\infty$, при $x \to -\infty$? Какую **функцию** называют **непрерывной в точке, непрерывной слева** в точке, **в интервале, на отрезке**? [I]

4. Дайте определение **производной** действительной **функции** действительного переменного. Дайте определение **производной** n**-го порядка**. [II]

5. Что такое **матрица**? Какую **матрицу** называют **нулевой, диагональной, единичной, симметрической**? Дайте определение **произведения** двух **матриц**. Какую **квадратную матрицу** называют **невырожденной, вырожденной, неотрицательно определенной**? Какую **матрицу** называют **обратной** по отношению к данной? Что называют **рангом**

матрицы? В чем состоит операция **транспонирования матриц**? Что называют **следом матрицы**, **алгебраическим дополнением**? [III]

6. Запишите неравенство **Коши — Буняковского**. В каком случае оно обращается в равенство? [IV]

7. Что такое **случайный эксперимент (опыт)** и из чего состоит **множество** его (**элементарных**) **исходов**? Что называют **пространством элементарных событий (исходов)**? Какое **событие** называют **случайным**? Перечислите операции над событиями и сформулируйте их свойства. Дайте **классическое**, **статистическое** и **аксиоматическое определения вероятности**. Что называют **вероятностным пространством**? [XVI]

8. Дайте определение **условной вероятности**. Какие **события** называют **независимыми**? [XVI]

9. Какую схему повторных испытаний называют **схемой Бернулли (биномиальной)**? Запишите **формулу Бернулли**. [XVI]

10. Дайте определения **(скалярной) случайной величины**, **n-мерного случайного вектора закона распределения (вероятностей) случайной величины** и **случайного вектора (векторной случайной величины)**. [XVI]

11. Что называют **рядом распределения дискретной случайной величины**? Дайте определения **функции распределения (вероятностей) одномерной** и **n-мерной случайных величин** (случайного вектора). Сформулируйте свойства функции распределения одномерной и n-мерной случайных величин. [XVI]

12. Дайте определения **плотности распределения (вероятностей) одномерной** и **n-мерной (векторной) случайных величин**. [XVI]

13. Дайте определение **квантили уровня α** (функции) распределения (случайной величины). [XVI]

14. Приведите определение **математического ожидания (среднего значения)** для дискретной и непрерывной случай-

ных величин. Перечислите свойства математического ожидания. [XVI]

15. Дайте определение **дисперсии и среднего квадратичного отклонения** случайной величины. Сформулируйте их свойства. [XVI]

16. Приведите определение **начального и центрального моментов k-го порядка** распределения случайной величины. [XVI]

17. Что называется **медианой** случайной величины? [XVI]

18. Дайте определение **условного закона распределения (условной функции распределения, условной плотности распределения)** случайной величины. [XVI]

19. Дайте определение **условного математического ожидания, условной дисперсии** случайной величины. Перечислите их свойства. Что такое **регрессия, функция регрессии**? Какую линию называют **линией регрессии**? [XVI]

20. Дайте определение **ковариации и коэффициента корреляции** двух случайных величин. Сформулируйте их свойства. Какие **случайные величины** называют **независимыми, некоррелированными, одинаково распределенными**? [XVI]

21. Какое распределение (закон распределения) называют **биномиальным, нормальным, гауссовым, стандартным нормальным, равномерным, экспоненциальным (показательным), гамма-распределением, распределением Коши, распределением Пуассона, распределением Стьюдента**? [XVI]

22. Какую **случайную величину** называют **нормально распределенной, равномерно распределенной, экспоненциально распределенной, биномиально распределенной**? Чему равны их математические ожидания и дисперсии? [XVI]

23. Какую **матрицу** называют **ковариационной, корреляционной**? [XVI]

24. Дайте определение **корреляционного отношения**. Какая связь существует между корреляционным отношением и **коэффициентом корреляции**? [XVI]

25. Что называют **функцией от случайной величины**? Как найти функцию распределения функции от случайной величины, зная закон распределения аргумента? Запишите выражение для плотности распределения монотонной функции от непрерывной случайной величины. Что называют **композицией (сверткой)** плотностей распределения случайных величин? [XVI]

26. Что можно сказать о распределении линейной комбинации случайных величин, распределенных по нормальному закону? [XVI]

27. Дайте определения **сходимости по вероятности** последовательности случайных величин и **слабой сходимости** последовательности функций распределения. [XVI]

28. Запишите **второе неравенство Чебышева**. Сформулируйте **закон больших чисел** в форме Бернулли и в форме Чебышева. [XVI]

29. Сформулируйте **центральную предельную теорему** (теорему Ляпунова) и **интегральную теорему Муавра — Лапласа**. [XVI]

ОСНОВНЫЕ ОБОЗНАЧЕНИЯ

◀ и ▶ — начало и окончание доказательства

$\#$ — окончание примера, замечания, теоремы без доказательства

$\mathbb{R}$ — множество вещественных чисел I-1.3

$\mathbb{R}^n$ — линейное арифметическое пространство IV

$p(x)$, $p(x;\theta)$ — плотность распределения вероятностей для непрерывной случайной величины X XVI, 1.1

θ — параметр функции (плотности) распределения 1.1

$\vec{\theta}=(\theta_1, \ldots, \theta_r)$ — вектор параметров функции (плотности) распределения 1.1

$\mathbf{P}\{A\}$ — вероятность события A XVI

$Y_n \xrightarrow[n\to\infty]{\mathbf{P}} Y$ — сходимость по вероятности последовательности $\{Y_n\}$ случайных величин к Y XYI

$F_n(x) \underset{n\to\infty}{\Longrightarrow} F(x)$ — сходимость по распределению (слабая сходимость) последовательности $\{F_n(x)\}$ функций распределения к функции $F(x)$ XVI

$\vec{X}_n=(X_1, \ldots, X_n)$ — случайная выборка объема n из генеральной совокупности X 1.1

$\vec{x}_n=(x_1, \ldots, x_n)$ — выборка объема n из генеральной совокупности X (реализация случайной выборки $\vec{X}_n$) 1.1

$\mathcal{X}_n$ — выборочное пространство (множество значений случайной выборки) 1.1

$\mathcal{P}$ — класс (множество) распределений случайной выборки 1.1

$F_X(x)$ — функция распределения генеральной совокупности X (случайной величины X) XVI, 1.1

$\{F(x)\}$ — статистическая модель **1.1**

$\{F(x,\vec{\theta});\vec{\theta}\in\Theta\}$ — параметрическая модель **1.1**

$X_{(i)}$ — i-й порядковый член вариационного ряда случайной выборки $\vec{X}_n$ **1.3**

$x_{(i)}$ — i-й член вариационного ряда выборки $\vec{X}_n$ **1.3**

$F_{\vec{X}}(t_1,\ldots,t_n)$ — функция распределения случайной выборки $\vec{X}_n$ объема n **1.1**

$\widehat{F}(x,\vec{X}_n)$ — выборочная функция распределения **1.3**

$F_n(x)$ — эмпирическая функция распределения **1.3**

$p_n(x)$ — эмпирическая плотность распределения **1.3**

$g(\vec{X}_n)$ — выборочная характеристика (статистика) **1.2**

$g_{\text{в}}=g(\vec{x}_n)$ — выборочное значение (значение выборочной характеристики $g(\vec{X}_n)$) **1.2**

$\mathrm{M}X$, μ — математическое ожидание случайной величины X **XVI**

$\mathrm{D}X$, σ^2 — дисперсия случайной величины X **XVI**

σ — среднее квадратичное отклонение случайной величины **XVI**

$\widehat{\mu}_k(\vec{X}_n)$ — начальный выборочный момент k-го порядка (оценка начального момента k-го порядка) **1.3**

$\widehat{\mu}_k$ — начальный момент k-го порядка выборки **1.3**

$\widehat{\nu}_k(\vec{X}_n)$ — центральный выборочный момент k-го порядка (оценка центрального момента k-го порядка) **1.3**

$\widehat{\nu}_k$ — центральный момент k-го порядка выборки **1.3**

$\overline{X}$ — выборочное среднее (оценка математического ожидания) случайной выборки $\vec{X}_n$ **1.3**

$\overline{x}$ — среднее (значение) выборки $\vec{x}_n$ **1.3**

$\widehat{\sigma}^2(\vec{X}_n)$ — выборочная дисперсия (оценка дисперсии) случайной выборки $\vec{X}_n$ **1.3**

$\widehat{\sigma}^2$ — дисперсия выборки **1.3**

$\widehat{\sigma}(\vec{X}_n)$ — выборочное среднее квадратичное отклонение (оценка среднего квадратичного отклонения) случайной выборки $\vec{X}_n$ **1.3**

$\widehat{\sigma}$ — среднеее квадратичное отклонение выборки **1.3**

$\widehat{S}^2(\vec{X}_n)$ — исправленная несмещенная оценка дисперсии **8.1**

S^2 — значение $\widehat{S}^2(\vec{X}_n)$ **8.1**

$\widetilde{S}^2(\vec{X}_n)$ — оценка дисперсии при известном математическом ожидании **8.1**

$\widehat{\theta}(\vec{X}_n)$ — точечная оценка параметра θ **1.2**

$\widehat{\theta}$ — значение оценки параметра $\vec{\theta}$ **1.2**

ρ — коэффициент корреляции **XVI-5.5, 1.3**

$\widehat{K}(\vec{X}_n, \vec{Y}_n)$ — выборочный корреляционный момент (оценка корреляционного момента) **1.3**

$\widehat{K}$ — корреляционный момент выборки **1.3**

$\widehat{\rho}(\vec{X}_n, \vec{Y}_n)$ — выборочный коэффициент корреляции (оценка коэффициента корреляции) **1.3**

$\widehat{\rho}$ — коэффициент корреляции выборки (значение оценки коэффициента корреляции) **1.3**

$L(\theta; \vec{X}_n)$ — функция правдоподобия **2.2**

$\varphi_n(X_1, \ldots, X_n)$ — отношение правдоподобия **4.3**

$I(\theta)$ — количество информации по Фишеру **2.1**

$e(\theta)$ — показатель эффективности **2.1**

$\underline{\theta}(\vec{X}_n)$ — нижняя граница интервальной оценки для параметра θ **1.2, 3.1**

$\overline{\theta}(\vec{X}_n)$ — верхняя граница интервальной оценки для параметра θ **1.2, 3.1**

$\big(\underline{\theta}(\vec{x}_n), \overline{\theta}(\vec{x}_n)\big)$, $\big(\underline{\theta}, \overline{\theta}\big)$ — доверительный интервал для параметра θ **3.1**

$D_{\vec{X}_n}$ — система γ-доверительных множеств **3.4**

$s(\theta)$ — оперативная характеристика **4.5**

W — критическое множество **4.2**

H — статистическая гипотеза **4.1**

$M(\theta)$ — функция мощности критерия **4.5**

γ — коэффициент доверия (доверительная вероятность) **3.1**

r, $r_{\xi\eta}$ — корреляционное отношение **6.2**

$\widehat{r}(\vec{X}_n, \vec{Y}_n)$ — оценка корреляционного отношения для пары случайных выборок $\vec{X}_n$ и $\vec{Y}_n$ **6.2**

$\widehat{r}$ — значение оценки корреляционного отношения **6.5**

$\rho_{ij(J(i,j))}$ — частный коэффициент корреляции **6.5**

$\widehat{\rho}_{ij(J(i,j))}(\vec{X}_{0n}, \ldots, \vec{X}_{Nn})$ — оценка частного коэффициента корреляции **6.5**

$\widehat{\rho}_{ij(J(i,j))}$ — значение оценки частного коэффициента корреляции **6.5**

R_η, R — множественный коэффициент корреляции (коэффициент детерминации) **6.5**

$R_\eta(\vec{X}_n, \vec{Y}_n)$ — оценка множественного коэффициента корреляции **6.5**

$\widehat{R}_\eta$ — значение оценки множественного коэффициента корреляции **6.5**

u_q — квантиль уровня q стандартного нормального распределения **6.5**

t_q — квантиль уровня q распределения Стьюдента **XVI**

χ^2_q — квантиль уровня q распределения χ^2 **Д.3.1**

f_q — квантиль уровня q распределения Фишера **Д.3.1**

$X \sim N(\mu, \sigma^2)$ — случайная величина X имеет нормальное распределение с параметрами μ и σ^2 **Д.3.1**

$X \sim S(m)$ — случайная величина X имеет распределение Стьюдента с m степенями свободы **Д.3.1**

$X \sim \chi^2(m)$ — случайная величина X имеет распределение χ^2 с m степенями свободы **Д.3.1**

$X \sim F(k,m)$ — случайная величина X имеет распределение Фишера с k и m степенями свободы **Д.3.1**

$X \sim \Gamma(\lambda,\alpha)$ — случайная величина X имеет γ-распределение с параметрами λ и α **Д.3.1**

$X \sim \Pi(\lambda)$ — случайная величина X имеет распределение Пуассона с параметром λ **Д.3.1**

H_θ — γ-зона для параметра θ **3.4**

$\mathcal{F}$ — класс допустимых моделей регрессии **Д.7.1**

Буквы латинского алфавита

Начертание	Произношение	Начертание	Произношение
A a *A a*	а	N n *N n*	эн
B b *B b*	бэ	O o *O o*	о
C c *C c*	цэ	P p *P p*	пэ
D d *D d*	дэ	Q q *Q q*	ку
E e *E e*	е	R r *R r*	эр
F f *F f*	эф	S s *S s*	эс
G g *G g*	же	T t *T t*	тэ
H h *H h*	аш	U u *U u*	у
I i *I i*	и	V v *V v*	вэ
J j *J j*	йот	W w *W w*	дубль-вэ
K k *K k*	ка	X x *X x*	икс
L l *L l*	эль	Y y *Y y*	игрек
M m *M m*	эм	Z z *Z z*	зэт

Представлен наиболее употребительный (но не единственный) вариант произношения (в частности, вместо „йот“ иногда говорят „жи“).

Буквы греческого алфавита

Начертание	Произношение	Начертание	Произношение	Начертание	Произношение
A α	альфа	I ι	йота	P ρ	ро
B β	бета	K κ	каппа	Σ σ	сигма
Γ γ	гамма	Λ λ	ламбда	T τ	тау
Δ δ	дельта	M μ	ми	Υ υ	ипсилон
E ε	эпсилон	N ν	ни	Φ φ	фи
Z ζ	дзета	Ξ ξ	кси	X χ	хи
H η	эта	O o	омикрон	Ψ ψ	пси
Θ θ ϑ	тэта	Π π	пи	Ω ω	омега

Наряду с указанным произношением также говорят „лямбда“, „мю“ и „ню“.

1. ОСНОВНЫЕ ПОНЯТИЯ ВЫБОРОЧНОЙ ТЕОРИИ

1.1. Генеральная совокупность. Выборка. Выборочные характеристики

Прежде чем ввести основные понятия математической статистики, рассмотрим пример. Некоторое стабильно работающее (т.е. работающее в одних и тех же условиях) предприятие изготавливает приборы, которые характеризуются некоторым количественным признаком. В силу влияния не поддающихся учету факторов значение количественного признака от прибора к прибору меняется. Например в случае, когда интерес представляет доля брака в производстве приборов, каждому изделию можно приписать значение 1, если прибор функционирует нормально, и значение 0, если прибор неисправен. Количественным признаком может быть также время бесперебойной работы прибора, точность измерительного прибора, чувствительность датчика и т.п.

В силу объективных причин обеспечивать контроль каждого прибора, как правило, не удается. Поэтому для контроля качества продукции поступают следующим образом. Выбирают наудачу некоторое количество n (конечное число) приборов и по их показателям судят о всей продукции в целом, например о доле бракованных изделий или о средней продолжительности бесперебойной работы прибора и т.д. В подобных ситуациях естественно предполагать, что наблюдения за контролируемым показателем (хотя бы мысленно) можно проводить сколько угодно раз. Результаты n наблюдений рассматриваются как значения случайной величины — рассматриваемого количественного признака. Эта случайная величина может быть как

дискретной, так и непрерывной. Например, она может принимать только два значения 0 и 1, если речь идет о проверке, является прибор бракованным или нет. В другой же ситуации, когда оценивается время бесперебойной работы прибора, естественно считать, что случайная величина может принимать любое неотрицательное значение и является непрерывной.

В математической статистике множество возможных значений случайной величины X называют ***генеральной совокупностью*** случайной величины X или просто генеральной совокупностью X. Под ***законом распределения*** (***распределением***) ***генеральной совокупности*** X будем понимать закон распределения вероятностей случайной величины X.

Исходным материалом для изучения свойств генеральной совокупности (т.е. некоторой случайной величины) являются ***экспериментальные*** (***статистические***) ***данные***, под которыми понимают значения случайной величины, полученные в результате повторений случайного эксперимента (наблюдений над случайной величиной).

Предполагаем, что эксперимент хотя бы теоретически может быть повторен сколько угодно раз в одних и тех же условиях. Под словами „в одних и тех же условиях“ будем понимать, что распределение случайной величины X_i, $i = 1, 2, \ldots$, заданной на множестве исходов i-го эксперимента, не зависит от номера испытания и совпадает с распределением генеральной совокупности X. В этом случае принято говорить о ***независимых повторных экспериментах*** (***испытаниях***) или о ***независимых повторных наблюдениях*** над случайной величиной.

Совокупность независимых случайных величин $X_1, \ldots, X_n$, каждая из которых имеет то же распределение, что и случайная величина X, будем называть ***случайной выборкой*** из генеральной совокупности X и записывать $\vec{X}_n = (X_1, \ldots, X_n)$ (иногда просто $X_1, \ldots, X_n$). При этом число n называют ***объемом случайной выборки***, а случайные величины X_i — ***элементами случайной выборки***.

Любое возможное значение $\vec{x_n} = (x_1, \ldots, x_n)$ случайной выборки $\vec{X_n}$ будем называть ***выборкой*** из генеральной совокупности X (также ***реализацией случайной выборки*** $\vec{X_n}$) Число n характеризует ***объем выборки***, а числа x_i, $i = \overline{1,n}$, представляют собой ***элементы выборки*** $\vec{x_n}$. Выборку $\vec{x_n}$ можно интерпретировать как совокупность n чисел $x_1, \ldots, x_n$, полученных в результате проведения n повторных независимых наблюдений над случайной величиной X.

Основой любых выводов о вероятностных свойствах генеральной совокупности X, т.е. ***статистических выводов***, является ***выборочный метод***, суть которого заключается в том, что свойства случайной величины X устанавливаются путем изучения тех же свойств на случайной выборке.

Множество возможных значений случайной выборки $\vec{X_n}$ содержит информацию о случайной величине, полученную в эксперименте. Это множество называют ***выборочным пространством*** и обозначают $\mathcal{X}_n$. Выборочным пространством может быть или *n-мерное линейное арифметическое пространство* $\mathbb{R}^n$, или его подмножество. Если X — дискретная случайная величина, то выборочное пространство — конечное или счетное.

Элементы X_i, $i = \overline{1,n}$, случайной выборки $\vec{X_n}$ независимы и имеют то же распределение, что и генеральная совокупность X. Таким образом, функция распределения $F_{\vec{X}}(t_1,\ldots,t_n)$ случайной выборки $\vec{X_n}$ имеет вид

$$F_{\vec{X}}(t_1,\ldots,t_n) = \mathbf{P}\{X_1 < t_1, \ldots, X_n < t_n\} = \prod_{i=1}^{n} \mathbf{P}\{X_i < t_i\} = \prod_{i=1}^{n} F(t_i), \quad (1.1)$$

где $F(t)$ — функция распределения случайной величины X (генеральной совокупности X).

О распределении случайной величины X в одних случаях у исследователя могут быть самые общие представления. Напри-

мер, X является непрерывной случайной величиной и только (о распределении практически ничего не известно!). В других случаях функция распределения (в случае непрерывной случайной величины — плотность распределения вероятностей) известна, но не известны параметры, от которых она зависит. Например, известно, что генеральная совокупность X имеет нормальный закон распределения

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где μ и σ — неизвестные параметры.

Значит, можно говорить только о семействе (классе) $\mathcal{P}$ распределений случайной выборки, в котором содержится ***априорная информация*** (информация до опыта) исследователя.

Выборочное пространство, на котором задан класс распределений $\mathcal{P}$, назовем ***статистической моделью****.

В случае повторных независимых испытаний статистическую модель будем обозначать $\{F(x)\}$, поскольку она полностью определена функцией распределения $F(x)$ генеральной совокупности X.

Если функция распределения (плотность распределения) задана с точностью до неизвестного вектора параметров $\vec{\theta} = (\theta_1, \ldots, \theta_r)$ с множеством возможных значений Θ, т.е. $\vec{\theta} \in \Theta$, то статистическую модель называют ***параметрической моделью***. Параметрическую модель обозначают $\{F(x;\vec{\theta}); \vec{\theta} \in \Theta\}$. Множество Θ называют ***параметрическим множеством***.

Следует отметить, что о параметрическом множестве исследователь может не иметь никакой априорной информации.

Статистическую модель называют ***непрерывной*** или ***дискретной***, если случайная величина X является, соответственно, непрерывной или дискретной. В дальнейшем мы будем

*Разумеется, это слишком узкое толкование термина, которое уместно лишь в рамках данной книги.

предполагать, что генеральная совокупность X с функцией распределения $F(x)$ является либо дискретной, либо непрерывной случайной величиной. В первом случае распределение X задают в виде таблицы (ряда распределений), а во втором — в виде плотности распределения $p_X(x)$. При этом будем использовать единое обозначение $p(x)$ (или $p(x;\theta)$ для параметрических моделей) как для плотности распределения случайной величины X, когда она непрерывная, так и для вероятности $\mathbf{P}\{X=x\}$ в случае дискретной случайной величины X.

Пример 1.1. Пусть известно, что генеральная совокупность случайной величины X распределена по нормальному закону с известной дисперсией и неизвестным средним θ. Тогда статистическая модель имеет вид $\{F(x;\theta);\ \theta\in\Theta=\mathbb{R}\}$ и может быть задана с помощью плотности распределения вероятностей

$$p(x;\theta)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\theta)^2}{2\sigma^2}},\quad x\in\mathbb{R}.$$

Если неизвестны оба параметра: среднее значение θ_1 и среднее квадратичное отклонение θ_2, то статистическая модель имеет вид $\left\{F(x;\vec{\theta});\ \vec{\theta}=(\theta_1,\theta_2)\in\Theta\right\}$, где $\Theta\subset\mathbb{R}^2$ $(\theta_1\in\mathbb{R},\ \theta_2\in\mathbb{R}^+)$ и плотность распределения вероятностей содержит два неизвестных параметра:

$$p(x;\theta_1,\theta_2)=\frac{1}{\sqrt{2\pi}\theta_2}e^{-\frac{(x-\theta_1)^2}{2\theta_2^2}},\quad x\in\mathbb{R}.$$

Пример 1.2. Пусть случайная величина X имеет распределение Пуассона с неизвестным параметром. Тогда статистическая модель имеет вид $\{F(x;\theta);\ \theta\in\Theta=(0,\infty)\}$, где $F(x;\theta)$ определяется равенством

$$p(x;\theta)=PrbX=x=\frac{\theta^x}{x!}e^{-\theta},\quad x=0,1,2,\ldots$$

Замечание 1.1. Наряду с генеральной совокупностью одномерной случайной величины можно рассматривать генеральную совокупность многомерной (векторной) случайной величины, распространяя введенные выше понятия на этот случай. При этом случайную выборку объема n из генеральной совокупности $(X, Y, \ldots, Z)$ будем обозначать $\left(\vec{X}_n, \vec{Y}_n, \ldots, \vec{Z}_n\right)$. #

В дальнейшем мы будем рассматривать различные функции $Y = g(X_1, \ldots, X_n)$ (или $Y = g(\vec{X}_n)$) случайной выборки $\vec{X}_n = (X_1, \ldots, X_n)$, например:

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad Y = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Любую функцию $g(\vec{X}_n)$ случайной выборки в математической статистике называют ***статистикой***, или ***выборочной характеристикой***. Распределение этой случайной величины называют ***выборочным распределением***. Выборочное распределение однозначно определяется совместным распределением случайных величин $X_1, \ldots, X_n$, т.е. распределением случайной выборки $\vec{X}_n$. Значение $g(\vec{x}_n)$ выборочной характеристики $g(\vec{X}_n)$, определенное по реализации $\vec{x}_n$ случайной выборки $\vec{X}_n$, называют ее ***выборочным значением***.

В математической статистике часто приходится рассматривать поведение выборочных характеристик при $n \to \infty$, где n — объем случайной выборки $(X_1, \ldots, X_n)$. При этом будем писать $Y_n = g(X_1, \ldots, X_n)$ и рассматривать последовательность случайных величин $\{Y_n\}$, сходящуюся в том или ином смысле к некоторому Y — случайной величине или константе.

В [XVI] были рассмотрены основные типы сходимости последовательности случайных величин и связь между ними. В этой книге будем использовать два вида сходимости: сходимость по вероятности и сходимость по распределению, или слабую сходимость.

Напомним, что последовательность $\{Y_n\}$ случайных величин называют сходящейся по вероятности к Y, т.е. $Y_n \xrightarrow{\mathbf{P}} Y$, если для любого $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\{|Y_n - Y| > \varepsilon\} \to 0.$$

Если имеет место равенство

$$\lim_{n\to\infty} F_{Y_n}(x) = F_Y(x)$$

в каждой точке непрерывности $F_Y(x)$, то говорят о *слабой сходимости* последовательности $F_{Y_n}(x)$ *функций распределения* (сходимости по распределению) и пишут

$$F_{Y_n}(x) \underset{n\to\infty}{\Longrightarrow} F_Y(x).$$

Уместно также говорить о слабой сходимости последовательности $\{Y_n\}$ случайных величин к Y. В этом смысле можно утверждать, что из сходимости по вероятности следует слабая сходимость.

Отметим, что в теории вероятностей и ее приложениях наиболее часто используемые законы распределения случайных величин имеют общепринятые названия и обозначения.

Например, нормальный закон со средним μ и дисперсией σ^2 обозначают символом $N(\mu, \sigma^2)$; распределение Пуассона со средним λ — символом $\Pi(\lambda)$ и т.д.

Если наблюдаемая в эксперименте случайная величина X имеет распределение некоторого стандартного типа, то соответствующая статистическая модель имеет такое же название: ***нормальная модель***, ***модель Коши***, ***биномиальная модель***, ***пуассоновская модель*** и т.д.

Для обозначения того, что случайная величина X имеет закон распределения $F(x)$, употребляют символическую запись $X \sim F(x)$. Например, запись $X \sim N(\mu, \sigma^2)$ означает, что случайная величина X имеет нормальный закон распределения с параметрами μ и σ^2.

1.2. Основные задачи математической статистики

При решении любой задачи математической статистики исследователь располагает двумя источниками информации. Первый и наиболее определенный (явный) — это результаты наблюдений (эксперимента) в виде *выборки* из некоторой *генеральной совокупности* скалярной или векторной случайной величины. При этом *объем выборки* n может быть фиксирован, а может увеличиваться в ходе эксперимента (т.е. могут использоваться так называемые последовательные процедуры статистического анализа).

Второй источник — это вся *априорная информация* об интересующих исследователя свойствах изучаемого объекта, которая накоплена к текущему моменту. Формально объем априорной информации отражается в той исходной *статистической модели*, которую исследователь выбирает при решении своей задачи.

В математической статистике всегда в той или иной мере используют априорную информацию об исследуемом объекте, но степень обоснованности такого использования лежит на совести (или зависит от компетентности) конкретного исследователя.

Если есть сомнения в том или ином исходном допущении при решении конкретной задачи, то его нужно проверять и обосновывать, а при невозможности это сделать — отбросить и попытаться найти решение задачи без привлечения сомнительных допущений.

Перечислим некоторые задачи математической статистики, наиболее часто встречающиеся в ее приложениях.

Оценка неизвестных параметров. Задача оценивания неизвестных параметров возникает в тех случаях, когда функция распределения генеральной совокупности известна с точностью до параметра θ. В этом случае необходимо найти такую

статистику $\widehat{\theta}(\vec{X}_n)$, *выборочное значение* $\widehat{\theta} = \widehat{\theta}(\vec{x}_n)$ которой для рассматриваемой *реализации* $\vec{x}_n$ *случайной выборки* можно было бы считать приближенным значением параметра $\vec{\theta}$.

Статистику $\widehat{\theta}(\vec{X}_n)$, выборочное значение $\widehat{\theta}$ которой для любой реализации $\vec{x}_n$ принимают за приближенное значение неизвестного параметра θ, называют его ***точечной оценкой*** или просто ***оценкой***, а $\widehat{\theta}$ — ***значением точечной оценки*** (просто ***оценки***).

Понятно, что точечная оценка $\widehat{\theta}(\vec{X}_n)$ должна удовлетворять вполне определенным требованиям для того, чтобы ее выборочное значение $\widehat{\theta}$ соответствовало истинному значению параметра $\vec{\theta}$. Свойства точечных оценок рассмотрены ниже (см. 2).

Для точечных оценок параметра θ будем использовать и другие обозначения, например $\widetilde{\theta}(\vec{X}_n)$, $\theta^*(\vec{X}_n)$.

Возможным является и иной подход к решению рассматриваемой задачи: найти такие статистики $\overline{\theta}(\vec{X}_n)$ и $\underline{\theta}(\vec{X}_n)$, чтобы с вероятностью γ выполнялось неравенство

$$\mathbf{P}\{\underline{\theta}(\vec{X}_n) \leqslant \theta \leqslant \overline{\theta}(\vec{X}_n)\} = \gamma.$$

В этом случае говорят об *интервальной оценке* для θ. Интервал

$$(\underline{\theta}(\vec{X}_n), \overline{\theta}(X_n))$$

называют *доверительным интервалом* для θ с *коэффициентом доверия* γ.

Доверительные интервалы обсуждаются в **3**.

Проверка статистических гипотез. ***Статистической гипотезой*** называют любое предположение о распределении вероятностей наблюдаемой случайной величины — скалярной или векторной.

В некотором смысле задача проверки статистической гипотезы является обратной к задаче оценивания параметра. При оценивании параметра мы ничего не знаем о его истинном значении. При проверке статистической гипотезы мы из каких-то

соображений предполагаем известным его значение и хотим по результатам эксперимента проверить наше предположение.

Примерами гипотез могут служить следующие предположения о вероятностных свойствах наблюдаемых случайных величин:

1) $\mu = \mu_0$, где μ — математическое ожидание случайной величины X (гипотеза о величине математического ожидания);

2) $\sigma_1^2 = \sigma_2^2$, где σ_1^2 и σ_2^2 — дисперсии случайных величин X_1 и X_2 (гипотеза об однородности дисперсий);

3) $F(x) = F_T(x)$, где $F(x)$ — неизвестная функция распределения наблюдаемой случайной величины X, а $F_T(x)$ — некоторая предполагаемая исследователем функция распределения (гипотеза о виде распределения).

Установление формы и степени связи между случайными величинами. Методы математической статистики, способствующие установлению формы и степени связи между случайными величинами, излагаются в таких разделах математической статистики, как корреляционный анализ, дисперсионный анализ, регрессионный анализ и др.

Смысл таких задач поясним на простом примере. Пусть Y — случайная величина, поведение которой мы хотели бы определять по значениям двух других случайных величин X_1 и X_2. Например, Y — это степень шума двигателя автомашины, а X_1 и X_2 — соответственно величина пробега автомобиля и вес груза в нем. Корреляционный и дисперсионный анализ позволяет нам ответить на вопрос: есть ли связь между X_1, X_2 и Y и насколько она существенна. На основе же регрессионного анализа мы можем построить так называемую регрессионную модель в виде зависимости

$$y = \varphi(x_1, x_2),$$

где y — среднее значение шума Y в зависимости от значений x_1 и x_2 случайных величин X_1 и X_2. Наличие такой модели

(которую строят, опираясь на результаты имеющихся *статистических данных* — результатов эксплуатации автомобилей) позволяет в дальнейшем выбрать наилучший режим эксплуатации и решать многие другие задачи.

Подобные задачи рассмотрены в 6–8.

1.3. Предварительная обработка результатов эксперимента

Прежде чем перейти к детальному анализу полученных в результате проведенного эксперимента *статистических данных*, обычно проводят их предварительную обработку. Иногда результаты такой обработки уже сами по себе дают ответы на многие вопросы. Но в большинстве случаев они служат исходным материалом для дальнейшего анализа.

Вариационный ряд. Одним из самых простых преобразований статистических данных является их упорядочивание по величине. Пусть $(x_1, \ldots, x_n)$ — *выборка объема* n *из генеральной совокупности* X. Ее можно упорядочить, расположив значения в неубывающем порядке:

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

где $x_{(1)}$ — наименьший, $x_{(n)}$ — наибольший из *элементов выборки.*

Определение 1.1. Последовательность чисел

$$x_{(1)}, \quad x_{(2)}, \quad \ldots, x_{(i)}, \quad \ldots, \quad x_{(n)},$$

удовлетворяющих условию (1.2), называют ***вариационным рядом выборки***, или, для краткости, просто ***вариационным рядом***; число $x_{(i)}$, $i = \overline{1, n}$, называют i-м ***членом вариационного ряда.***

Обозначим $X_{(i)}$, $i=\overline{1,n}$, случайную величину, которая при каждой *реализации случайной выборки* $\vec{X}_n$ принимает значение, равное i-му члену вариационного ряда.

Определение 1.2. Последовательность случайных величин

$$X_{(1)},\quad X_{(2)},\quad \ldots,\quad X_{(i)},\quad \ldots,\quad X_{(n)}$$

называют ***вариационным рядом случайной выборки.*** При этом $X_{(i)}$, $i=\overline{1,n}$, называют i-м ***членом вариационного ряда случайной выборки.***

Переход от случайной выборки $\vec{X}_n$ к ее вариационному ряду не приводит к потере информации, содержащейся в случайной выборке, поскольку их совместная функция распределения (1.1) остается одной и той же. Однако функция распределения каждой случайной величины $X_{(i)}$, $i=\overline{1,n}$, уже не совпадает с функцией распределения $F(x)$ генеральной совокупности X, хотя и может быть через нее выражена. Например, можно показать (см. пример 2.20), что для ***крайних членов вариационного ряда*** случайной выборки $X_{(1)}$ и $X_{(n)}$ их функции распределения имеют вид

$$\mathbf{P}\{X_{(1)}<x\}=1-\big(1-F(x)\big)^n$$

и

$$\mathbf{P}\{X_{(n)}<x\}=F^n(x).$$

Эти соотношения позволяют находить неизвестную функцию распределения $F(x)$ генеральной совокупности X, имея в эксперименте лишь результаты измерений либо величины $X_{(1)}$, либо $X_{(n)}$.

Пример 1.3. В результате пяти повторных независимых наблюдений некоторой случайной величины X (например, X — давление в газовом баллоне, измеряемое в мегапаскалях) полу-

чены следующие ее значения:

$$x_1 = 10{,}4; \quad x_2 = 9{,}5; \quad x_3 = 10{,}7; \quad x_4 = 9{,}3; \quad x_5 = 10{,}1.$$

Для данной выборки объема $n = 5$ вариационный ряд имеет вид

$$x_{(1)} = 9{,}3; \quad x_{(2)} = 9{,}5; \quad x_{(3)} = 10{,}1; \quad x_{(4)} = 10{,}4; \quad x_{(5)} = 10{,}7.$$

Статистический ряд. Среди *элементов выборки* $x_1, \ldots, x_n$ (а значит, и среди членов вариационного ряда $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$) могут быть одинаковые. Так бывает, либо когда наблюдаемая случайная величина X — дискретная, либо когда X — непрерывная, но ее значения при измерениях округляют.

Пусть среди элементов выборки $x_1, \ldots, x_n$ выделены $m < n$ их различных значений, расположенных в порядке возрастания. Обозначим их $z_{(1)}, \ldots, z_{(m)}$. Предположим, что каждое из них повторяется соответственно $n_1, \ldots, n_m$ раз, причем, разумеется, $\sum\limits_{i=1}^{m} n_i = n$.

Определение 1.3. ***Статистическим рядом*** для выборки называют таблицу, которая в первой строке содержит значения $z_{(1)}, \ldots, z_{(m)}$ (напомним, что $z_{(1)} < \ldots < z_{(m)}$), а во второй — числа их повторений (табл. 1.1). Число n_i, $i = \overline{1, m}$, показывающее, сколько раз встречался элемент $z_{(i)}$ в выборке, называют ***частотой***, а отношение n_i/n — ***относительной частотой*** этого значения.

Таблица 1.1

$z_{(1)}$	$z_{(2)}$	$\ldots$	$z_{(m)}$
n_1	n_2	$\ldots$	n_m

Статистические данные, представленные в виде статистического ряда, называют ***группированными.***

Исходные данные группируют обычно при больших объемах выборки (свыше 50), причем не только в виде статистического ряда, но и следующим образом: отрезок $J = [x_{(1)}, x_{(n)}]$, содержащий все выборочные значения, разбивают на m промежутков J_i, как правило одинаковой длины Δ. При этом считают, что каждый промежуток содержит свой левый конец, но лишь последний промежуток содержит и свой правый конец. При таком соглашении каждая точка отрезка J содержится в одном и только в одном промежутке J_i. Далее, для каждого промежутка J_i, $i = \overline{1, m}$, подсчитывают число n_k элементов выборки, попавших в него (при этом $n = n_1 + \ldots + n_m$), а результаты представляют в виде табл. 1.2, которую называют ***интервальным статистическим рядом.***

Таблица 1.2

J_1	J_2	...	J_m	
n_1	n_2	...	n_m	$\sum\limits_{i=1}^{m} n_k$

Иногда в верхней строке табл. 1.2 указывают не интервал, а его середину $\widetilde{x}_k$, а в нижней строке вместо частоты n_k записывают относительную частоту n_k/n.

Число промежутков m, на которые разбивают отрезок J, выбирают в зависимости от объема выборки n. Для ориентировочной оценки величины m можно пользоваться следующей формулой*:

$$m \approx \log_2 n + 1,$$

которая дает нижнюю оценку величины m и наиболее точна при больших значениях n. Например, при $n = 100$ она дает $m \geqslant 6$, а при $n = 1000$ — $m \geqslant 9$.

*См.: *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.*, 1983.

Пример 1.4. В течение суток измеряют напряжение X тока в электросети в вольтах. В результате опыта получена выборка объема $n = 30$:

107;	108;	110;	109;	110;	111;	109;	110;	111;	107;
108;	109;	110;	108;	107;	110;	109;	111;	111;	110;
109;	112;	113;	110;	106;	110;	109;	110;	108;	112.

Построим статистический ряд этой выборки.

Наименьшее значение в выборке $x_{(1)} = 106$, наибольшее — $x_{(8)} = 113$. Подсчитываем частоту n_k, $k = \overline{1,8}$, каждого из восьми различных значений в выборке и строим табл. 1.3.

Таблица 1.3

$z_{(k)}$	106	107	108	109	110	111	112	113
n_k	1	3	4	6	9	4	2	1

Эмпирическая и выборочная функции распределения. Рассмотрим функцию $n(x, \vec{X}_n)$, которая для каждого значения $x \in \mathbb{R}$ и каждой реализации $\vec{x}_n$ случайной выборки $\vec{X}_n$ принимает значение, равное числу элементов в выборке $\vec{x}_n$, меньших x.

Определение 1.4. Функцию

$$\widehat{F}(x, \vec{X}_n) = \frac{n(x, \vec{X}_n)}{n}, \tag{1.3}$$

где n — объем случайной выборки, будем называть ***выборочной функцией распределения.***

Согласно определению 1.4, при любом фиксированном x функция $\widehat{F}(x; \vec{X}_n)$ есть случайная величина, которая принимает одно из значений

$$0, \quad \frac{1}{n}, \quad \frac{2}{n}, \quad \ldots, \quad \frac{n-1}{n}, \quad \frac{n}{n} = 1$$

и имеет биномиальное распределение с параметром p, равным значению функции распределения генеральной совокупности X в точке x, т.е. $p = F(x)$.

Теорема 1.1. Для любого фиксированного x последовательность случайных величин $\{\widehat{F}(x;\vec{X}_n)\}$ сходится по вероятности при $n \to \infty$ к значению $F(x)$ функции распределения генеральной совокупности X в точке x.

◀ При любом фиксированном x выборочная функция распределения $\widehat{F}(x;\vec{X}_n)$ есть относительная частота события $\{X < x\}$. В соответствии с законом больших чисел в форме Бернулли, относительная частота при $n \to \infty$ сходится по вероятности к вероятности события $\{X < x\}$. Следовательно,

$$\widehat{F}(x;\vec{X}_n) \xrightarrow[n\to\infty]{\mathbf{P}} \mathbf{P}\{X < x\} = F(x). \qquad \blacktriangleright$$

Для каждой реализации $\vec{x}_n$ функцию $n(x,\vec{x}_n)$ аргумента x в дальнейшем будем обозначать $n(x)$.

Определение 1.5. ***Эмпирической функцией распределения*** называют скалярную функцию $F_n(x)$, которая определена для любого $x \in \mathbb{R}$ следующим образом:

$$F_n(x) = \frac{n(x)}{n}, \tag{1.4}$$

где n — объем выборки.

Функция $F_n(x)$ обладает всеми свойствами функции распределения. При этом она кусочно постоянна и изменяется скачками в каждой точке $x_{(i)}$ ($x_{(i)}$ — i-й член вариационного ряда).

Если все выборочные значения $x_1, \ldots, x_n$ различны, то функцию $F_n(x)$ можно записать в следующем виде:

$$F_n(x) = \begin{cases} 0, & x \leqslant x_{(1)}; \\ \dfrac{i}{n}, & x_{(i)} < x \leqslant x_{(i+1)}, \quad i = \overline{1, n-1}; \\ 1, & x > x_{(n)}, \end{cases}$$

т.е. в каждой точке $x_{(i)}$ функция $F_n(x)$ имеет скачок величиной $1/n$.

График функции $F_n(x)$ изображен на рис. 1.1.

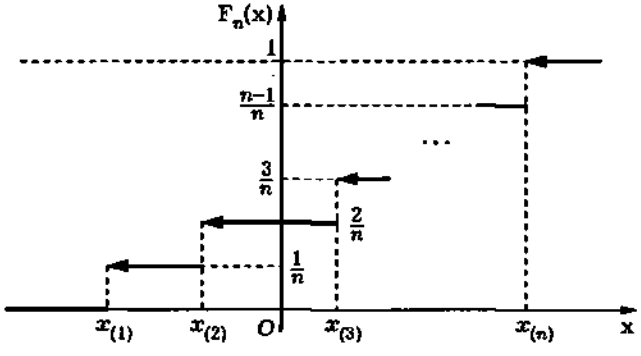


Рис. 1.1

Замечание 1.2. Функция $F_n(x)$ позволяет любую выборку $(x_1, \ldots, x_n)$ интерпретировать как генеральную совокупность $\widetilde{X}$, все значения которой равновероятны, т.е.

$$\mathrm{P}\{\widetilde{X} = x_i\} = \frac{1}{n}, \quad i = \overline{1, n}.$$

Такая интерпретация позволит в дальнейшем рассматривать числовые характеристики случайной величины $\widetilde{X}$ как приближенные значения соответствующих числовых характеристик исходной генеральной совокупности X. #

Из сказанного выше следует, что функция $F_n(x)$ является статистическим аналогом функции распределения $F(x)$ генеральной совокупности X. Функцию распределения $F(x)$ генеральной совокупности X в математической статистике называют иногда ***теоретической функцией распределения.***

В случае *непрерывной статистической модели* и большого объема выборки (свыше 50) экспериментальные данные удобнее представлять в виде интервального статистического ряда

(см. табл. 1.2). Разделив частоты $n_i/n = \widehat{p}_i$ на длину Δ интервалов J_i получим значения $n_i/(n\Delta)$, $i = \overline{1, m}$.

Определение 1.6. ***Эмпирической плотностью распределения***, соответствующей реализации $\vec{x}_n$ случайной выборки $\vec{X}_n$ из генеральной совокупности X, называют функцию $p_n(x)$, которая во всех точках интервала J_i, $i = \overline{1, m}$, принимает значение $\frac{n_i}{n\Delta}$, а вне интервала J равна нулю, т.е.

$$p_n(x) = \begin{cases} \dfrac{n_i}{n\Delta}, & x \in J_i; \\ 0, & x \notin J. \end{cases} \tag{1.5}$$

График функции $p_n(x)$, представляющий собой кусочно постоянную функцию на промежутке $J = [z_{(1)}, z_{(m)}]$, называют ***гистограммой*** (рис. 1.2).

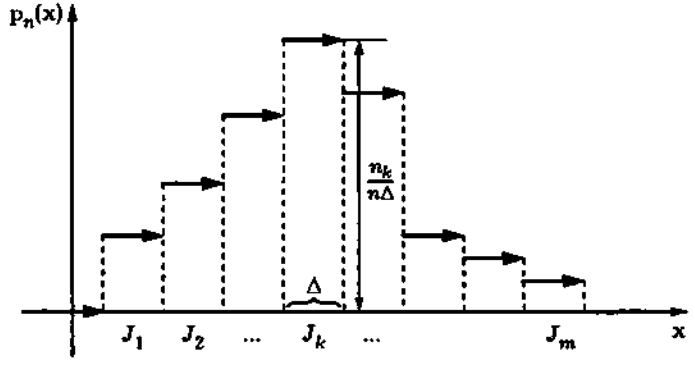


Рис. 1.2

Часто гистограммой называют диаграмму, составленную из прямоугольников с основанием Δ и высотами $n_i/(n\Delta)$, $i = \overline{1, m}$. Нетрудно увидеть, что суммарная площадь всех прямоугольников, образующих такую диаграмму, равна 1, так как

$$\sum_{i=1}^{m} \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{m} n_i = 1.$$

Кроме того, площадь каждого прямоугольника n_i/n есть частота попадания элементов выборки в соответствующий интервал J_i статистического ряда.

Рассмотрим случайную величину $n_i(\vec{X}_n)/n$, которая для каждой реализации $\vec{x}_n$ случайной выборки $\vec{X}_n$ равна частоте n_i/n. В соответствии с законом больших чисел в форме Бернулли $n_i(\vec{X}_n)/n$ при $n \to \infty$ будет сходиться по вероятности к вероятности попадания случайной величины X в промежуток J_i, $i = \overline{1,m}$, т.е.

$$\frac{n_i(\vec{X}_n)}{n} \xrightarrow[n\to\infty]{\mathbf{P}} \mathbf{P}\{X \in J_i\} = \int\limits_{J_i} p(x)\,dx,$$

где $p(x)$ — плотность распределения генеральной совокупности X. Если длина Δ промежутков достаточно мала и объем выборки n велик, то с вероятностью, близкой к 1, можно утверждать, что

$$\frac{n_i}{n} \approx p(\widetilde{x}_i)\Delta,$$

или

$$\frac{n_i}{n\Delta} \approx p(\widetilde{x}_i),$$

где $\widetilde{x}_i$ — середина промежутка J_i, $i = \overline{1,m}$. Таким образом, при большом объеме выборки n и достаточно малом Δ с вероятностью, близкой к 1, можно считать, что $p_n(x) \approx p(x)$. Иными словами, функция $p_n(x)$ является статистическим аналогом плотности распределения $p(x)$, наблюдаемой в эксперименте случайной величины X.

Наряду с гистограммой часто используют другое графическое представление для приближенного описания функции $p(x)$, которое называют ***полигоном частот***. По определению полигон частот — это ломаная, отрезки которой соединяют середины горизонтальных отрезков, образующих прямоугольники в гистограмме (рис. 1.3). Полигон частот используют

также в том случае, когда в эксперименте наблюдают дискретную случайную величину X. В этом случае по оси абсцисс откладывают все возможные (различные) значения случайной величины X, полученные в эксперименте, а по оси ординат — соответствующие частоты $\widehat{p}_i = n_i/n$, и соседние точки соединяют отрезками прямой.

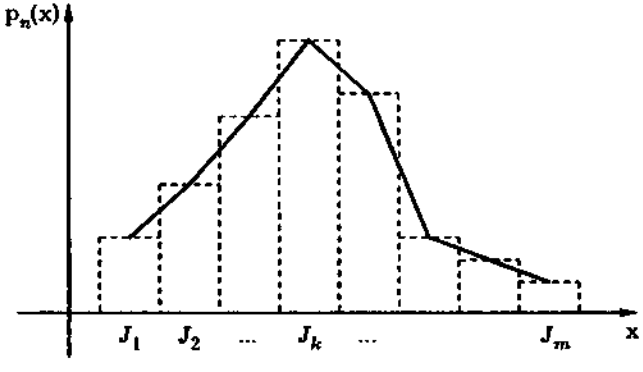


Рис. 1.3

Пример 1.5. Измерен рост $n = 500$ студентов. Результаты измерений представлены в виде интервального статистического ряда (табл. 1.4).

Таблица 1.4

[145, 150)	[150, 155)	[155, 160)	[160, 165)	[165, 170)
1	2	28	90	169
[170, 175)	[175, 180)	[180, 185)	[185, 190)	[190, 195]
132	55	16	6	1

Построим гистограмму — график эмпирической плотности распределения роста студентов.

Для построения гистограммы (рис. 1.4) нужно найти выборочную плотность распределения, используя формулу (1.5) и

учитывая, что $\Delta = 5$:

$$p_n(x) = \begin{cases} \frac{1}{2500}, & x \in [145, 150); \\ \frac{2}{2500}, & x \in [150, 155); \\ \frac{28}{2500}, & x \in [155, 160); \\ \frac{90}{2500}, & x \in [160, 165); \\ \frac{169}{2500}, & x \in [165, 170); \\ \frac{132}{2500}, & x \in [170, 175); \\ \frac{55}{2500}, & x \in [175, 180); \\ \frac{16}{2500}, & x \in [180, 185); \\ \frac{6}{2500}, & x \in [185, 190); \\ \frac{1}{2500}, & x \in [190, 195]. \end{cases}$$

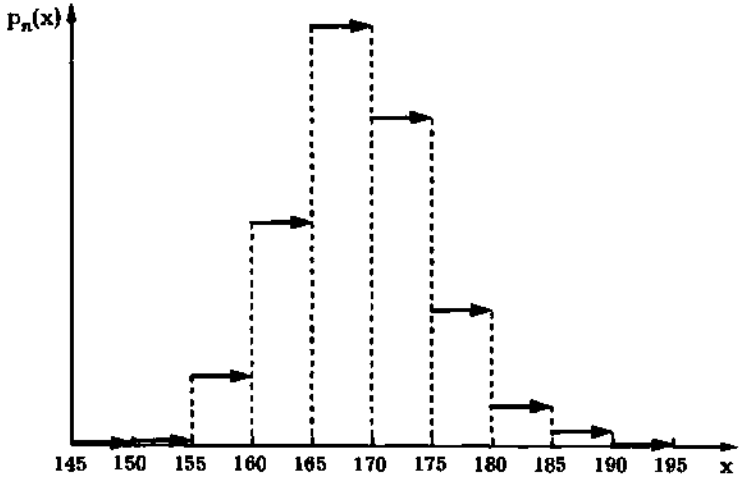


Рис. 1.4

Пример 1.6. В условиях примера 1.4 построим полигон частот. Для этого найдем относительные частоты каждого из элементов выборки и представим результаты в виде таблицы (табл. 1.5).

Таблица 1.5

$z_{(k)}$	106	107	108	109	110	111	112	113
$\frac{n_k}{n}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{4}{20}$	$\frac{6}{20}$	$\frac{9}{20}$	$\frac{4}{20}$	$\frac{2}{20}$	$\frac{1}{20}$

Построим точки с координатами $(z_{(i)}; n_i/n)$, $i = \overline{1,8}$, соединим их отрезками прямых (рис. 1.5).

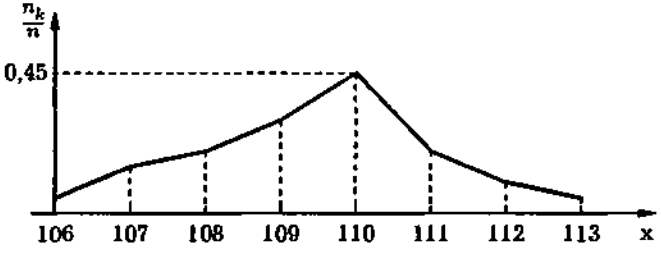


Рис. 1.5

Выборочные числовые моменты. Пусть $\vec{X}_n$ — случайная выборка из генеральной совокупности X с функцией распределения $F(x)$ (и плотностью распределения $p(x)$ в случае непрерывной статистической модели). Напомним [XVI], что, зная $p(x)$ или $F(x)$, можно записать математическое ожидание функции $g(X)$ в виде

$$\mathrm{M}g(X) = \int_{-\infty}^{\infty} g(x)\,p(x)\,dx$$

или

$$\mathrm{M}g(X) = \int_{-\infty}^{\infty} g(x)\,dF(x),$$

где последний интеграл есть интеграл Римана — Стилтьеса [XI]. При $g(X)=X^k$ или $g(X)=(X-\mathrm{M}X)^k$, $k\geqslant 1$, получаем соответственно начальные моменты m_k и центральные моменты $\mathring{m}_k$ k-го порядка случайной величины X:

$$m_k=\mathrm{M}(X^k)=\int\limits_{-\infty}^{\infty}x^k\,dF(x)=\int\limits_{-\infty}^{\infty}x^k p(x)\,dx,$$

$$\mathring{m}_k=\mathrm{M}(X-\mathrm{M}X)^k=\int\limits_{-\infty}^{\infty}(x-\mathrm{M}X)^k dF(x)=\int\limits_{-\infty}^{\infty}(x-\mathrm{M}X)^k p(x)\,dx.$$

В частности, при $g(X)=X$ и $g(X)=(X-\mathrm{M}X)^2$ получаем формулы соответственно для математического ожидания и дисперсии случайной величины X.

Все эти ***числовые характеристики*** в математической статистике называют ***теоретическими*** (или ***генеральными числовыми характеристиками***, т.е. относящимися к генеральной совокупности).

Так же, как функциям $F(x)$ или $p(x)$ мы сопоставили их статистические аналоги — эмпирические функции $F_n(x)$ и $p_n(x)$, построенные по выборке $\vec{x}_n$, можно каждой теоретической числовой характеристике сопоставить ее статистический аналог, если в соответствующих формулах, приведенных выше, заменить $F(x)$ на $F_n(x)$, а $p(x)$ на $p_n(x)$.

Статистические аналоги теоретических числовых характеристик можно получить из следующих соображений.

Как отмечалось в замечании 1.2, любую выборку $\vec{x}_n$ можно рассматривать в качестве генеральной совокупности дискретной случайной величины $\widetilde{X}$, все значения которой равновероятны, т.е.

$$\mathrm{P}\{\widetilde{X}=x_i\}=\frac{1}{n},\quad i=\overline{1,n}.$$

По определению начальные и центральные моменты k-го порядка такой генеральной совокупности соответственно равны

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \qquad \widehat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k,$$

где

$$\overline{x} = \widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.6}$$

Числа $\widehat{\mu}_k$, $\widehat{\nu}_k$ есть значения *выборочных характеристик* (*статистик*)

$$\widehat{\mu}_k(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad \text{и} \quad \widehat{\nu}_k(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k,$$

где

$$\overline{X} = \widehat{\mu}_1(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{1.7}$$

т.е. $\widehat{\mu}_k = \widehat{\mu}_k(\vec{x}_n)$, $\widehat{\nu}_k = \widehat{\nu}_k(\vec{x}_n)$.

Выборочную характеристику

$$\widehat{\mu}_k(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

называют ***выборочным начальным моментом k-го порядка.*** В частности, выборочный начальный момент первого порядка $\overline{X} = \widehat{\nu}_1(\vec{X}_n)$ называют ***выборочным средним.***

Выборочную характеристику

$$\widehat{\nu}_k(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k \tag{1.8}$$

называют ***выборочным центральным моментом k-го порядка.*** В частности, выборочный центральный момент 2-го порядка $\widehat{\sigma}^2(\vec{X}_n) = \widehat{\nu}_2(\vec{X}_n)$ называют ***выборочной дисперсией.***

Выборочную характеристику $\widehat{\sigma}(\vec{X}_n) = \sqrt{\widehat{\sigma}^2(\vec{X}_n)}$ называют ***выборочным средним квадратичным отклонением.*** Величины

$$\widehat{\sigma}^2 = \widehat{\sigma}^2(\vec{x}_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \tag{1.9}$$

$$\widehat{\sigma} = \widehat{\sigma}(\vec{x}_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1.10}$$

являются статистическими аналогами соответственно дисперсии $\sigma^2 = \mathbf{D}X$ и среднего квадратичного отклонения генеральной совокупности X.

Числа $\overline{x}$, $\widehat{\sigma}^2$, $\widehat{\sigma}$, $\widehat{\mu}_k$, $\widehat{\nu}_k$ будем называть соответственно ***средним значением*** (или средним), ***дисперсией***, ***средним квадратичным отклонением***, ***начальным моментом*** и ***центральным моментом k-го порядка выборки.***

Выборочные характеристики можно ввести также и при рассмотрении выборок из многомерных генеральных совокупностей.

Так, например, рассмотрим случайную выборку $(\vec{X}_n, \vec{Y}_n)$ объема n из двумерной генеральной совокупности (X, Y).

Выборочную характеристику

$$\widehat{K}(\vec{X}_n, \vec{Y}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{1.11}$$

называют ***выборочным корреляционным моментом.***

Выборочную характеристику

$$\widehat{\rho}(\vec{X}_n, \vec{Y}_n) = \frac{\widehat{K}(\vec{X}_n, \vec{Y}_n)}{\widehat{\sigma}_x(\vec{X}_n)\,\widehat{\sigma}_y(\vec{Y}_n)}, \tag{1.12}$$

где

$$\widehat{\sigma}_x^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad \widehat{\sigma}_y^2(\vec{Y}_n) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

называют ***выборочным коэффициентом корреляции.***

Значения $\widehat{K}(\vec{x}_n, \vec{y}_n)$ и $\widehat{\rho}(\vec{x}_n, \vec{y}_n)$ выборочного корреляционного момента и выборочного коэффициента корреляции, где $(\vec{x}_n, \vec{y}_n)$ — реализация случайной выборки $(\vec{X}_n, \vec{Y}_n)$, будем соответственно обозначать $\widehat{K}_{xy}$ и $\widehat{\rho}_{xy}$, называя ***корреляционным моментом выборки*** $(\vec{x}_n, \vec{y}_n)$ и ***коэффициентом корреляции выборки*** $(\vec{x}_n, \vec{y}_n)$.

Замечание 1.3. При больших n от выборки $\vec{x}_n$ часто переходят к интервальному статистическому ряду. При этом значения $\overline{x}$, $\widehat{\sigma}^2$ и $\widehat{K}_{XY}$ соответственно вычисляют по формулам

$$\overline{x} = \sum_{i=1}^{m} \widehat{p}_i \widetilde{x}_i, \tag{1.13}$$

$$\widehat{\sigma}^2 = \sum_{i=1}^{m} \widehat{p}_i \left(\widetilde{x}_i - \overline{x}\right)^2, \tag{1.14}$$

$$\widehat{K}_{XY} = \sum_{i=1}^{m} \widehat{p}_i \left(\widetilde{x}_i - \overline{x}\right)\left(\widetilde{y}_i - \overline{y}\right), \tag{1.15}$$

где $\widehat{p}_i = n_i/n$ — относительная частота события $\{X \in J_i\}$, $i = \overline{1, m}$, а $\widetilde{y}_i$ и $\overline{y}$ имеют тот же самый смысл, что и $\widetilde{x}_k$ и $\overline{x}$, но для случайной величины Y. #

Основное свойство выборочных моментов, как начальных, так и центральных, и в том числе выборочного среднего $\overline{X}$ и выборочной дисперсии $\widehat{\sigma}^2(\vec{X}_n)$, состоит в том, что при увеличении объема выборки n они сходятся по вероятности к соответствующим теоретическим (генеральным) моментам*. В частности, при $n \to \infty$ имеем $\vec{X} \xrightarrow[n\to\infty]{\mathrm{P}} \mathrm{M}X$, а $\widehat{\sigma}^2(\vec{X}_n) \xrightarrow[n\to\infty]{\mathrm{P}} \mathrm{D}X$.

Более того, можно показать, что распределение выборочных моментов является асимптотически (при $n \to \infty$) нормальным. Точные формулировки этих утверждений в некоторых частных случаях будут приведены в дальнейшем изложении.

*См.: *Крамер Г.*

1.4. Решение типовых примеров

Пример 1.7. В результате эксперимента получена *выборка объема* $n = 79$:

2; 4; 2; 4; 3; 3; 3; 2; 0; 6; 1; 2; 3; 2; 2;
4; 3; 3; 5; 1; 0; 2; 4; 3; 2; 2; 3; 3; 1; 3;
3; 3; 1; 1; 2; 3; 1; 4; 3; 1; 7; 4; 3; 4; 2;
3; 2; 3; 3; 1; 4; 3; 1; 4; 5; 3; 4; 2; 4; 5;
3; 6; 4; 1; 3; 2; 4; 1; 3; 1; 0; 0; 4; 6; 4;
7; 4; 1; 3.

Построим *статистический ряд*, *полигон частот*, *эмпирическую функцию распределения* и нарисуем ее график, найдем $\overline{x}$, $\widehat{\sigma}^2$, $\widehat{\sigma}$.

Наименьший *элемент выборки* (первый *член вариационного ряда*) $x_{(1)} = 0$, наибольший — $x_{(79)} = 7$. Составим статистический ряд, расположив все элементы выборки в порядке возрастания (табл. 1.6).

Таблица 1.6

$x_{(k)}$	0	1	2	3	4	5	6	7	
n_k	4	13	14	24	16	3	3	2	$\sum\limits_{k=1}^{8} n_k = 79$

Статистический ряд содержит восемь элементов: 0, 1, 2, 3, 4, 5, 6, 7. Для построения полигона частот (рис. 1.6) следует вычислить относительные частоты n_k/n каждого из элементов статистического ряда:

$$\frac{n_1}{n} = \frac{4}{79} \approx 0{,}0506; \quad \frac{n_2}{n} = \frac{13}{79} \approx 0{,}1646; \quad \frac{n_3}{n} = \frac{14}{79} \approx 0{,}1772;$$

$$\frac{n_4}{n} = \frac{24}{79} \approx 0{,}3038; \quad \frac{n_5}{n} = \frac{16}{79} \approx 0{,}2025; \quad \frac{n_6}{n} = \frac{3}{79} \approx 0{,}0380;$$

$$\frac{n_7}{n} = \frac{3}{79} \approx 0{,}0380; \quad \frac{n_8}{n} = \frac{2}{79} \approx 0{,}0253.$$

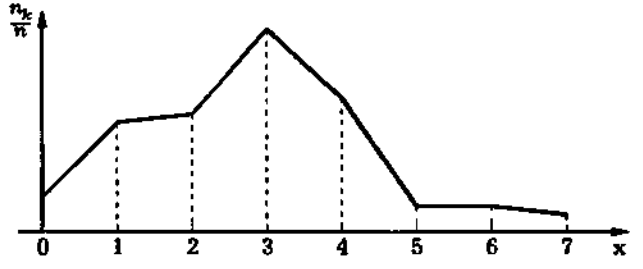


Рис. 1.6

Чтобы найти выборочную функцию распределения, нужно последовательно суммировать относительные частоты:

$$F_n(x) = \begin{cases} 0{,}0000, & x \leqslant 0; \\ 0{,}0506, & 0 < x \leqslant 1; \\ 0{,}2152, & 1 < x \leqslant 2; \\ 0{,}3924, & 2 < x \leqslant 3; \\ 0{,}6962, & 3 < x \leqslant 4; \\ 0{,}8987, & 4 < x \leqslant 5; \\ 0{,}9367, & 5 < x \leqslant 6; \\ 0{,}9747, & 6 < x \leqslant 7; \\ 1{,}0000, & x > 7. \end{cases}$$

График функции $F_n(x)$ — ступенчатая кривая (рис. 1.7).

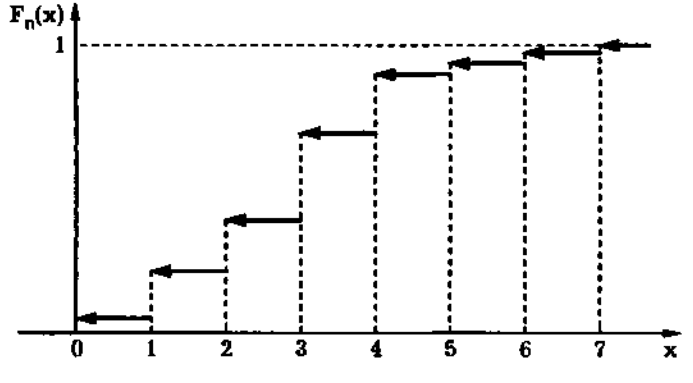


Рис. 1.7

Учитывая, что элементы выборки повторяются, с помощью формулы (1.6) находим *среднее значение выборки*:

$$\overline{x} = \frac{1}{79}(0\cdot 4 + 1\cdot 13 + 2\cdot 14 + 3\cdot 24 + 4\cdot 16 + 5\cdot 3 + 6\cdot 3 + 7\cdot 2) \approx 2{,}835.$$

С помощью формулы (1.9) находим *дисперсию выборки*

$$\widehat{\sigma}^2 = \frac{1}{79}((0-2{,}84)^2\cdot 4 + (1-2{,}84)^2\cdot 13 + (2-2{,}84)^2\cdot 14 + (3-2{,}84)^2\cdot 24 + (4-2{,}84)^2\cdot 16 + (5-2{,}84)^2\cdot 3 + (6-2{,}84)^2\cdot 3 + (7-2{,}84)^2\cdot 2) \approx 2{,}3668;$$

а с помощью формулы (1.10) — *среднее квадратичное отклонение выборки*

$$\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} \approx 1{,}54.$$

Пример 1.8. Измерена максимальная емкость 20 подстроечных конденсаторов, и результаты измерений (в пикофарадах) приведены в табл. 1.7. Составим статистический ряд и построим *гистограмму*.

Таблица 1.7

Номер конденсатора	1	2	3	4	5	6	7	8	9	10
Емкость, пФ	4,40	4,31	4,40	4,40	4,65	4,56	4,71	4,54	4,36	4,56
Номер конденсатора	11	12	13	14	15	16	17	18	19	20
Емкость, пФ	4,31	4,42	4,60	4,35	4,50	4,40	4,43	4,48	4,42	4,45

Статистический ряд представлен в табл. 1.8. Наименьшее значение выборки $x_{(1)} = 4{,}31$, наибольшее — $x_{(20)} = 4{,}71$.

Для построения гистограммы результаты наблюдений представим в виде *интервального статистического ряда*, разбив отрезок [4,31, 4,71] на пять равных промежутков (табл. 1.9).

Таблица 1.8

$x_{(i)}$	4,31	4,35	4,36	4,40	4,42	4,43	4,45	
n_i	2	1	1	4	2	1	1	
$x_{(i)}$	4,48	4,50	4,54	4,56	4,60	4,65	4,71	
n_i	1	1	1	2	1	1	1	$\sum\limits_{i=1}^{14} n_i = 20$

Таблица 1.9

J_k	[4,31, 4,39)	[4,39, 4,47)	[4,47, 4,55)	[4,55, 4,63)	[4,63, 4,71]
n_k	4	8	3	3	2

Длина Δ каждого полученного промежутка равна 0,08. Определим *эмпирическую плотность распределения*, используя формулу (1.5):

$$p_n(x) = \begin{cases} \frac{4}{20 \cdot 0{,}08} = 2{,}500, & x \in [4{,}31,\ 4{,}39); \\ \frac{8}{20 \cdot 0{,}08} = 5{,}000, & x \in [4{,}39,\ 4{,}47); \\ \frac{3}{20 \cdot 0{,}08} = 1{,}875, & x \in [4{,}47,\ 4{,}55); \\ \frac{3}{20 \cdot 0{,}08} = 1{,}875, & x \in [4{,}55,\ 4{,}63); \\ \frac{2}{20 \cdot 0{,}08} = 1{,}250, & x \in [4{,}63,\ 4{,}71]; \\ 0, & x \notin [4{,}31,\ 4{,}71]. \end{cases}$$

График функции $p_n(x)$ (гистограмма) представлен на рис. 1.8.

Пример 1.9. В результате измерения диаметров 200 валиков из партии, изготовленной одним станком-автоматом, получены отклонения измеренных диаметров от номинала (в микрометрах). *Группированные данные* представлены в виде интервального статистического ряда (табл. 1.10). Найдем среднее значение $\overline{x}$ и дисперсию $\widehat{\sigma}^2$ выборки.

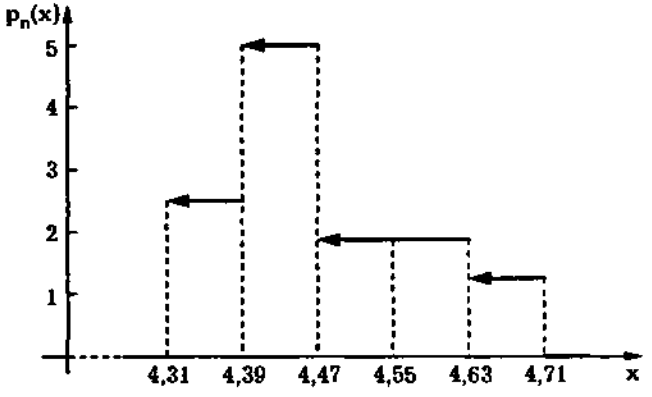


Рис. 1.8

Таблица 1.10

J_i	$[-20,-15)$	$[-15,-10)$	$[-10,-5)$	$[-5,0)$	$[0,5)$	
n_i	7	11	15	24	49	
J_i	$[5,10)$	$[10,15)$	$[15,20)$	$[20,25)$	$[25,30]$	
n_i	41	26	17	7	3	$\sum\limits_{i=1}^{10} n_i = 200$

Обозначив через $\widetilde{x}_{(i)}$ середины промежутков J_i, $i = \overline{1,10}$, представим группированные данные в виде табл. 1.11.

Таблица 1.11

$\widetilde{x}_i$	$-17{,}5$	$-12{,}5$	$-7{,}5$	$-2{,}5$	$2{,}5$	$7{,}5$	$12{,}5$	$17{,}5$	$22{,}5$	$27{,}5$	
n_i	7	11	15	24	49	41	26	17	7	3	$\sum\limits_{i=1}^{10} n_i = 200$

Среднее значение, согласно формуле (1.13), можно представить следующим образом:

$$\overline{x} = \sum_{i=1}^{m} \frac{n_i}{n} \widetilde{x}_{(i)} = \frac{1}{n} \sum_{i=1}^{m} n_i \widetilde{x}_{(i)}.$$

В данном случае $n = 200$, $m = 10$, а значения n_i и $\widetilde{x}_{(i)}$ даны в табл. 1.11. Вычисляя, находим

$$\overline{x} = \frac{1}{200}\Big(-7 \cdot 17{,}5 - 11 \cdot 12{,}5 - 15 \cdot 7{,}5 - 24 \cdot 2{,}5 + 49 \cdot 2{,}5 + \\ + 41 \cdot 7{,}5 + 26 \cdot 12{,}5 + 17 \cdot 17{,}5 + 7 \cdot 22{,}5 + 3 \cdot 27{,}5\Big) = 4{,}3.$$

Дисперсию выборки находим по тем же данным с помощью формулы (1.14):

$$\widehat{\sigma}^2 = \sum_{i=1}^{m} \frac{n_i}{n}\big(\widetilde{x}_{(i)} - \overline{x}\big) = \frac{1}{200}\sum_{i=1}^{10} n_i\big(\widetilde{x}_{(i)} - 4{,}3\big)^2 = \frac{1}{200}\Big(7 \cdot 21{,}8^2 + \\ + 11 \cdot 16{,}8^2 + 15 \cdot 11{,}8^2 + 24 \cdot 6{,}8^2 + 49 \cdot 1{,}8^2 + 41 \cdot 3{,}2^2 + \\ + 26 \cdot 8{,}2^2 + 17 \cdot 13{,}2^2 + 7 \cdot 18{,}2^2 + 3 \cdot 23{,}2^2\Big) \approx 83{,}84.$$

Пример 1.10. Из двумерной *генеральной совокупности* (X, Y) получена выборка объема $n = 20$:

(1365, 0,28);	(1375, 0,38);	(1375, 0,42);	(1375, 0,31);
(1405, 0,33);	(1410, 0,47);	(1410, 0,60);	(1420, 0,47);
(1425, 0,50);	(1415, 0,66);	(1440, 0,65);	(1385, 0,37);
(1390, 0,53);	(1395, 0,38);	(1450, 0,85);	(1450, 0,93);
(1455, 0,60);	(1475, 1,68);	(1480, 1,45);	(1485, 1,80).

Найдем значение *корреляционного момента выборки*

$$\widehat{K}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

По выборке

1365; 1375; 1375; 1375; 1405; 1410; 1410;
1420; 1425; 1415; 1440; 1385; 1390; 1395;
1450; 1450; 1455; 1475; 1480; 1485

находим

$$\overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 1419.$$

По выборке

0,28;	0,38;	0,42;	0,31;	0,33;	0,47;	0,60;
0,47;	0,50;	0,66;	0,65;	0,37;	0,53;	0,38;
0,85;	0,93;	0,60;	0,68;	1,45;	1,80	

находим

$$\overline{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 0{,}683 \approx 0{,}68.$$

В результате получаем

$$\begin{aligned}\widehat{K}_{xy} \approx \frac{1}{20}\Big((1365-1419)(0{,}28-0{,}68)+(1375-1419)(0{,}38-0{,}68)+\\ +(1375-1419)(0{,}42-0{,}68)+(1375-1419)(0{,}31-0{,}68)+\\ +(1405-1419)(0{,}33-0{,}68)+(1410-1419)(0{,}47-0{,}68)+\\ +(1410-1419)(0{,}60-0{,}68)+(1420-1419)(0{,}47-0{,}65)+\\ +(1425-1419)(0{,}50-0{,}68)+(1415-1419)(0{,}66-0{,}68)+\\ +(1440-1419)(0{,}65-0{,}68)+(1385-1419)(0{,}37-0{,}68)+\\ +(1390-1419)(0{,}53-0{,}68)+(1395-1419)(0{,}38-0{,}68)+\\ +(1450-1419)(0{,}85-0{,}68)+(1450-1419)(0{,}93-0{,}68)+\\ +(1455-1419)(0{,}60-0{,}68)+(1475-1419)(1{,}68-0{,}68)+\\ +(1480-1419)(1{,}45-0{,}68)+(1485-1419)(1{,}80-0{,}68)\Big) \approx 10{,}955.\end{aligned}$$

Вопросы и задачи

1.1. Что называют случайной выборкой, объемом выборки, элементом выборки, реализацией случайной выборки (выборкой)?

1.2. Что называют генеральной совокупностью?

1.3. Какие повторные наблюдения (эксперименты) называют независимыми?

1.4. Укажите связь между функцией распределения случайной выборки и функцией распределения генеральной совокупности.

1.5. Что такое статистика, выборочная характеристика?

1.6. Что такое выборочные распределения?

1.7. Что называют вариационным рядом случайной выборки, вариационным рядом выборки?

1.8. Что называют статистическим рядом?

1.9. Что такое интервальный статистический ряд?

1.10. Дайте определение выборочной и эмпирической функций распределения.

1.11. Дайте определение эмпирической плотности распределения.

1.12. Что такое гистограмма?

1.13. Что такое полигон?

1.14. Что называют выборочным средним, выборочной дисперсией, выборочными моментами, выборочным корреляционным моментом, выборочным коэффициентом корреляции?

1.15. Напишите выражения для среднего значения, дисперсии, начального и центрального моментов, корреляционного момента, коэффициента корреляции выборки.

1.16. По результатам измерений имеем выборку 2781, 2836, 2807, 2763, 2858. Составьте вариационный ряд, постройте эмпирическую функцию распределения и ее график. Вычислите $\overline{x}$, $\widehat{\sigma}^2$.

Ответ: $\overline{x} = 2809$; $\widehat{\sigma}^2 = 1206{,}8$.

1.17. Докажите, что имеет место равенство

$$\widehat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2.$$

1.18. По результатам измерений задана выборка

3,7;	6,2;	5,2;	5,7;	6,2;	4,7;	4,2;	6,7;
7,2;	5,2;	6,2;	4,7;	7,2;	5,2;	4,7;	5,7;
5,2;	4,7;	5,2;	5,7;	4,2;	6,7;	5,2;	6,2;
5,7;	6,7;	5,2;	5,7;	5,2;	4,2;	5,2;	4,7;
5,7;	4,2;	5,2;	6,2;	5,7;	6,2;	5,7;	4,2;
5,2;	5,7;	4,2;	5,2;	6,2;	7,2;	5,2;	4,7;
5,7;	6,2;	5,2;	4,7;	5,7;	6,7;	7,2;	6,7;
7,2;	3,7;	7,7;	3,2;	3,7;	7,7;	5,2;	4,7.

По выборке составьте статистический ряд, постройте гистограмму, эмпирическую функцию распределения и ее график. Вычислите значения числовых характеристик $\overline{x}$, $\widehat{\sigma}^2$, $\widehat{\sigma}$.

Ответ: $\overline{x} = 5{,}73$; $\widehat{\sigma}^2 = 1{,}187$; $\widehat{\sigma} = 1{,}06$.

1.19. При сверлении отверстий одним и тем же сверлом и последующем измерении диаметров отверстий получены данные, представленные в виде интервального статистического ряда (табл. 1.12). Найдите значения $\overline{x}$ и $\widehat{\sigma}_x$.

Таблица 1.12

$y_{(k)}$	[40,25, 40,28)	[40,28, 40,31)	[40,31, 40,34)	[40,34, 40,37)
n_k	2	10	18	25
$y_{(k)}$	[40,37, 40,40)	[40,40, 40,43)	[40,43, 40,46)	
n_k	12	8	5	$\sum_{i=1}^{n} n_i = 80$

Ответ: $\overline{x} \approx 40{,}355$; $\widehat{\sigma}_x \approx 0{,}04$.

1.20. Из двумерной генеральной совокупности сделана выборка объема $n = 60$ (данные приведены в табл. 1.13). Найдите значение выборочного коэффициента корреляции.

Таблица 1.13

	4100	4300	4500	4700	4900	5100	5300	5500
6,75		1						
6,25	1	2	2	1				
5,75		1	3	4	2	3		
5,25		3	5	7	1	1		
4,75			2	5	5	3	2	
4,25					1	2	2	
3,75								1

Ответ: $\widehat{\rho} = 0{,}63$.

2. ТОЧЕЧНЫЕ ОЦЕНКИ

Одной из задач математической статистики (см. 1.2) является *оценка* неизвестных параметров выбранной *параметрической модели.*

Очень часто в приложениях рассматривают параметрическую модель. В этом случае предполагают, что закон распределения *генеральной совокупности* принадлежит множеству $\left\{F(x;\vec{\theta})\colon \vec{\theta}\in\Theta\right\}$, где вид функции распределения задан, а вектор параметров $\vec{\theta}=(\theta_1, \ldots, \theta_r)$ неизвестен. Требуется найти оценку для $\vec{\theta}$ или некоторой функции от него (например, математического ожидания, дисперсии) по *случайной выборке* $(X_1, \ldots, X_n)$ из генеральной совокупности X.

Например, предположим, что масса X детали имеет нормальный закон распределения, но его параметры $\theta_1=\mu$ и $\theta_2=\sigma^2$ неизвестны. Нужно найти приближенное значение параметров по результатам наблюдений $x_1, \ldots, x_n$, полученным в эксперименте (по реализации случайной выборки).

Как уже отмечалось (см. 1.2), в математической статистике существуют два вида *оценок*: *точечные* и *интервальные.* В этой главе будут рассмотрены точечные оценки, а интервальным оценкам посвящена следующая глава.

2.1. Состоятельные, несмещенные и эффективные оценки

Пусть $\vec{X}_n=(X_1, \ldots, X_n)$ — *случайная выборка* из *генеральной совокупности* X, функция распределения $F(x;\theta)$ которой известна, а θ — неизвестный параметр, т.е. рассматривается *параметрическая модель* $\{F(x;\theta), \theta\in\Theta\}$ (для простоты изложения будем считать пока, что θ — скаляр).

Требуется построить *статистику* $\widehat{\theta}(\vec{X}_n)$, которую можно было бы принять в качестве *точечной оценки* параметра θ.

Интуитивно ясно, что в качестве *оценки* параметра θ можно использовать различные статистики. Например, в качестве точечной оценки для $\mu = \mathrm{M}\,X$ можно предложить такие статистики:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \theta^*(\vec{X}_n) = \frac{X_{(1)} + X_{(n)}}{2},$$

$$\widetilde{\theta}(\vec{X}_n) = \begin{cases} \dfrac{1}{2}\left(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right), & n \text{ — четное;} \\ X_{(\frac{n+1}{2})}, & n \text{ — нечетное.} \end{cases}$$

Какую же из этих статистик предпочесть? В общем случае нужно дать ответ на вопрос: какими свойствами должна обладать статистика $\widehat{\theta}(X_1, \ldots, X_n) = \theta(\vec{X}_n)$, чтобы она была в некотором смысле наилучшей оценкой параметра θ? Рассмотрению требований к оценкам и методам их нахождения посвящена настоящая глава.

Заметим, что в дальнейшем, как правило, будем говорить об оценке параметра θ параметрической модели, хотя все сказанное можно перенести и на функцию от θ.

Определение 2.1. Статистику $\widehat{\theta}(\vec{X}_n)$ называют ***состоятельной оценкой*** параметра $\theta \in \Theta$, если с ростом *объема выборки* n она сходится по вероятности к оцениваемому параметру θ, т.е.

$$\widehat{\theta}(\vec{X}_n) \xrightarrow[n\to\infty]{\mathrm{P}} \theta.$$

Иными словами, для состоятельной оценки $\widehat{\theta}(\vec{X}_n)$ отклонение ее от θ на величину ε и более становится маловероятным при большом объеме выборки. Это свойство оценки является очень важным, ибо несостоятельная оценка практически бесполезна. Однако следует отметить, что на практике приходится

оценивать неизвестные параметры и при малых объемах выборки.

Естественным является то требование, при выполнении которого оценка не дает систематической погрешности в сторону завышения (или занижения) истинного значения параметра θ.

Определение 2.2. Статистику $\widehat{\theta}(\vec{X}_n)$ называют ***несмещенной оценкой*** параметра θ, если ее математическое ожидание совпадает с θ, т.е. $\mathrm{M}\widehat{\theta}(\vec{X}_n) = \theta$ для любого фиксированного n.

Если оценка является ***смещенной*** (т.е. последнее равенство не имеет места), то величина смещения $b_n(\theta) = \mathrm{M}\widehat{\theta}(\vec{X}_n) - \theta$. Как мы увидим далее, смещение оценки часто можно устранить, введя соответствующую поправку.

Говорят также, что ***оценка*** $\widehat{\theta}(\vec{X}_n)$ является ***асимптотически несмещенной***, если при $n \to \infty$ она сходится по вероятности к своему математическому ожиданию, т.е. для любого $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\widehat{\theta}(\vec{X}_n) - \mathrm{M}\widehat{\theta}(\vec{X}_n)\right| < \varepsilon\right\} = 1.$$

Предположим, что имеются две несмещенные оценки $\widehat{\theta}(\vec{X}_n)$ и $\widetilde{\theta}(\vec{X}_n)$ для параметра θ. Если дисперсии $\mathbf{D}\widehat{\theta}(\vec{X}_n)$ и $\mathbf{D}\widetilde{\theta}(\vec{X}_n)$ удовлетворяют условию

$$\mathbf{D}\widehat{\theta}(\vec{X}_n) \leqslant \mathbf{D}\widetilde{\theta}(\vec{X}_n) \tag{2.1}$$

для любого фиксированного n и $\theta \in \Theta$, то следует предпочесть оценку $\widehat{\theta}(\vec{X}_n)$, поскольку разброс статистики $\widehat{\theta}(\vec{X}_n)$ относительно параметра θ меньше, чем разброс статистики $\widetilde{\theta}(\vec{X}_n)$.

Определение 2.3. Если в некотором классе несмещенных оценок параметра θ, имеющих конечную дисперсию, существует такая оценка $\widehat{\theta}(\vec{X}_n)$, что неравенство (2.1) выполняется для всех оценок $\widetilde{\theta}(\vec{X}_n)$ из этого класса, то говорят, что ***оценка*** $\widehat{\theta}(\vec{X}_n)$ является ***эффективной в данном классе оценок***.

Иными словами, дисперсия эффективной оценки параметра в некотором классе является минимальной среди дисперсий всех оценок из рассматриваемого класса несмещенных оценок.

Замечание 2.1. Эффективную оценку в классе всех несмещенных оценок будем называть ***эффективной оценкой***, не добавляя слов „в классе несмещенных оценок".

Замечание 2.2. В литературе по математической статистике при рассмотрении параметрических моделей вместо термина „эффективная оценка" в классе всех несмещенных оценок используют и другие: „несмещенная оценка с минимальной дисперсией", „оптимальная оценка".

Теорема 2.1. Оценка

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

(*выборочное среднее*) математического ожидания

$$\theta = \mathrm{M}X = \mu$$

генеральной совокупности X с конечной дисперсией является несмещенной, состоятельной и эффективной в классе всех ***линейных оценок***, т.е. оценок вида

$$\widetilde{\theta}(\vec{X}_n) = \sum_{i=1}^{n} \alpha_i X_i,$$

где $\sum\limits_{i=1}^{n} \alpha_i = 1$, для произвольной параметрической модели.

◀ Напомним, что элементы X_i, $i = \overline{1,n}$, случайной выборки $\vec{X}_n$ являются независимыми случайными величинами и распределенными так же, как и сама генеральная совокупность X. Следовательно, $\mathrm{M}X_i = \mathrm{M}X = \mu$ и $\mathrm{D}X_i = \mathrm{D}X = \sigma^2$, $i = \overline{1,n}$.

В силу свойств математического ожидания имеем

$$\mathrm{M}\overline{X} = \mathrm{M}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}X_i = \frac{1}{n}n\mu = \mu,$$

что и доказывает несмещенность оценки $\overline{X}$.

Далее, поскольку последовательность $X_1, \ldots, X_n$ состоит из независимых одинаково распределенных случайных величин с конечной дисперсией, то в силу закона больших чисел в форме Чебышева для любого $\varepsilon > 0$

$$\mathbf{P}\{|\overline{X} - \mu| < \varepsilon\} \to 1, \quad n \to \infty,$$

т.е. оценка $\overline{X}$ сходится по вероятности к оцениваемому параметру, а это и означает ее состоятельность.

Покажем теперь, что

$$\mathbf{D}\,\widetilde{\theta}(\vec{X}_n) = \mathbf{D}\sum_{i=1}^{n}\alpha_i X_i = \sum_{i=1}^{n}\mathbf{D}(\alpha_i X_i) = \sum_{i=1}^{n}\alpha_i^2\,\mathbf{D}\,X_i = \sigma^2\sum_{i=1}^{n}\alpha_i^2$$

достигает своего минимального значения при $\alpha_i = 1/n$, т.е. когда оценка $\widetilde{\theta}(\vec{X}_n) = \overline{X}$, что и означает эффективность оценки $\overline{X}$ в классе линейных оценок.

Для отыскания условного минимума функции [V]

$$g(\alpha_1, \ldots, \alpha_n) = \sum_{i=1}^{n}\alpha_i^2$$

при ограничении

$$\sum_{i=1}^{n}\alpha_i = 1$$

составим функцию Лагранжа [V]

$$L(\alpha_1, \ldots, \alpha_n; \lambda) = \sum_{i=1}^{n}\alpha_i^2 + \lambda\Big(\sum_{i=1}^{n}\alpha_i - 1\Big),$$

где λ — множитель Лагранжа. Необходимые условия существования условного экстремума имеют вид

$$\begin{cases} \dfrac{\partial L}{\partial \alpha_i} = 2\alpha_i + \lambda = 0, & i = \overline{1,n}, \\ \dfrac{\partial L}{\partial \lambda} = \displaystyle\sum_{i=1}^{n} \alpha_i - 1 = 0. \end{cases}$$

Решив эту систему, находим $\lambda = -2/n$ и $\alpha_i = 1/n$, $i = \overline{1,n}$, и убеждаемся в том, что при этих значениях аргументов функция $g(\alpha_1,\ldots,\alpha_n)$ имеет условный минимум. ▶

Замечание 2.3. Можно доказать состоятельность оценки $\overline{X}$ для математического ожидания (если оно существует), не предполагая существования конечной дисперсии $\mathbf{D}X$. #

Свойства *выборочной дисперсии*

$$\widehat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

отражены в следующей теореме.

Теорема 2.2. Если $\vec{X}_n$ — случайная выборка из генеральной совокупности X с конечной дисперсией σ^2, то выборочная дисперсия $\widehat{\sigma}^2(\vec{X}_n)$ — смещенная состоятельная оценка σ^2.

◀ Действительно,

$$\widehat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\Big((X_i - \mu) - (\overline{X} - \mu)\Big)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}\sum_{i=1}^{n}(\overline{X} - \mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)^2 + \frac{1}{n}n(\overline{X} - \mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2.$$

Используя свойства математического ожидания, получим

$$\mathrm{M}\,\widehat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\mathrm{M}\sum_{i=1}^{n}(X_i-\mu)^2 - \mathrm{M}(\overline{X}-\mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}(X_i-\mu)^2 - \mathrm{M}(\overline{X}-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{D}\,X_i - \mathrm{D}\,\overline{X} =$$

$$= \frac{1}{n}n\sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \neq \sigma^2,$$

т.е. $\widehat{\sigma}^2(\vec{X}_n)$ — смещенная оценка для дисперсии.

Докажем, что $\widehat{\sigma}^2(\vec{X}_n)$ является состоятельной оценкой. Доказательство проведем для случая, когда генеральная совокупность имеет моменты до четвертого порядка включительно и нулевое математическое ожидание. Последнее допущение не является принципиальным, так как дисперсия не зависит от значения ее математического ожидания (от точки отсчета). Применяя второе неравенство Чебышева, имеем

$$\mathbf{P}\left\{\left|\widehat{\sigma}^2(\vec{X}_n) - \frac{n-1}{n}\sigma^2\right| < \varepsilon\right\} \geqslant 1 - \frac{\mathrm{D}\,\widehat{\sigma}^2(\vec{X}_n)}{\varepsilon^2}.$$

Найдем дисперсию $\widehat{\sigma}^2(\vec{X}_n)$:

$$\widehat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X\overline{X} + \overline{X}^2) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}\frac{1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}n\overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2.$$

Воспользуемся известным равенством, согласно которому дисперсия скалярной случайной величины равна математическому ожиданию ее квадрата минус квадрат ее математического ожидания:

$$\mathrm{D}\,\widehat{\sigma}^2(\vec{X}_n) = \mathrm{M}\left(\widehat{\sigma}^2(\vec{X}_n)\right)^2 - \left(\mathrm{M}\,\widehat{\sigma}^2(\vec{X}_n)\right)^2.$$

Поскольку

$$\bigl(\widehat{\sigma}^2(\vec{X}_n)\bigr)^2 = \Bigl(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\Bigr)^2 = \\ = \frac{1}{n^2}\Bigl(\sum_{i=1}^{n}X_i^2\Bigr)^2 - 2\overline{X}^2\frac{1}{n}\sum_{i=1}^{n}X_i^2 + \overline{X}^4,$$

заключаем, что

$$\mathrm{M}\bigl(\widehat{\sigma}^2(\vec{X}_n)\bigr)^2 = \frac{1}{n^2}\mathrm{M}\Bigl(\sum_{i=1}^{n}X_i^2\Bigr)^2 - \frac{2}{n}\mathrm{M}\Bigl(\overline{X}^2\sum_{i=1}^{n}X_i^2\Bigr) + \mathrm{M}\,\overline{X}^4.$$

Получим выражения для математических ожиданий трех слагаемых, используя свойства математического ожидания и независимость случайных величин X_i, $i=\overline{1,n}$, каждая из которых имеет нулевое математическое ожидание и дисперсию σ^2. Для первого слагаемого имеем

$$\mathrm{M}\Bigl(\sum_{i=1}^{n}X_i^2\Bigr)^2 = \mathrm{M}\Bigl(\sum_{i=1}^{n}X_i^4 + \sum_{\substack{i,j=1\\ i\neq j}}^{n}X_i^2X_j^2\Bigr) = \sum_{i=1}^{n}\mathrm{M}\,X_i^4 + \\ + \sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathrm{M}\,X_i^2\,\mathrm{M}\,X_j^2 = \sum_{i=1}^{n}\mathring{m}_4 + \sum_{\substack{i,j=1\\ i\neq j}}^{n}\sigma^2\sigma^2 = n\,\mathring{m}_4 + n(n-1)\sigma^4.$$

Чтобы вычислить второе слагаемое, преобразуем его:

$$\mathrm{M}\Bigl(\overline{X}^2\sum_{i=1}^{n}X_i^2\Bigr) = \frac{1}{n^2}\mathrm{M}\Bigl(\Bigl(\sum_{j=1}^{n}X_j\Bigr)^2\sum_{i=1}^{n}X_i^2\Bigr) = \\ = \frac{1}{n^2}\mathrm{M}\Bigl(\Bigl(\sum_{j=1}^{n}X_j^2 + \sum_{\substack{j,k=1\\ j\neq k}}^{n}X_jX_k\Bigr)\sum_{i=1}^{n}X_i^2\Bigr) = \\ = \frac{1}{n^2}\mathrm{M}\Bigl(\sum_{j=1}^{n}X_j^2\sum_{i=1}^{n}X_i^2\Bigr) + \frac{1}{n^2}\Bigl(\sum_{i=1}^{n}X_i^2\sum_{\substack{j,k=1\\ j\neq k}}^{n}X_jX_k\Bigr).$$

Так как

$$\mathrm{M}\sum_{\substack{i=1\\ i\neq k}}^{n} X_i X_k = \sum_{\substack{i=1\\ i\neq k}}^{n} \mathrm{M}X_i \,\mathrm{M}X_k = 0, \quad k=\overline{1,n},$$

то

$$\mathrm{M}\sum_{i=1}^{n} X_i^2 \sum_{\substack{j,k=1\\ j\neq k}}^{n} X_j X_k = \sum_{\substack{i,j,k=1\\ i\neq j,\, i\neq k,\, j\neq k}}^{n} \mathrm{M}(X_i^2 X_j X_k) + 2\sum_{\substack{i,j=1\\ i\neq j}}^{n} \mathrm{M}(X_i^3 X_j) = 0.$$

Следовательно,

$$\mathrm{M}\Big(\overline{X}^2 \sum_{i=1}^{n} X_i^2\Big) = \frac{1}{n^2}\mathrm{M}\Big(\sum_{j=1}^{n} X_j^2 \sum_{i=1}^{n} X_i^2\Big) =$$

$$= \frac{1}{n^2}\mathrm{M}\Big(\sum_{i=1}^{n} X_i^4 + \sum_{\substack{i,j=1\\ i\neq j}}^{n} X_i^2 X_j^2\Big) = \frac{1}{n^2}\big(n\,\mathring{m}_4 + n(n-1)\sigma^4\big).$$

Аналогично можно показать, что

$$\mathrm{M}\overline{X}^4 = \frac{1}{n^4}\Big(\sum_{i=1}^{n} X_i\Big)^4 = \frac{\mathring{m}_4 + 3(n-1)\sigma^4}{n^3}.$$

В итоге получаем

$$\mathrm{M}\big(\widehat{\sigma}^2(\vec{X}_n)\big)^2 = \frac{n\,\mathring{m}_4 + n(n-1)\sigma^4}{n^2} - \frac{2(n\,\mathring{m}_4 + n(n-1)\sigma^4)}{n^3} +$$

$$+ \frac{\mathring{m}_4 - 3\sigma^4}{n^3} = \sigma^4 + \frac{\mathring{m}_4 - 3\sigma^4}{n} - \frac{2\,\mathring{m}_4 - 5\sigma^4}{n^2} + \frac{\mathring{m}_4 + 3(n-1)\sigma^4}{n^3}.$$

Поскольку $\mathrm{M}\widehat{\sigma}^2(\vec{X}_n) = \sigma^2 - \sigma^2/n$, окончательно находим

$$\mathrm{D}\widehat{\sigma}^2(\vec{X}_n) = \frac{\mathring{m}_4 - \sigma^4}{n} - \frac{2(\mathring{m}_4 - 2\sigma^4)}{n^2} + \frac{\mathring{m}_4 - 3\sigma^4}{n^3},$$

откуда с учетом второго неравенства Чебышева и следует состоятельность оценки $\widehat{\sigma}^2(\vec{X}_n)$ для дисперсии σ^2 генеральной совокупности X. ▶

Замечание 2.4. Из теоремы 2.2 следует, что статистика

$$S^2(\vec{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

является несмещенной и состоятельной оценкой дисперсии σ^2 генеральной совокупности. Ее называют ***исправленной выборочной дисперсией***.

Действительно,

$$S^2(\vec{X}_n) = \frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{n}{n-1}\sigma^2(\vec{X}_n).$$

Имеем

$$\mathrm{M}S^2(\vec{X}_n) = \frac{n}{n-1}\mathrm{M}\widehat{\sigma}^2(\vec{X}_n) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2$$

и

$$\mathrm{D}S^2(\vec{X}_n) = \frac{n^2}{(n-1)^2}\mathrm{D}\widehat{\sigma}^2(\vec{X}_n) \to 0$$

при $n \to \infty$, откуда и следует несмещенность и состоятельность $S^2(\vec{X}_n)$.

Отметим, что в дальнейшем ее выборочное значение будем обозначать S^2.

Замечание 2.5. Можно доказать, что *выборочные начальные* и *центральные моменты* являются состоятельными оценками соответствующих моментов генеральной совокупности, если только они существуют*. Однако эти оценки, кроме $\overline{X}$, являются смещенными.

*См.: *Крамер Г.*, а также: *Ивченко Г.И., Медведев Ю.И.*

Пример 2.1. Пусть n — число испытаний по схеме Бернулли с неизвестной вероятностью успеха θ. Рассмотрим случайную выборку $(X_1, \ldots, X_n)$, где X_i, $i = \overline{1,n}$, — случайная величина, которая с вероятностью θ принимает значение 1 („успех" в i-м испытании) и с вероятностью $1-\theta$ — значение 0 („неудача" в i-м испытании).

В качестве оценки θ возьмем относительную частоту успехов, т.е. $\widehat{\theta}(\vec{X}_n) = k(\vec{X}_n)/n$, где

$$k(\vec{X}_n) = \sum_{i=1}^{n} X_i$$

есть суммарное число успехов в n испытаниях. Эта оценка является несмещенной, так как

$$\mathrm{M}\widehat{\theta}(\vec{X}_n) = \frac{1}{n}\mathrm{M}(X_1 + \ldots + X_n) = \frac{1}{n}(\mathrm{M}X_1 + \ldots + \mathrm{M}X_n) = \frac{1}{n}n\theta = \theta,$$

и состоятельной, что непосредственно вытекает из закона больших чисел в форме Бернулли, согласно которому для любого $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{k(\vec{X}_n)}{n} - p\right| < \varepsilon\right\} = 1. \quad \#$$

В дальнейшем в соответствии с установившейся традицией статистику $k(\vec{X}_n)$, так же как и ее значение, часто будем обозначать просто символом k. В каждом конкретном случае должно быть ясно, о чем идет речь: о случайной величине или ее реализации.

Пример 2.2. Пусть $X_1, \ldots, X_n$ — случайная выборка из генеральной совокупности X, имеющей нормальное распределение с неизвестным средним значением θ и известной дисперсией σ^2.

Оценка $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n) = X_1$ является несмещенной для θ, ибо $\mathrm{M}X_1 = \mathrm{M}X = \theta$, но не является состоятельной, так как, во-первых, X_1 не зависит от объема выборки и, следовательно,

ее распределение не меняется с ростом n, а во-вторых,

$$\mathrm{P}\{|X_1-\theta|<\varepsilon\}=\frac{2}{\sigma\sqrt{2\pi}}\int\limits_0^{\varepsilon}e^{-\frac{t^2}{2\sigma^2}}\,dt\neq 1.$$

Пример 2.3. Имеем случайную выборку $\vec{X}_n$ из генеральной совокупности X с равномерным законом распределения

$$p(t;\theta)=\begin{cases}\frac{1}{b-a}, & t\in[a,b],\\ 0, & t\notin[a,b],\end{cases}$$

где $b-a=l$ — известная величина, $\theta=(a+b)/2$ — неизвестный параметр.

Возьмем в качестве оценки параметра θ среднее арифметическое *крайних членов вариационного ряда*

$$\theta^*(\vec{X}_n)=\frac{X_{(1)}+X_{(n)}}{2}.$$

Убедимся, что $\theta^*(\vec{X}_n)$ является несмещенной оценкой параметра θ и в классе всех несмещенных оценок $\overline{X}$ не является эффективной оценкой параметра θ для заданной параметрической модели.

Плотности распределения $X_{(1)}$ и $X_{(n)}$ на отрезке $[a,b]$ соответственно равны

$$p_{X_{(1)}}(x)=n\Big(1-\frac{x-a}{b-a}\Big)^{n-1}\frac{1}{b-a},\qquad p_{X_{(n)}}(x)=n\Big(\frac{x-a}{b-a}\Big)^{n-1}\frac{1}{b-a}$$

(см. пример 2.20). Вычислив

$$\mathrm{M}\,X_{(1)}=\int\limits_a^b xn\Big(\frac{b-x}{b-a}\Big)^{n-1}\frac{1}{b-a}\,dx=b-\frac{n}{n+1}(b-a),$$

$$\mathrm{M}\,X_{(n)}=\int\limits_a^b xn\left(\frac{x-a}{b-a}\right)^{n-1}\frac{1}{b-a}\,dx=a+\frac{n}{n+1}(b-a),$$

получим

$$\mathrm{M}\theta^*(\vec{X}_n) = \frac{1}{2}\big(\mathrm{M}\,X_{(1)} + \mathrm{M}\,X_{(n)}\big) = \frac{a+b}{2},$$

что и доказывает несмещенность оценки $\theta^*(\vec{x}_n)$.

Далее, используя совместную плотность распределения вероятностей случайных величин* $X_{(1)}$ и $X_{(n)}$

$$p_{X_{(1)}X_{(n)}}(x,y) = \frac{n(n-1)(y-x)^{n-2}}{(b-a)^n}, \quad a < x \leqslant y \leqslant b,$$

и равенство

$$\mathrm{D}\,\theta^*(\vec{X}_n) = \frac{1}{4}\mathrm{D}\big(X_{(1)} + X_{(n)}\big) = \\ = \frac{1}{4}\big(\mathrm{D}\,X_{(1)} + \mathrm{D}\,X_{(n)}\big) + \frac{1}{2}\,\mathrm{cov}\big(X_{(1)}, X_{(n)}\big),$$

можно получить

$$\mathrm{D}\,\theta^*(\vec{X}_n) = \frac{(b-a)^2}{2(n+1)(n+2)}.$$

Поскольку

$$\mathrm{D}\,\theta^*(\vec{X}_n) < \mathrm{D}(\overline{X}) = \frac{\sigma^2}{n} = \frac{(b-a)^2}{12n}, \quad n \geqslant 3,$$

то, следовательно, в классе всех несмещенных оценок $\overline{X}$ не является эффективной оценкой параметра θ для рассматриваемой параметрической модели.

Теорема 2.3 (о единственности эффективной оценки). Пусть $\widehat{\theta}(\vec{X}_n)$ и $\widetilde{\theta}(\vec{X}_n)$ — две эффективные оценки для параметра θ рассматриваемой параметрической модели. Тогда

$$\widehat{\theta}(\vec{X}_n) = \widetilde{\theta}(\vec{X}_n),$$

*См.: *Емельянов Г.В., Скитович В.П.*

где равенство следует понимать в вероятностном смысле:

$$\mathbf{P}\Big\{\vec{X}_n \in \{\vec{x}_n\colon \widehat{\theta}(\vec{x}_n) \neq \widetilde{\theta}(\vec{x}_n)\}\Big\} = 0.$$

◀ Действительно, рассмотрим статистику

$$\theta^*(\vec{X}_n) = \frac{1}{2}\big(\widehat{\theta}(\vec{X}_n) + \widetilde{\theta}(\vec{X}_n)\big).$$

По условию $\mathbf{D}\widehat{\theta}(\vec{X}_n) = \mathbf{D}\widetilde{\theta}(\vec{X}_n)$. Значит,

$$\mathbf{D}\theta^*(\vec{X}_n) = \frac{1}{4}\Big(\mathbf{D}\widehat{\theta}(\vec{X}_n) + \mathbf{D}\widetilde{\theta}(\vec{X}_n)\Big) + \frac{1}{2}\mathbf{cov}\big(\widehat{\theta}(\vec{X}_n), \widetilde{\theta}(\vec{X}_n)\big) =$$
$$= \frac{1}{2}\Big(\mathbf{D}\widehat{\theta}(\vec{X}_n) + \mathbf{cov}\big(\widehat{\theta}(\vec{X}_n), \widetilde{\theta}(\vec{X}_n)\big)\Big).$$

Поскольку

$$\Big|\mathbf{cov}\big(\widehat{\theta}(\vec{X}_n), \widetilde{\theta}(\vec{X}_n)\big)\Big| \leqslant \sqrt{\mathbf{D}\widehat{\theta}(\vec{X}_n)\,\mathbf{D}\widetilde{\theta}(\vec{X}_n)} = \mathbf{D}\widehat{\theta}(\vec{X}_n),$$

то

$$\mathbf{D}\theta^*(\vec{X}_n) = \frac{1}{2}\Big|\mathbf{D}\widehat{\theta}(\vec{X}_n) + \mathbf{cov}\big(\widehat{\theta}(\vec{X}_n), \widetilde{\theta}(\vec{X}_n)\big)\Big| \leqslant$$
$$\leqslant \frac{1}{2}\Big(\mathbf{D}\widehat{\theta}(\vec{X}_n) + \mathbf{cov}\big(\widehat{\theta}(\vec{X}_n), \widetilde{\theta}(\vec{X}_n)\big)\Big) \leqslant \mathbf{D}\widehat{\theta}(\vec{X}_n).$$

А так как $\widehat{\theta}(\vec{X}_n)$ — эффективная оценка, то

$$\mathbf{D}\theta^*(\vec{X}_n) = \mathbf{D}\widehat{\theta}(\vec{X}_n) = \mathbf{D}\widetilde{\theta}(\vec{X}_n)$$

и, как следствие,

$$\mathbf{cov}\big(\widehat{\theta}(\vec{X}_n), \widetilde{\theta}(\vec{X}_n)\big) = \mathbf{D}\widehat{\theta}(\vec{X}_n) = \mathbf{D}\widetilde{\theta}(\vec{X}_n) = \sqrt{\mathbf{D}\widehat{\theta}(\vec{X}_n)\,\mathbf{D}\widetilde{\theta}(\vec{X}_n)}.$$

Из последнего равенства следует [XVI], что

$$\widehat{\theta}(\vec{X}_n) = k\widetilde{\theta}(\vec{X}_n) + b.$$

Так как

$$\mathbf{D}\,\widehat{\theta}(\vec{X}_n) = \mathbf{cov}\big(k\widetilde{\theta}(\vec{X}_n) + b, \widetilde{\theta}(\vec{X}_n)\big) = k\,\mathbf{D}\,\widetilde{\theta}(\vec{X}_n) = k\,\mathbf{D}\,\widehat{\theta}(\vec{X}_n),$$

то получаем $k = 1$. Из условия несмещенности оценок следует, что $b = 0$:

$$\mathbf{M}\,\widetilde{\theta}(\vec{X}_n) = \mathbf{M}\,\widehat{\theta}(\vec{X}_n) = \mathbf{M}\big(\widetilde{\theta}(\vec{X}_n) + b\big) = \mathbf{M}\,\widetilde{\theta}(\vec{X}_n) + b.$$

Таким образом, $\widehat{\theta}(\vec{X}_n) = \widetilde{\theta}(\vec{X}_n)$. ▶

В дальнейшем изложении при рассмотрении параметрических моделей будем использовать дифференцирование по параметру под знаком *интеграла, зависящего от параметра.* Параметрические модели, для которых выполнены условия, обеспечивающие законность указанных операций, называют ***регулярными моделями.***

Теорема 2.4 (*неравенство Рао — Крамера).** Пусть рассматриваемая параметрическая модель является регулярной и $\widehat{\theta}(\vec{X}_n)$ — несмещенная оценка неизвестного параметра θ. Тогда имеет место неравенство

$$\mathbf{D}\,\widehat{\theta}(\vec{X}_n) \geqslant \frac{1}{nI(\theta)}, \tag{2.2}$$

где

$$I(\theta) = \mathbf{M}\left(\frac{\partial \ln p(X;\theta)}{\partial \theta}\right)^2.$$

Здесь $I(\theta)$ — ***количество информации по Фишеру***** в одном наблюдении, а $p(t;\theta)$ — плотность распределения генеральной совокупности X в случае *непрерывной статистической модели* и вероятность события $\{X = t\}$ в случае *дискретной статистической модели.*

*С.Р. Рао — индийский математик, К.Х. Крамер — шведский математик.

**Р.Э. Фишер (1890–1962) — английский статистик и генетик.

◀ Доказательство проведем для непрерывной модели. Пусть $p(t;\theta) > 0$ при $t \in A \subset \mathbb{R}$ и $p(t;\theta) = 0$ при $t \notin A$. Тогда плотность распределения

$$p_{\vec{X}_n}(T,\theta) \equiv p_{\vec{X}_n}(t_1,\ldots,t_n,\theta) = \prod_{i=1}^{n} p(t_i;\theta)$$

случайной выборки $\vec{X}_n$ отлична от нуля на множестве

$$B = A \times A \times \ldots \times A \subset \mathbb{R}^n,$$

где $T = (t_1, \ldots, t_n)$ — векторный аргумент. Поскольку

$$\int\limits_{\mathbb{R}^n} p_{\vec{X}_n}(T,\theta)\,dT = \int\limits_{B} p_{\vec{X}_n}(T,\theta)\,dT = 1,$$

имеем

$$\frac{\partial}{\partial\theta}\int\limits_{\mathbb{R}^n} p_{\vec{X}_n}(T,\theta)\,dT = \frac{\partial}{\partial\theta}\int\limits_{B} p_{\vec{X}_n}(T,\theta)\,dT = \int\limits_{B} \frac{\partial p_{\vec{X}_n}(T,\theta)}{\partial\theta}\,dT = 0,$$

или

$$\int\limits_{B} \frac{\partial \ln p_{\vec{X}_n}(T,\theta)}{\partial\theta}\, p_{\vec{X}_n}(T,\theta)\,dT = 0. \tag{2.3}$$

Так как $\widehat{\theta}(X_1,\ldots,X_n)$ — несмещенная оценка параметра θ, то

$$\mathrm{M}\widehat{\theta}(\vec{X}_n) = \int\limits_{\mathbb{R}^n} \widehat{\theta}(T)\, p_{\vec{X}_n}(T,\theta)\,dT = \int\limits_{B} \widehat{\theta}(T)\, p_{\vec{X}_n}(T,\theta)\,dT = \theta.$$

Таким образом,

$$\frac{\partial}{\partial\theta}\int\limits_{\mathbb{R}^n} \widehat{\theta}(T)\, p_{\vec{X}_n}(T,\theta)\,dT = \int\limits_{B} \widehat{\theta}(T)\, p_{\vec{X}_n}(T,\theta)\,dT = 1,$$

или, что то же самое,

$$\int\limits_B \widehat{\theta}(T)\frac{\partial \ln p_{\vec{X}_n}(T,\theta)}{\partial\theta}p_{\vec{X}_n}(T,\theta)\,dT = 1. \tag{2.4}$$

Умножив равенство (2.3) на параметр θ и вычтя его из равенства (2.4), приходим к равенству

$$\int\limits_B \bigl(\widehat{\theta}(T)-\theta\bigr)\frac{\partial \ln p_{\vec{X}_n}(T,\theta)}{\partial\theta}p_{\vec{X}_n}(T,\theta)\,dT = 1. \tag{2.5}$$

Согласно неравенству Коши — Буняковского, имеем

$$1 \leqslant \int\limits_B \bigl(\widehat{\theta}(T)-\theta\bigr)^2 p_{\vec{X}_n}(T,\theta)\,dT \int\limits_B \left(\frac{\partial \ln p_{\vec{X}_n}(T,\theta)}{\partial\theta}\right)^2 p_{\vec{X}_n}(T,\theta)\,dT =$$

$$= \mathbf{D}\,\widehat{\theta}(\vec{X}_n)\,\mathbf{M}\left(\frac{\partial \ln p_{\vec{X}_n}(\vec{X}_n,\theta)}{\partial\theta}\right)^2,$$

откуда и следует неравенство (2.2), так как

$$\mathbf{M}\left(\frac{\partial \ln p_{\vec{X}_n}(\vec{X}_n,\theta)}{\partial\theta}\right)^2 = \int\limits_B \left(\frac{\partial \ln p_{\vec{X}_n}(\vec{X}_n,\theta)}{\partial\theta}\right)^2 p_{\vec{X}_n}(T,\theta)\,dT =$$

$$= \int\limits_B \left(\sum_{i=1}^n \frac{\partial \ln p(t_i,\theta)}{\partial\theta}\right)^2 p_{\vec{X}_n}(T,\theta)\,dT =$$

$$= \sum_{i=1}^n \int\limits_B \left(\frac{\partial \ln p(t_i,\theta)}{\partial\theta}\right)^2 p_{\vec{X}_n}(T,\theta)\,dT = \sum_{i=1}^n \mathbf{M}\left(\frac{\partial \ln p(X_i,\theta)}{\partial\theta}\right)^2 =$$

$$= \sum_{i=1}^n \mathbf{M}\left(\frac{\partial \ln p(X,\theta)}{\partial\theta}\right)^2 = n\,\mathbf{M}\left(\frac{\partial \ln p(X,\theta)}{\partial\theta}\right)^2 = nI(\theta). \quad \blacktriangleright$$

Неравенство (2.2) определяет нижнюю границу дисперсий несмещенных оценок параметра θ для регулярных моделей.

Величину

$$e(\theta) = \frac{1}{nI(\theta)\,\mathrm{D}\,\widehat{\theta}(\vec{X}_n)}$$

называют ***показателем эффективности по Рао — Крамеру***. Из (2.2) следует, что для любой несмещенной оценки параметра θ величина $e(\theta)$ удовлетворяет условию $0 < e(\theta) \leqslant 1$.

Определение 2.4. Несмещенную ***оценку*** $\widehat{\theta}(\vec{X}_n)$ параметра $\theta \in \Theta \subset \mathbf{R}$ называют ***эффективной по Рао — Крамеру***, если показатель эффективности $e(\theta) = 1$.

Замечание 2.6. Равенство

$$\mathrm{D}\,\widehat{\theta}(\vec{X}_n) = \frac{1}{nI(\theta)}$$

имеет место тогда и только тогда, когда

$$\frac{\partial \ln p_{\vec{X}_n}(\vec{X}_n, \theta)}{\partial \theta} = a(\theta)\,(\widehat{\theta}(\vec{X}_n) - \theta),$$

что является необходимым и достаточным условием обращения неравенства Коши — Буняковского в равенство. Следовательно, это равенство является ***критерием эффективности для регулярных моделей***. При этом из равенства (2.5) следует, что $a(\theta) = 1/\mathrm{D}\,\widehat{\theta}(\vec{X}_n)$.

Замечание 2.7. Эффективная оценка по Рао — Крамеру для рассматриваемой регулярной модели является эффективной (см. определение 2.3). Утверждение следует из теоремы 2.3 о единственности эффективной оценки в классе несмещенных оценок. Обратное утверждение неверно, поскольку не любая параметрическая модель является регулярной (см. пример 2.21).

Пример 2.4. Рассмотрим *нормальную модель* $N(\theta, \sigma^2)$ в предположении, что дисперсия σ^2 известна. Оценка

$$\widehat{\theta}(\vec{X}_n) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

является несмещенной для неизвестного среднего значения $\theta = \mu$ (см. теорему 2.1). Убедимся в ее эффективности по Рао — Крамеру. Во-первых, в силу независимости элементов случайной выборки $\vec{X}_n = (X_1, \ldots, X_n)$ имеем

$$\mathbf{D}(\vec{X}_n) = \mathbf{D}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{D}\,X_i = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Во-вторых,

$$I(\theta) = \mathbf{M}\left(\frac{\partial}{\partial\theta}\ln\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(X-\theta)^2}{2\sigma^2}}\right)\right)^2 =$$

$$= \mathbf{M}\left(\frac{\partial}{\partial\theta}\left(\ln\frac{1}{\sqrt{2\pi}} - \ln\sigma - \frac{(X-\theta)^2}{2\sigma^2}\right)\right)^2 = \mathbf{M}\frac{(X-\theta)^2}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Следовательно,

$$e(\theta) = \frac{1}{nI(\theta)\,\mathbf{D}\overline{X}} = \frac{n}{n\frac{1}{\sigma^2}\cdot\sigma^2} = 1,$$

т.е. для нормальной модели $\overline{X}$ — эффективная оценка параметра μ.

Пример 2.5. Рассмотрим модель $N(\mu, \theta)$ в предположении, что среднее значение μ генеральной совокупности известно, а $\theta = \sigma^2$ — неизвестный параметр.

Покажем, что

$$\widetilde{S}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

является несмещенной и эффективной по Рао — Крамеру оценкой параметра σ^2. Действительно,

$$\mathbf{M}\widetilde{S}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}(X_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}X_i = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n}n\sigma^2 = \sigma^2,$$

т.е. $\widetilde{S}^2(\vec{X}_n)$ — несмещенная оценка.

Вычислим дисперсию $S^2(\vec{X}_n)$:

$$\mathbf{D}\widetilde{S}^2(\vec{X}_n) = \mathbf{M}\widetilde{S}^4(\vec{X}_n) - \left(\mathbf{M}\widetilde{S}^2(\vec{X}_n)\right)^2 =$$

$$= \mathbf{M}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right)^2 - \sigma^4 =$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbf{M}(X_i-\mu)^4 + \sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbf{M}\left((X_i-\mu)^2(X_j-\mu)^2\right)\right) - \sigma^4 =$$

$$= \frac{n\,\mathring{m}_4}{n^2} + \frac{n(n-1)}{n^2}\sigma^4 - \sigma^4 = \frac{3\sigma^4}{n} + \frac{n-1}{n}\sigma^4 - \sigma^4 = \frac{2\sigma^4}{n}.$$

Затем определим информацию по Фишеру:

$$I(\theta) = \mathbf{M}\left(\frac{\partial}{\partial\sigma^2}\left(\ln\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}\right)\right)^2 =$$

$$= \mathbf{M}\left(\frac{\partial}{\partial\sigma^2}\left(\ln\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\ln\sigma^2 - \frac{(X-\mu)^2}{2\sigma^2}\right)\right)^2 =$$

$$= \mathbf{M}\left(-\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4}\right)^2 = \frac{1}{4\sigma^4} - \frac{\mathbf{M}(X-\mu)^2}{2\sigma^6} + \frac{\mathbf{M}(X-\mu)^4}{4\sigma^8} =$$

$$= \frac{1}{4\sigma^2} - \frac{\sigma^2}{2\sigma^6} + \frac{3\sigma^4}{4\sigma^8} = \frac{1}{2\sigma^4},$$

поскольку для нормальной модели $\mathring{m}_4 = 3\sigma^4$ [XVI]. В результате получим

$$\mathbf{D}\widetilde{S}^2(\vec{X}_n) = \frac{1}{nI(\theta)} \quad \text{и} \quad \mathbf{e}(\theta) = 1,$$

т.е. $\widetilde{S}^2(\vec{X}_n)$ — эффективная оценка параметра θ для нормальной модели. #

Заметим, что

$$S^2(\vec{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

является несмещенной оценкой параметра σ^2 (см. замечание 2.4), но для нормальной модели $N(\mu,\theta)$ эта оценка не является эффективной. Это вытекает из теоремы 2.3 о единственности существования эффективной оценки. Можно показать, что

$$\mathbf{D}\widetilde{S}^2(\vec{X}_n)=\frac{2\sigma^4}{n-1}.$$

Следовательно,

$$e(\theta)=\frac{1}{nI(\theta)\,\mathbf{D}S^2(\vec{X}_n)}=\frac{n-1}{n}<1.$$

Пример 2.6. Рассмотрим экспоненциальную модель

$$p(x;\theta)=\begin{cases}\frac{1}{\theta}e^{-\frac{1}{\theta}x}, & x\geqslant 0;\\ 0, & x<0.\end{cases}$$

Покажем, что $\overline{X}$ является эффективной по Рао — Крамеру оценкой неизвестного параметра θ. Действительно,

$$\mathbf{D}(\overline{X})=\mathbf{D}\Big(\frac{1}{n}\sum_{i=1}^{n}X_i\Big)=\frac{1}{n}\mathbf{D}X=\frac{\theta^2}{n},$$

$$I(\theta)=\mathbf{M}\left(\frac{\partial}{\partial\theta}\left(\ln\frac{1}{\theta}e^{-\frac{X}{\theta}}\right)\right)^2=\mathbf{M}\Big(-\frac{1}{\theta}+\frac{X}{\theta^2}\Big)^2=$$

$$=\mathbf{M}\Big(\frac{X-\theta}{\theta^2}\Big)^2=\frac{\mathbf{M}(X-\theta)^2}{\theta^4}=\frac{\mathbf{D}X}{\theta^4}=\frac{\theta^2}{\theta^4}=\frac{1}{\theta^2},$$

откуда заключаем, что

$$e(\theta)=\frac{1}{n\frac{1}{\theta^2}\cdot\frac{\theta^2}{n}}=1.$$

2.2. Понятие достаточных статистик

Применение в реальных прикладных задачах методов математической статистики, как правило, связано с обработкой и хранением больших массивов *статистических данных*, относящихся к изучаемому объекту или процессу. Поэтому в этой области существует проблема сокращения объемов исходных данных без потери информации о *статистической модели*. Именно в связи с этой проблемой рассматривают так называемые *достаточные статистики*, к изучению которых мы приступаем.

Пусть $\vec{X}_n$ — *случайная выборка* из *генеральной совокупности* X с функцией распределения $F(x;\theta)$, где θ — неизвестный параметр.

Пусть, далее, $T = T(\vec{X}_n)$ — некоторая *статистика* (функция случайной выборки). Предположим, что нам известна не *выборка* $\vec{x}_n$, являющаяся *реализацией случайной выборки* $\vec{X}_n$, а только значение $T(\vec{x}_n) = t$ статистики T.

В дальнейших рассуждениях нас будет интересовать условная функция распределения

$$F_{\vec{X}_n}(z_1, \ldots, z_n \mid T(\vec{X}_n) = t)$$

случайной выборки X_1, ..., X_n при условии, что статистика $T(\vec{X}_n)$ приняла значение t. Заметим, что в общем случае это условное распределение зависит от параметра θ.

Определение 2.5. ***Статистику*** $T(\vec{X}_n)$ называют ***достаточной*** для параметра θ, если условная функция распределения $F_{\vec{X}_n}(z_1, \ldots, z_n \mid T(\vec{X}_n) = t)$ случайной выборки $\vec{X}_n$ при условии $T(\vec{X}_n) = t$ не зависит от параметра θ при любом возможном значении t.

Согласно определению 2.5, при фиксированном значении t изменение параметра θ не влияет на условный закон распределения случайной выборки $\vec{X}_n$ при условии $T(\vec{X}_n) = t$. Это

означает, что значение t статистики $T(\vec{X}_n)$ дает полную информацию о параметре θ.

Замечание 2.8. Поскольку для *непрерывной статистической модели*

$$F_{\vec{X}_n}\left(z_1,\ldots,z_n \mid T(\vec{X}_n)=t\right)=$$

$$=\int_{-\infty}^{z_1}\ldots\int_{-\infty}^{z_n} p_{\vec{X}_n}\left(u_1,\ldots,u_n \mid T(\vec{X}_n)=t\right)du_1\ldots du_n,$$

а для дискретной

$$F_{\vec{X}_n}\left(z_1,\ldots,z_n \mid T(\vec{X}_n)=t\right)=$$

$$=\sum_{i=1}^{n}\sum_{j:\,x_i^j<z_i}\mathbf{P}\left\{X_{11}=x_1^j,\ldots,X_{1n}=x_n^j \mid T(\vec{X}_n)=t\right\},$$

то в случае достаточной статистики $T(\vec{X}_n)$ соответственно условная плотность распределения $p_{\vec{X}_n}\left(z_1,\ldots,z_n \mid T(\vec{X}_n)=t\right)$ и условная вероятность $\mathbf{P}\left\{X_1=x_1^j,\ldots,X_n=x_n^j \mid T(\vec{X}_n)=t\right\}$ не зависят от θ.

Пример 2.7. Пусть X_i, $i=\overline{1,n}$, — число успехов в i-м испытании по схеме Бернулли. Рассмотрим статистику

$$T(\vec{X}_n)=X_1+\ldots+X_n,$$

имеющую смысл числа успехов в n испытаниях по схеме Бернулли. Покажем, что она является достаточной для параметра θ — вероятности успеха в одном испытании.

Найдем условное распределение вероятностей, которое для случая дискретной модели будем записывать в виде

$$\mathbf{P}\left\{X_1=x_1,\ldots,X_n=x_n \mid T(\vec{X}_n)=t\right\}.$$

Согласно определению условной вероятности, имеем

$$\mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n \mid T(\vec{X}_n) = t\} = \\ = \frac{\mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n, T(\vec{X}_n) = t\}}{\mathbf{P}\{T(\vec{X}_n) = t\}}.$$

Если $x_1 + \ldots + x_n = t$, то

$$\mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n, T(\vec{X}_n) = t\} = \\ = \mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n\} = \theta^t(1-\theta)^{n-t}.$$

Напомним, что случайные величины X_i, $i = \overline{1, n}$, могут принимать здесь только значения 1 или 0, причем $X_1 + \ldots + X_n = t$. Поскольку вероятность $\mathbf{P}\{T(\vec{X}_n) = t\}$ определяется формулой Бернулли

$$\mathbf{P}\{T(\vec{X}_n) = t\} = C_n^t \theta^t (1-\theta)^{n-t},$$

то условную вероятность можно переписать в виде

$$\mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n \mid T(\vec{X}_n) = t\} = \frac{\theta^t(1-\theta)^{n-t}}{C_n^t \theta^t (1-\theta)^{n-t}} = \frac{1}{C_n^t},$$

т.е. она не зависит от θ. Если же $x_1 + \ldots + x_n \neq t$, то

$$\mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n, T(\vec{X}_n) = t\} = 0,$$

а следовательно, и

$$\mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n \mid T(\vec{X}_n) = t\} = 0,$$

т.е. опять-таки условная вероятность не зависит от θ, а значит, согласно определению 2.5, $T(\vec{X}_n) = X_1 + \ldots + X_n$ — достаточная статистика для параметра θ. #

Проверять достаточность конкретных статистик, основываясь на определении 2.5, довольно сложно. Следующая теорема

дает критерий достаточности статистики, который помогает выполнять такую проверку.

Предварительно введем функцию

$$L(X_1,\ldots,X_n;\theta)=p(x_1;\theta)\ldots p(x_n;\theta), \tag{2.6}$$

которую называют ***функцией правдоподобия***. Здесь $p(x;\theta)$ обозначает плотность распределения непрерывной случайной величины или вероятность события $\{X=x\}$ в случае дискретной случайной величины, а X_i, $i=\overline{1,n}$, — *элементы случайной выборки* $\vec{X}_n$.

Теорема 2.5 (***критерий факторизации Неймана — Пирсона****). Статистика $T=T(x_1,\ldots,x_n)$ является достаточной для параметра θ тогда и только тогда, когда для любой реализации $(x_1,\ \ldots,\ x_n)$ случайной выборки $(X_1,\ \ldots,\ X_n)$ *выборочное значение* функции правдоподобия имеет вид

$$L(x_1,\ldots,x_n;\theta)=g(T(x_1,\ldots,x_n),\theta)\,h(x_1,\ldots,x_n), \tag{2.7}$$

т.е. может быть представлено в виде произведения двух сомножителей, из которых второй не зависит от θ, а первый (зависящий от θ) зависит от результатов наблюдений x_1, ..., x_n только через статистику $T=T(x_1,\ldots,x_n)$.

◀ Приведем доказательство для дискретной модели и учтем, что в рассматриваемом случае вероятность

$$p(x_i;\theta)=\mathbf{P}_\theta\{X_i=x_i\},\quad i=\overline{1,n},$$

зависит от θ. Поэтому будем использовать следующую форму записи:

$$\mathbf{P}_\theta\{\vec{X}_n=\vec{x}_n\}=\prod_{i=1}^{n}\mathbf{P}_\theta\{X_i=x_i\}=\prod_{i=1}^{n}p(x_i;\theta),$$

*Е. Нейман (1894–1981) — американский математик и статистик; Э. Пирсон (1857–1936) — английский математик, биолог и философ.

что в соответствии с (2.6) приводит к равенству

$$L(\vec{x}_n;\theta) = \mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n\}.$$

Если статистика $T = T(\vec{X}_n)$ достаточна, то при любом фиксированном значении t из области возможных значений условное распределение выборки

$$\mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n \mid T(\vec{X}_n) = t\}$$

не зависит от θ и, следовательно, его можно записать в виде $h(\vec{x}_n, t)$ или $h(\vec{x}_n)$, так как t — фиксированная величина.

Пусть $T(\vec{X}_n) = t$. Тогда для любой реализации $\vec{x}_n$ случайной выборки, удовлетворяющей условию $T(\vec{x}_n) = t$, событие $\{\vec{X}_n = \vec{x}_n\}$ включено в событие $\{T(\vec{X}_n) = t\}$, т.е. $\{\vec{X}_n = \vec{x}_n\} \subset \subset \{T(\vec{X}_n) = t\}$ и, следовательно,

$$L(\vec{x}_n;\theta) = \mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n\} = \mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n, T(\vec{X}_n) = t\} =$$
$$= \mathbf{P}_\theta\{T(\vec{X}_n) = t\}\,\mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n \mid T(\vec{X}_n) = t\} = g(t,\theta)\,h(\vec{x}_n),$$

т.е. имеет место равенство (2.7).

Наоборот, пусть имеет место представление (2.7). Тогда при любом $\vec{x}_n$, для которого $T(\vec{x}_n) = t$, с учетом равенств

$$\mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n, T(\vec{X}_n) = t\} = \mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n\} = L(\vec{x}_n;\theta)$$

имеем

$$\mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n \mid T(\vec{X}_n) = t\} = \frac{\mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n, T(\vec{X}_n) = t\}}{\mathbf{P}_\theta\{T(\vec{X}_n) = t\}} =$$
$$= \frac{L(\vec{x}_n;\theta)}{\sum\limits_{T(\vec{x}_n)=t} L(\vec{x}_n;\theta)} = \frac{g(t,\theta)\,h(\vec{x}_n)}{\sum\limits_{T(\vec{x}_n)=t} g(t,\theta)\,h(\vec{x}_n)} = \frac{h(\vec{x}_n)}{\sum\limits_{T(\vec{x}_n)=t} h(\vec{x}_n)},$$

т.е. условное распределение выборки не зависит от θ. Если же $\vec{x}_n$ таково, что $T(\vec{x}_n) \neq t$, то очевидно, что

$$\mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n \mid T(\vec{X}_n) = t\} = 0.$$

Таким образом, в любом случае условная вероятность

$$\mathbf{P}_\theta\{\vec{X}_n = \vec{x}_n \mid T(\vec{X}_n) = t\}$$

не зависит от θ, а это и означает достаточность статистики $T(\vec{X}_n)$ согласно определению (2.5). ▶

Заметим, что всякая *эффективная по Рао — Крамеру оценка* $\widehat{\theta} = \widehat{\theta}(\vec{X}_n)$ параметра θ является достаточной статистикой. Это следует из равенства

$$a(\theta)\big(\widehat{\theta}(\vec{X}_n) - \theta\big) = \frac{\partial\big(\ln p_{\vec{X}_n}(\vec{X}_n, \theta)\big)}{\partial\theta}$$

(*критерия эффективности для регулярных моделей*, см. замечание 2.6) и соотношения (2.7). Обратное утверждение неверно (см. пример 2.27).

Приведем без доказательства следующие утверждения.

1°. Если существует *эффективная оценка* параметра, то она является функцией от достаточной статистики.

Из этого утверждения следует, что эффективную оценку следует искать среди функций от достаточных статистик.

2°. Если $T(\vec{X}_n)$ — достаточная статистика для параметра θ, то таковой же является и любая взаимно однозначная функция от $T(\vec{X}_n)$.

Нахождение эффективных оценок с помощью достаточных статистик связано с понятием полноты достаточной статистики, которое здесь мы не будем рассматривать, а отсылаем заинтересованного читателя к специальной литературе*.

Замечание 2.9. Определение 2.5 достаточной статистики можно обобщить на случай вектора параметров $\vec{\theta} = (\theta_1, \ldots, \theta_r)$. Векторную статистику

$$\vec{T} = (T_1, \ldots, T_r) = (T_1(X_1, \ldots, X_n), \ldots, T_r(X_1, \ldots, X_n))$$

*См.: *Ивченко Г.И., Медведев Ю.И.*

будем называть достаточной для вектора параметров $\vec{\theta}$, если условное распределение выборки $\vec{X}_n = (X_1, \ldots, X_n)$ при условии $\vec{T}(\vec{X}_n) = \vec{t}$, где $\vec{t} = (t_1, \ldots, t_r)$ — некоторое фиксированное значение, не зависит от параметра $\vec{\theta}$. При этом критерий факторизации — теорема 2.5 — обобщается на случай векторной статистики. #

Как уже отмечалось выше, достаточные статистики позволяют сократить объем исходных данных, сохраняя всю содержащуюся в этих данных информацию.

Кроме того, один из наиболее универсальных методов нахождения оценок для неизвестных параметров — *метод максимального правдоподобия*, который приводит к *оценкам* параметров через достаточные статистики.

Приведем примеры, поясняющие смысл и свойства достаточных статистик.

Пример 2.8. Пусть $(x_1, \ldots, x_n)$ — реализация случайной выборки $(X_1, \ldots, X_n)$, и случайная величина X имеет экспоненциальное распределение, т.е.

$$p(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, & x \geqslant 0; \\ 0, & x < 0. \end{cases}$$

В этом случае функция правдоподобия имеет вид

$$L(X_1,\ldots,X_n;\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{X_i}{\theta}} = \frac{1}{\theta^n} \exp\Big(-\frac{1}{\theta}\sum_{i=1}^{n} X_i\Big),$$

откуда в соответствии с критерием факторизации Неймана — Пирсона следует, что статистика

$$T = \sum_{i=1}^{n} X_i$$

является достаточной. Здесь роль множителя $g\big(T(x_1,\ldots,x_n),\theta\big)$ играет все выражение для $L\big(x_1,\ldots,x_n;\theta\big)$, а $h(x_1,\ldots,x_n)=1$.

В данном случае существует эффективная оценка параметра θ, выражающаяся через достаточную статистику, а именно: оценка

$$\widehat{\theta}=\frac{1}{n}T=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X},$$

как было показано в примере 2.6, является эффективной по Рао — Крамеру.

Пример 2.9 (общая нормальная модель). Пусть в эксперименте наблюдается случайная величина $X\sim N(\theta_1,\theta_2^2)$ с неизвестными параметрами θ_1, θ_2. Так как плотность распределения X имеет вид

$$p\big(x;\theta_1,\theta_2\big)=\frac{1}{\theta_2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Big(\frac{x-\theta_1}{\theta_2}\Big)^2\right),\quad \theta_1\in\mathbb{R},\quad \theta_2\in(0,\infty),$$

то значение функции правдоподобия для выборки x_1, ..., x_n из генеральной совокупности X в данном случае имеет вид

$$L\big(x_1,\ldots,x_n;\theta_1,\theta_2\big)=\prod_{i=1}^{n}p\big(x;\theta_1,\theta_2\big)=$$

$$=\frac{1}{\big(\theta_2\sqrt{2\pi}\big)^n}\exp\left(-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\overline{x})^2-\frac{n(\overline{x}-\theta_1)}{2\theta_2^2}\right),$$

откуда в силу критерия факторизации Неймана — Пирсона (множитель $h(x_1,\ldots,x_n)=1$) заключаем, что двумерная статистика $T=(T_1,T_2)$, где

$$T_1=\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\qquad T_2=\sum_{i=1}^{n}(X_i-\overline{X})^2,$$

является достаточной для вектора параметров (θ_1,θ_2).

Пример 2.10. Пусть дана случайная выборка $(X_1, \ldots, X_n)$ из генеральной совокупности $X \sim R(0,\theta)$, т.е. X имеет равномерное распределение на интервале $(0,\theta)$, где θ — неизвестный параметр.

Покажем, что *крайний член вариационного ряда* $X_{(n)}$ случайной выборки является достаточной статистикой для параметра θ, т.е. $T(X_1,\ldots,X_n) = X_{(n)}$ — достаточная статистика.

Действительно, так как плотность равномерного распределения имеет вид

$$p(x;\theta) = \begin{cases} \frac{1}{\theta}, & x \in [0,\theta]; \\ 0, & x \notin [0,\theta], \end{cases}$$

то выборочное значение функции правдоподобия имеет вид

$$L(x_1,\ldots,x_n;\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n, & x_i \in [0,\theta],\ i = \overline{1,n}; \\ 0 & \text{в противном случае.} \end{cases}$$

Мы видим, что область изменения каждого аргумента x_i функции $L(x_1,\ldots,x_n;\theta)$ зависит от параметра θ. Рассмотрим статистику

$$T(X_1,\ldots,X_n) = X_{(n)}$$

и положим

$$g(t,\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n, & x_i \in [0,\theta],\ i = \overline{1,n}; \\ 0 & \text{в противном случае.} \end{cases}$$

$$h(x_1,\ldots,x_n) = \begin{cases} 1, & x_i > 0,\ i = \overline{1,n}; \\ 0 & \text{в противном случае.} \end{cases}$$

Тогда выборочное значение функции правдоподобия для выборки $\vec{x}_n$ можно представить в виде

$$L(x_1,\ldots,x_n;\theta) = g(T,\theta)\,h(x_1,\ldots,x_n).$$

Заметим, что при определении функции $h(x_1,\ldots,x_n)$ на x_i не наложены ограничения, поскольку

$$x_i \leqslant T(\vec{x}_n) = x_{(n)} \leqslant \theta, \quad i = \overline{1,n}.$$

Это значит, что функция $h(x_1,\ldots,x_n)$ не зависит от параметра θ. Согласно критерию факторизации, статистика $T(\vec{x}_n) = X_{(n)}$ является достаточной для параметра θ.

Пример 2.11 (модель Коши). Пусть имеется случайная выборка $\vec{X}_n$ из генеральной совокупности $X \sim K(\theta)$, т.е. X имеет *распределение Коши*:

$$p(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}, \quad x \in \mathbb{R}.$$

Функция правдоподобия в рассматриваемом случае имеет вид

$$L(X_1,\ldots,X_n;\theta) = \prod_{i=1}^{n} p(X_i;\theta) = \frac{1}{\pi^n} \prod_{i=1}^{n} \frac{1}{1+(X_i-\theta)^2}.$$

Из этого равенства следует, что существует лишь одна статистика $T(X_1,\ldots,X_n)$, которая для выборочного значения функции правдоподобия дает представление (2.7), а именно: тривиальная статистика $(T_1(\vec{X}_n),\ldots,T_n(\vec{X}_n)) = (X_1, \ldots, X_n)$, совпадающая с самой случайной выборкой. #

Отметим, что из определений эффективности по Рао — Крамеру и достаточных статистик вытекает, что существование эффективных оценок по Рао — Крамеру или достаточных статистик можно ожидать для специальных классов параметрических моделей. Если существование таких оценок установлено, то их можно найти с помощью метода максимального правдоподобия, который изложен в следующем параграфе.

2.3. Методы получения точечных оценок

Рассмотрим методы определения *точечных оценок* параметров $\theta_1, \ldots, \theta_r$, от которых зависит распределение $p(x;\theta_1,\ldots,\theta_r)$ *генеральной совокупности* X.

В математической статистике разработано большое число методов оценивания неизвестных параметров по данным *случайной выборки*, из которых в приложениях наиболее часто используются:

- *метод моментов*;
- *метод максимального правдоподобия*;
- *графический метод* (или *метод номограмм*);
- *метод наименьших квадратов*.

Рассмотрим первые три из них (последний рассмотрен ниже, см. **7**).

Метод моментов. ***Метод моментов*** был предложен английским статистиком К. Пирсоном и является одним из первых общих методов оценивания. Он состоит в следующем.

Пусть имеется случайная выборка $\vec{X}_n = (X_1, \ldots, X_n)$ из генеральной совокупности X, распределение которой $p(x;\vec{\theta})$ известно с точностью до вектора параметров $\vec{\theta} = (\theta_1, \ldots, \theta_r)$. Требуется найти *оценку* параметра $\vec{\theta}$ по случайной выборке $\vec{X}_n$.

Будем предполагать, что у случайной величины X существуют первые r моментов: $m_k = \mathrm{M}X^k$, $k = \overline{1,r}$. Ясно, что величины m_k являются функциями неизвестного вектора параметров $\vec{\theta}$, т.е. $m = m_k(\vec{\theta})$.

Рассмотрим *выборочные моменты* $\widehat{\mu}_k(\vec{X}_n)$ (или же $\widehat{\nu}_k(\vec{X}_n)$, см. **1.3**).

Выборочные моменты являются *состоятельными оценками* соответствующих моментов генеральной совокупности X (см. замечание 2.5), поэтому при большом *объеме выборки* m_k и $\mathring{m}_k$, $k = \overline{1,r}$, можно заменить соответственно моментами $\widehat{\mu}_k$ и $\widehat{\nu}_k$ выборки $\vec{x}_n$.

В методе моментов в качестве точечной оценки $\widehat{\vec{\theta}}(\vec{X}_n) = \left(\widehat{\theta}_1(\vec{X}_n), \ldots, \widehat{\theta}_r(\vec{X}_n)\right)$ вектора параметров $\vec{\theta}$ берут *статистику*, значение которой для любой реализации $\vec{x}_n$ случайной выборки $\vec{X}_n$ получают как решение системы уравнений

$$\widehat{\mu}_k = \mu_k(\vec{\theta}), \quad k = \overline{1, r}. \tag{2.8}$$

Можно показать*, что при условии непрерывной зависимости решения этой системы от $\widehat{\mu}_k$, $k = \overline{1, r}$, *оценка*, полученная методом моментов, является состоятельной и имеет ***асимптотически нормальное распределение***, т.е. ее распределение при $n \to \infty$ стремится к нормальному. При этом уравнения (2.8) во многих случаях просты и их решение не вызывает больших вычислительных сложностей.

Понятно, что метод моментов не применим, когда моменты генеральной совокупности нужного порядка не существуют (например, для *распределения Коши*, у которого не существует даже начальный момент первого порядка — математическое ожидание [XVI]).

Пример 2.12. Пусть случайная величина X имеет *гамма-распределение* с плотностью

$$f(x, \lambda, \alpha) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x}, & x \geqslant 0; \\ 0, & x < 0, \end{cases}$$

где λ и α — два неизвестных параметра.

Заметим, что этому распределению подчиняется время X до отказа системы из $\alpha = m$ (m — натуральное число) однотипных элементов, если каждый из $m - 1$ элементов включается в работу после отказа предыдущего, и время до отказа X_i, $i = \overline{1, m}$,

*См.: *Ивченко Г.И., Медведев Ю.И.*

любого элемента имеет экспоненциальное распределение

$$p(x;\lambda)=\begin{cases}\dfrac{1}{\lambda}e^{-\frac{x}{\lambda}}, & x\geqslant 0;\\ 0, & x<0.\end{cases}$$

Найдем с помощью метода моментов оценки неизвестных параметров λ и α.

В данном случае, используя определение *гамма-функции*

$$\Gamma(\alpha)=\int\limits_0^\infty t^{\alpha-1}e^{-t}dt,$$

а также рекуррентное соотношение $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$, получим следующие выражения для первого, второго начальных моментов и дисперсии:

$$m_1=\int\limits_0^\infty\frac{\lambda^\alpha x^\alpha}{\Gamma(\alpha)}e^{-\lambda x}dx=\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\lambda}=\frac{\alpha}{\lambda},$$

$$m_2=\int\limits_0^\infty\frac{\lambda^\alpha x^{\alpha+1}}{\Gamma(\alpha)}e^{-\lambda x}dx=\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\lambda^2}=\frac{\alpha(\alpha+1)}{\lambda^2},$$

$$\mathbf{D}X=\mathbf{M}(X^2)-(\mathbf{M}X)^2=m_2-m_1^2=\frac{\alpha}{\lambda^2}.$$

Пусть $\vec{x}_n$ — *выборка* объема n из генеральной совокупности X. Находим моменты выборки $\widehat{\mu}_1=\overline{x}$ и $\widehat{\nu}_2=\widehat{\sigma}^2$. Приравнивая моменты $m_1=\mathbf{M}X$ и $\mathring{m}_2=\mathbf{D}X$ к соответствующим моментам выборки, получаем систему уравнений

$$\begin{cases}\dfrac{\alpha}{\lambda}=\overline{x},\\ \dfrac{\alpha}{\lambda^2}=\widehat{\sigma}^2,\end{cases}$$

откуда находим значения оценок

$$\widehat{\lambda}=\frac{\overline{x}}{\widehat{\sigma}^2},\qquad \widehat{a}=\left(\frac{\overline{x}}{\widehat{\sigma}}\right)^2.$$

Следовательно, оценками неизвестных параметров будут статистики

$$\widehat{\lambda}(\vec{X}_n)=\frac{\overline{X}}{\widehat{\sigma}^2(\vec{X}_n)},\qquad \widehat{a}(\vec{X}_n)=\left(\frac{\overline{X}}{\widehat{\sigma}(\vec{X}_n)}\right)^2.$$

Пример 2.13. Методом моментов найдем оценку параметра $\theta=p$ в *биномиальной модели*, где p есть вероятность „успеха“ в любом из n независимых повторных наблюдений, а случайная величина $k(\vec{X}_n)$ — число „успехов“. Случайной выборкой $\vec{X}_n$ в данном случае являются n дискретных случайных величин X_i, каждая из которых принимает значение 1 с вероятностью p и 0 с вероятностью $1-p$. При этом $k(\vec{X}_n)=X_1+\ldots+X_n$, а математическое ожидание $\mathrm{M}k(\vec{X}_n)=np$ [XVI].

Если в результате n независимых наблюдений мы получили *выборочное значение* $k(\vec{X}_n)=k$, то уравнение, которое нужно составить согласно методу моментов, имеет вид

$$np=k.$$

Получаем $\widehat{p}=k/n$. Следовательно, точечной оценкой параметра p является *относительная частота*.

Метод максимального правдоподобия. Одним из наиболее универсальных методов оценивания параметров является ***метод максимального правдоподобия*** (предложенный Р. Фишером), суть которого состоит в следующем.

Рассмотрим *функцию правдоподобия* случайной выборки $\vec{X}_n$ из генеральной совокупности X, распределение $p(x;\vec{\theta})$ которой известно с точностью до параметра $\vec{\theta}\in\Theta$:

$$L(X_1,\ldots,X_n;\vec{\theta})=\prod_{i=1}^{n}p(X_i;\theta).$$

По определению, ***оценкой максимального правдоподобия*** параметра θ называют статистику $\widehat{\vec{\theta}}(\vec{X}_n)$, значения $\widehat{\vec{\theta}}$ которой для любой выборки $\vec{x}_n$ удовлетворяют условию

$$L(\vec{x}_n;\widehat{\vec{\theta}}) = \max_{\vec{\theta}\in\Theta} L(\vec{x}_n;\vec{\theta}), \tag{2.9}$$

т.е. для выборки функция правдоподобия, как функция аргумента $\vec{\theta}$, достигает максимума.

Если функция $L(\vec{x}_n;\vec{\theta})$ дифференцируема как функция аргумента $\vec{\theta}$ при любом значении $\vec{x}_n$ из множества $\mathcal{X}_n$ значений случайной выборки $\vec{X}_n$ и максимум $L(\vec{x}_n;\vec{\theta})$ достигается во внутренней точке из Θ, то значение точечной оценки максимального правдоподобия в случае скалярного параметра удовлетворяет уравнению (необходимому условию экстремума [II])

$$\frac{\partial L(\vec{x}_n;\theta)}{\partial\theta} = 0, \quad \text{или} \quad \frac{\partial \ln L(\vec{x}_n;\theta)}{\partial\theta} = 0, \tag{2.10}$$

так как при логарифмировании точки экстремума остаются теми же, а уравнение, как правило, упрощается.

Если распределение случайной величины X зависит от вектора параметров $\vec{\theta} = (\theta_1, \ldots, \theta_r)$, то второе из уравнений (2.10) заменяется системой уравнений

$$\frac{\partial \ln L(\vec{x}_n;\theta)}{\partial\theta_k} = 0, \quad k = \overline{1,r}. \tag{2.11}$$

Уравнения (2.10) и (2.11) называют ***уравнениями правдоподобия***. Для наиболее важных семейств распределений $p(x;\vec{\theta})$ уравнение правдоподобия имеет единственное решение $\widehat{\vec{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_r)$. Во многих случаях решение системы (2.11), являющейся, как правило, нелинейной, приходится искать численными методами*.

*См.: *Ивченко Г.И., Медведев Ю.И.*

Пример 2.14. Применим метод максимального правдоподобия для оценки параметра $\theta = p$ в биномиальной модели, где p имеет смысл вероятности „успеха“ в любом из n независимых повторных испытаний (испытаний по схеме Бернулли), в которых было зафиксировано k „успехов“.

В рассматриваемом случае значения функции правдоподобия $L(k;p)$ есть вероятность появления k „успехов“ в серии из n испытаний. Эта вероятность, как известно, определяется по формуле Бернулли, т.е.

$$L(k;p) = C_n^k p^k (1-p)^{n-k}.$$

Находя

$$\ln L(k;p) = \ln C_n^k + k \ln p + (n-k)\ln(1-p),$$

получаем уравнение правдоподобия (2.10) в виде

$$\frac{\partial \ln L(k;p)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0,$$

откуда получаем $\widehat{p} = k/n$. Нетрудно убедиться в том, что $\widehat{p}$ есть точка максимума $L(k;p)$. Следовательно, оценка максимального правдоподобия вероятности p совпадает с относительной частотой „успеха“ в n испытаниях.

Пример 2.15. Пусть наблюдаемая в эксперименте случайная величина X — время работы прибора до отказа — имеет экспоненциальное распределение с плотностью

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0; \\ 0, & x < 0, \end{cases}$$

где λ — неизвестный параметр.

Применяя метод максимального правдоподобия, найдем точечную оценку для параметра λ. Пусть $\vec{x}_n = (x_1, \ldots, x_n)$ —

любая *реализация случайной выборки* $\vec{X}_n$ из генеральной совокупности X.

В рассматриваемом случае

$$L(x_1,\ldots,x_n;\lambda)=\prod_{i=1}^{n}\lambda e^{-\lambda x_i}=\lambda^n\exp\Big(-\lambda\sum_{i=1}^{n}x_i\Big),$$

$$\ln L(x_1,\ldots,x_n;\lambda)=n\ln\lambda-\lambda\sum_{i=1}^{n}x_i.$$

Следовательно, уравнение правдоподобия (2.10) имеет вид

$$\frac{\partial\ln L(\vec{x};\lambda)}{\partial\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i=0,$$

откуда следует, что

$$\widehat{\lambda}=\Big(\frac{1}{n}\sum_{i=1}^{n}x_i\Big)^{-1}.$$

Итак, точечной оценкой неизвестного параметра λ является $\widehat{\lambda}(\vec{X}_n)=1/\overline{X}$.

Если учесть, что $\mathbf{M}X=1/\lambda$, а наилучшей оценкой $\mathbf{M}X=\mu$ является *выборочное среднее*

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

то полученный ответ представляется вполне естественным.

Пример 2.16. Для общей *нормальной модели* $N(\theta_1,\theta_2^2)$ методом максимального правдоподобия найдем оценку вектора параметров $\vec{\theta}=(\theta_1,\theta_2)$.

В этом случае функция правдоподобия

$$L(\vec{X}_n;\theta_1,\theta_2)=\frac{1}{(\theta_2\sqrt{2\pi})^n}\exp\Big(-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(X_i-\theta_1)^2\Big)$$

и, как следствие,

$$\ln L(\vec{x}_n;\theta_1,\theta_2) = -n\ln\sqrt{2\pi} - n\ln\theta_2 - \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1)^2.$$

Поскольку число неизвестных параметров $r=2$, система уравнений правдоподобия (2.11) будет состоять из двух уравнений:

$$\begin{cases} \dfrac{\partial}{\partial\theta_1}\ln L = \dfrac{1}{\theta_2^2}\displaystyle\sum_{i=1}^{n}(x_i-\theta_1) = 0, \\ \dfrac{\partial}{\partial\theta_2}\ln L = -\dfrac{n}{\theta_2} + \dfrac{1}{\theta_2^3}\displaystyle\sum_{i=1}^{n}(x_i-\theta_1) = 0. \end{cases}$$

Решая систему, получаем

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n}x_i, \qquad \widehat{\theta}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Следовательно, оценками максимального правдоподобия для математического ожидания $\mathrm{M}X=\theta_1$ и дисперсии $\mathrm{D}X=\theta_2^2$ случайной величины, распределенной по нормальному закону, являются соответственно выборочное среднее

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$$

и *выборочная дисперсия*

$$\widehat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2. \quad \#$$

Оценки максимального правдоподобия могут быть *смещенными* (см. примеры 2.16, 2.27) и не являться *эффективными* (см. пример 2.27). Однако, как показывают примеры, часто

смещенность можно устранить. Кроме того, во многих случаях для несмещенной и не являющейся *эффективной по Рао — Крамеру оценки* $\widehat{\theta}(\vec{X}_n)$ параметра θ выполняется условие

$$\lim_{n\to\infty} e(\theta) = 1$$

(см. пример 2.27). В этом случае ***оценку*** $\widehat{\theta}(\vec{X}_n)$ параметра θ называют ***асимптотически эффективной.***

Приведем без доказательства основные свойства оценок максимального правдоподобия для *регулярных моделей.*

1. Если для скалярного параметра θ существует эффективная оценка, то уравнение правдоподобия (2.10) имеет единственное решение, которое является выборочным значением этой оценки.

2. Если существует *достаточная статистика* параметра θ, то решения уравнения правдоподобия являются функциями от выборочного значения этой статистики.

Следовательно, если, кроме того, существует эффективная по Рао — Крамеру оценка $\widehat{\theta}(\vec{X}_n)$, то единственное решение уравнения правдоподобия является функцией от выборочного значения достаточной статистики.

3. Если параметрическая модель $\{F(x;\theta), \theta \in \Theta\}$ удовлетворяет некоторым общим условиям*, то уравнение правдоподобия имеет решение $\widehat{\theta}$, которое является выборочным значением состоятельной оценки $\widehat{\theta}(\vec{X}_n)$ параметра θ. Оценка $\widehat{\theta}(\vec{X}_n)$ является асимптотически эффективной и имеет асимптотически нормальное распределение $N\left(\theta, 1/\sqrt{nI(\theta)}\right)$.

Графический метод (метод номограмм). ***Графический метод*** позволяет не только достаточно просто найти значения оценок неизвестных параметров распределения вероятностей $F(x;\theta_1,\theta_2)$ наблюдаемой в эксперименте случайной величины, но и сделать предварительное заключение о правильности выбора вида распределения. Окончательное заключение

*См.: *Ивченко Г.И., Медведев Ю.И.*

о правильности такого выбора проводят с помощью так называемых *критериев согласия*, которые рассмотрены подробно в 5.

Идея графического метода состоит в следующем. С помощью некоторого нелинейного преобразования $u = u(y)$ семейство уравнений $y = F(x;\theta_1,\theta_2)$ приводится к виду $u = ax + b$.

По выборке $\vec{x}_n = (x_1, \ldots, x_n)$ из генеральной совокупности X строится *эмпирическая функция распределения* $F_n(x)$, являющаяся, как известно, статистическим аналогом для *теоретической функции распределения* $F(x;\theta_1,\theta_2)$.

Если в результате преобразования $u = u(y)$, которое применяется к функции $y = F_n(x)$, точки $(x_i,\ u(F_n(x_i)))$ будут достаточно „тесно" концентрироваться около некоторой прямой, то можно говорить о правильности выбора семейства распределений $F(x;\theta_1,\theta_2)$. В этом случае остается найти приближенные значения $\widetilde{\theta}_1$ и $\widetilde{\theta}_2$ параметров θ_1 и θ_2.

Для реализации идеи графического метода строят ***вероятностную бумагу*** — бумагу, разграфленную (специальным образом) так, чтобы график функции $F(x;\theta_1,\theta_2)$ изображался на ней прямой линией. С этой целью на оси ординат отмечают не значения переменной u, а соответствующие им значения y. Тем самым равноотстоящим точкам на оси ординат соответствуют значения y, связанные с u нелинейной зависимостью $u = u(y)$.

Из пояснений к вероятностной бумаге всегда ясно, как связаны параметры a и b с параметрами θ_1 и θ_2 рассматриваемого семейства.

Проиллюстрируем сказанное на примере нормального закона распределения $\Phi\left(\frac{x-\mu}{\sigma}\right)$, где μ и σ — неизвестные параметры. Графики функций этого семейства при $\mu = 2$ и $\sigma = 1/2, 1, 2$ изображены на рис. 2.1.

Рассмотрим преобразование $u = \Phi^{-1}(y)$, в результате которого получим $u = \frac{x-\mu}{\sigma}$, или $u = ax + b$, где

$$a = \frac{1}{\sigma}, \qquad b = -\frac{\mu}{\sigma}. \tag{2.12}$$

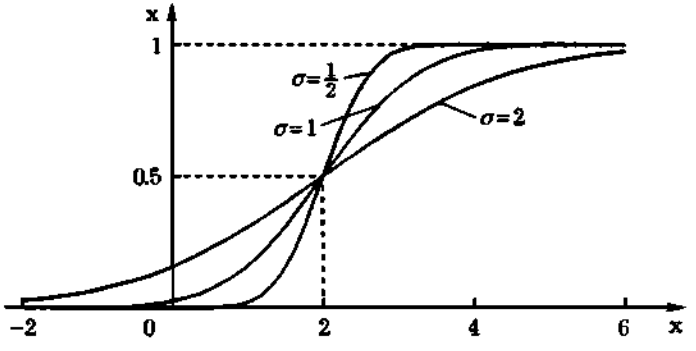


Рис. 2.1

По выборке $(x_1, \ldots, x_n)$ из генеральной совокупности X построим эмпирическую функцию распределения $F_n(x)$. Если точки $(x_i,\ u(F_n(x_i)))$ достаточно „тесно" концентрируются около некоторой прямой, то предположение о нормальном законе распределения принимаем. Затем на глаз проводим (на вероятностной бумаге) прямую линию $u = ax + b$, проходящую как можно ближе ко всем точкам $(x_i,\ u(F_n(x_i)))$, и определяем приближенные значения a и b.

Используя равенства (2.12), находим приближенные значения неизвестных параметров:

$$\widetilde{\sigma} = \frac{1}{a}, \qquad \widetilde{\mu} = -\frac{b}{a}.$$

Пример 2.17. Для определения предела прочности стекловолокна, изготовленного по новой технологии, проведены испытания на разрыв $n = 17$ образцов. Получены следующие значения предела прочности X (в мегапаскалях): $x_1 = 181$, $x_2 = 194$, $x_3 = 173$, $x_4 = 153$, $x_5 = 168$, $x_6 = 176$, $x_7 = 163$, $x_8 = 152$, $x_9 = 155$, $x_{10} = 156$, $x_{11} = 178$, $x_{12} = 160$, $x_{13} = 164$, $x_{14} = 169$, $x_{15} = 155$, $x_{16} = 122$, $x_{17} = 144$.

Предел прочности образцов, изготовленных по старой технологии, хорошо согласовывался с нормальным законом распределения. Требуется проверить согласие результатов экспе-

римента с нормальным законом распределения и оценить его параметры.

Для решения поставленной задачи воспользуемся графическим методом. Перейдем от выборки к вариационному ряду $x_{(1)}, x_{(2)}, \ldots, x_{(17)}$ и нанесем значения x_i, $i=\overline{1,17}$, на ось Ox. Далее с помощью таблицы квантилей нормального распределения (см. табл. П.2) находим значения функции $\Phi^{-1}(y_i)$, обратной к функции $\Phi(x)$, при $y_i=\frac{2i-1}{2n}$, $i=\overline{1,17}$:

$$\Phi^{-1}(17/34)=\Phi^{-1}(1/2)=0,$$
$$\Phi^{-1}(19/34)=-\Phi^{-1}(15/34)=0{,}1479,$$
$$\Phi^{-1}(21/34)=-\Phi^{-1}(13/34)=0{,}2993,$$
$$\Phi^{-1}(23/34)=-\Phi^{-1}(11/34)=0{,}4578,$$
$$\Phi^{-1}(25/34)=-\Phi^{-1}(9/34)=0{,}6289,$$
$$\Phi^{-1}(27/34)=-\Phi^{-1}(7/34)=0{,}8208,$$
$$\Phi^{-1}(29/34)=-\Phi^{-1}(5/34)=1{,}0494,$$
$$\Phi^{-1}(31/34)=-\Phi^{-1}(3/34)=1{,}3517,$$
$$\Phi^{-1}(33/34)=-\Phi^{-1}(1/34)=1{,}8895.$$

На рис. 2.2 приведены значения $F_n(x)$ в плоскости переменных x и $u=\Phi^{-1}(y)$.

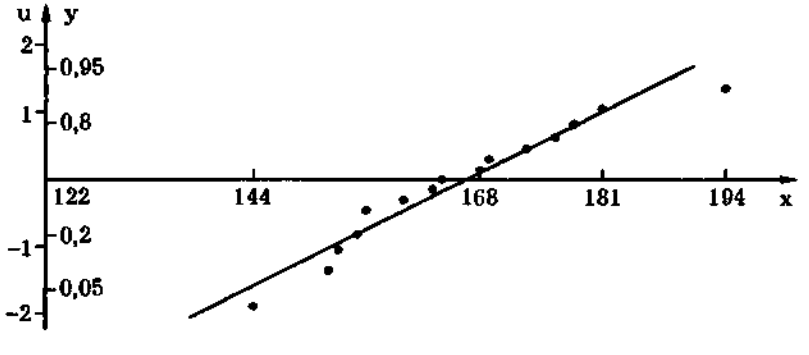


Рис. 2.2

На рис. 2.2 видно, что точки графика функции $F_n(x)$ расположены достаточно близко от прямой $u = ax + b$ при

$$a = \operatorname{tg}\alpha \approx 0{,}58, \qquad b = -d/a \approx -62{,}7,$$

где d — расстояние от 0 до точки пересечения прямой с осью Ox. Следовательно, оценки параметров μ и σ нормального распределения $F(x;\mu,\sigma) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ равны

$$\widetilde{\sigma} = 1/a = 1{,}75, \qquad \widetilde{\mu} = -b/a = d = 110.$$

Для сравнения приведем оценки параметров μ и σ, полученные методом максимального правдоподобия: $\widehat{\mu} = 162{,}5$, $\widehat{\sigma} =$, т.е. оценки весьма близки.

2.4. Решение типовых примеров

Пример 2.18. В результате пяти измерений длины стержня одним прибором (без систематических ошибок) получены следующие данные: 92; 94; 103; 105; 106. Найдем *выборочное значение несмещенной оценки* $S^2(\vec{X}_5)$ дисперсии ошибок прибора.

Выборочное значение несмещенной оценки вычисляется по формуле

$$S^2(\vec{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

где n — *объем выборки*. В данном случае среднее значение $\overline{x}$ выборки равно

$$\overline{x} = \frac{92 + 94 + 103 + 105 + 106}{5} = 100.$$

Используя это значение, находим

$$S^2(\vec{X}_5) = \frac{1}{4}\Big((92-100)^2 + (94-100)^2 + (103-100)^2 + \\ + (105-100)^2 + (106-100)^2\Big) = \\ = \frac{(-8)^2 + (-6)^2 + 3^2 + 5^2 + 6^2}{4} = 42{,}5.$$

Пример 2.19. Убедимся в том, что *выборочная функция распределения* является несмещенной оценкой для функции распределения $F(x)$ *генеральной совокупности* X в точке $x \in \mathbb{R}$. По определению выборочная функция распределения имеет вид

$$\widehat{F}(x; \vec{X}_n) = \frac{n(x, \vec{X}_n)}{n},$$

где $n(x, \vec{X}_n)$ — число *элементов случайной выборки*, меньших x.

Используя функцию Хевисайда [XI]

$$h(x) = \begin{cases} 1, & x \geqslant 0; \\ 0, & x < 0, \end{cases}$$

получим

$$\widehat{F}(x; \vec{X}_n) = \frac{n(x, \vec{X}_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} h(x - X_i).$$

Так как

$$\mathrm{M}\, h(x - X_i) = \int\limits_{-\infty}^{\infty} h(x-t)\, p(t)\, dt = \int\limits_{-\infty}^{x} p(t)\, dt = F(x),$$

где $p(x)$ — плотность распределения генеральной совокупности X, *оценка* $\widehat{F}(x; \vec{X}_n)$ является несмещенной:

$$\mathrm{M}\, \widehat{F}(x; \vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{M}\, h(x - X_i) = \frac{1}{n} n F(x) = F(x).$$

Пример 2.20. Рассмотрим *случайную выборку* $(X_1, \ldots, X_5)$ объема $n = 5$ из генеральной совокупности X, распределенной по *показательному закону*

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0; \\ 0, & x < 0. \end{cases}$$

В качестве *точечной оценки* математического ожидания возьмем среднее арифметическое *крайних членов вариационного ряда*

$$\widehat{\theta}(\vec{X}_5) = \frac{X_{(1)} + X_{(5)}}{2}.$$

Покажем, что *оценка* является *смещенной*.

Из свойств математического ожидания получаем

$$\mathrm{M}\widehat{\theta}(\vec{X}_5) = \frac{\mathrm{M}X_{(1)} + \mathrm{M}X_{(5)}}{2},$$

откуда заключаем, что для вычисления математического ожидания точечной оценки необходимо знать законы распределения случайных величин $X_{(1)}$ и $X_{(5)}$. Для первой из них имеем

$$\begin{aligned} F_{X_{(1)}}(x) &= \mathbf{P}\{X_{(1)} < x\} = \\ &= 1 - \mathbf{P}\{X_{(1)} \geqslant x\} = 1 - \mathbf{P}\{X_1 \geqslant x, \ldots, X_5 \geqslant x\} = \\ &= 1 - \mathbf{P}\{X_1 \geqslant x\} \ldots \mathbf{P}\{X_5 \geqslant x\} = 1 - \big(1 - F(x)\big)^5, \end{aligned}$$

где $F(x)$ — функция распределения генеральной совокупности. Аналогичны вычисления для случайной величины $X_{(5)}$:

$$F_{X_{(5)}}(x) = \mathbf{P}\{X_{(5)} < x\} = \mathbf{P}\{X_1 < x, \ldots, X_5 < x\} = (F(x))^n.$$

Учитывая вид функции распределения генеральной совокупности (вид показательного закона распределения), заключаем, что

$$F_{X_{(1)}}(x) = \begin{cases} 1 - e^{-5\lambda x}, & x \geqslant 0; \\ 0, & x < 0, \end{cases} \qquad F_{X_{(5)}}(x) = \begin{cases} \big(1 - e^{-\lambda x}\big)^5, & x \geqslant 0; \\ 0, & x < 0. \end{cases}$$

Следовательно,

$$\mathrm{M}X_{(1)} = \int\limits_0^\infty x p_{X_{(1)}}(x)\,dx = 5\lambda \int\limits_0^\infty x e^{-5\lambda x}\,dx = \frac{1}{5\lambda},$$

$$\mathrm{M}X_{(5)} = \int\limits_0^\infty x p_{X_{(5)}}(x)\,dx = 5\lambda \int\limits_0^\infty x\left(1-e^{-5\lambda x}\right)^4 e^{-\lambda x}\,dx = \frac{287}{60\lambda}.$$

Из найденных формул окончательно находим

$$\mathrm{M}\widehat{\theta}(\vec{X}_5) = \frac{1}{2}\left(\frac{1}{5\lambda} + \frac{287}{60\lambda}\right) = \frac{299}{120\lambda}.$$

Сравнивая найденное значение с математическим ожиданием для рассматриваемой генеральной совокупности X, убеждаемся, что $\mathrm{M}\widehat{\theta}(\vec{X}_5) \neq \mathrm{M}X$, т.е. точечная оценка $\widehat{\theta}(\vec{X}_5)$ смещенная.

Пример 2.21. Пусть задана случайная выборка объема n из генеральной совокупности X с плотностью распределения

$$p(x;\alpha) = \begin{cases} e^{\alpha - x}, & x \geqslant \alpha; \\ 0, & x < \alpha. \end{cases}$$

В качестве точечной оценки неизвестного параметра α возьмем $\widehat{\theta}(\vec{X}_n) = X_{(1)}$. Убедимся, что эта оценка смещенная. Найдем несмещенную оценку и покажем, что обе *оценки* являются *состоятельными.*

Предварительно найдем плотность распределения случайной величины $\widehat{\theta}(\vec{X}_n)$. Если $F(x)$ — функция распределения генеральной совокупности X, то функция распределения случайной величины $X_{(1)}$ имеет вид $F_{X_{(1)}}(x) = 1 - (1 - F(x))^n$ (см. пример 2.20). Поскольку

$$F(x) = \int\limits_\alpha^x e^{\alpha - x}\,dx = 1 - e^{\alpha - x}$$

при $x \geqslant \alpha$, то

$$F_{X_{(1)}}(x) = \begin{cases} 1 - e^{n(\alpha - x)}, & x \geqslant \alpha; \\ 0, & x < \alpha. \end{cases}$$

Исходя из функции распределения, можем найти плотность распределения оценки $\widehat{\theta}(\vec{X}_n)$

$$p_{X_{(1)}}(x) = \begin{cases} ne^{n(\alpha - x)}, & x \geqslant \alpha; \\ 0, & x < \alpha. \end{cases}$$

Таким образом,

$$\mathrm{M}\widehat{\theta}(\vec{X}_n) = \mathrm{M}X_{(1)} = \int\limits_{\alpha}^{\infty} xne^{n(\alpha - x)}\,dx = \alpha + \frac{1}{n},$$

т.е. оценка $\widehat{\theta}(\vec{X}_n) = X_{(1)}$ — смещенная. Из последнего равенства легко увидеть, как нужно модифицировать оценку $\widehat{\theta}(\vec{X}_n)$, чтобы получить несмещенную оценку. Несмещенной оценкой параметра α является $\theta^*(\vec{X}_n) = \widehat{\theta}(\vec{X}_n) - 1/n$.

Покажем, что оценка $\widehat{\theta}^*(\vec{X}_n)$ является состоятельной. Для этого применим второе неравенство Чебышева. Так как

$$\mathrm{M}X_{(1)}^2 = \int\limits_{\alpha}^{\infty} x^2 ne^{n(\alpha - x)}\,dx = \alpha^2 + \frac{2\alpha}{n} + \frac{2}{n^2}$$

и

$$\mathrm{D}X_{(1)} = \mathrm{M}X_{(1)}^2 - \left(\mathrm{M}X_{(1)}\right)^2 = \frac{1}{n^2},$$

то

$$\mathrm{P}\left\{\left|X_{(1)} - \frac{1}{n} - \alpha\right| < \varepsilon\right\} \geqslant 1 - \frac{1}{n^2}.$$

Отсюда следует, что обе оценки состоятельные.

Пример 2.22. Пусть дана случайная выборка $(X_1, \ldots, X_n)$ из генеральной совокупности X с плотностью распределения

$$p(x;\theta) = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\theta)^2}{2\sigma^2}}, & x \geqslant 0; \\ 0, & x < 0, \end{cases}$$

где σ — известный параметр. Покажем, что *эффективной по Рао — Крамеру оценкой* неизвестного параметра θ является

$$\widehat{\theta}(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \ln X_i.$$

Прежде всего убедимся, что $\widehat{\theta}(\vec{X}_n)$ является несмещенной оценкой параметра θ. Имеем

$$\mathrm{M}\widehat{\theta}(\vec{X}_n) = \mathrm{M}\frac{1}{n}\sum_{i=1}^{n} \ln X_i = \frac{1}{n}\sum_{i=1}^{n} \mathrm{M}\ln X_i = \frac{1}{n} n \mathrm{M}\ln X = \mathrm{M}\ln X.$$

Поскольку

$$\mathrm{M}\ln X = \int_0^{\infty} \ln x \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\theta)^2}{2\sigma^2}}\, dx,$$

то, введя новое переменное $t = \dfrac{\ln x - \theta}{\sigma}$, получим

$$\mathrm{M}\ln X = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma t + \theta) e^{-\frac{t^2}{2}}\, dt =$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}}\, dt + \frac{\theta}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\, dt = \theta.$$

Чтобы убедиться в том, что *показатель эффективности по Рао — Крамеру* $e(\theta)$ равен единице, найдем дисперсию $\mathrm{D}\widehat{\theta}(\vec{X}_n)$

и *количество информации по Фишеру* $I(\theta)$. Для дисперсии имеем

$$\mathbf{D}\widehat{\theta}(\vec{X}_n) = \mathbf{D}\frac{1}{n}\sum_{i=1}^{n}\ln X_i = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{D}\ln X_i =$$

$$= \frac{1}{n^2} n\,\mathbf{D}\ln X = \frac{1}{n}\int_0^{\infty}(\ln x-\theta)^2\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\theta)^2}{2\sigma^2}}\,dx.$$

С помощью замены переменного $t = \dfrac{\ln x-\theta}{\sigma}$ приходим к следующему результату:

$$\mathbf{D}\widehat{\theta}(\vec{X}_n) = \frac{\sigma^2}{n}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}}\,dt = \frac{\sigma^2}{n}.$$

Для $I(\theta)$ имеем

$$I(\theta) = \mathbf{M}\Big(\frac{\partial \ln p(X;\theta)}{\partial\theta}\Big)^2 =$$

$$= \mathbf{M}\left(\frac{\partial}{\partial\theta}\left(\ln\frac{1}{X\sigma\sqrt{2\pi}} - \frac{(\ln X-\theta)^2}{2\sigma^2}\right)\right)^2 = \frac{1}{\sigma^4}\mathbf{M}(\ln X-\theta)^2 =$$

$$= \frac{1}{\sigma^4}\int_0^{\infty}(\ln x-\theta)^2\cdot\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\theta)^2}{2\sigma^2}}\,dx = \frac{1}{\sigma^4}\sigma^2 = \frac{1}{\sigma^2}.$$

В результате окончательно находим

$$e(\theta) = \frac{1}{nI(\theta)\,\mathbf{D}\widehat{\theta}(\vec{X}_n)} = \frac{1}{n\cdot\frac{\sigma^2}{n}\cdot\frac{1}{\sigma^2}} = 1.$$

Пример 2.23. В условиях примера 2.21 найдем нижнюю границу $\dfrac{1}{nI(\theta)}$ *неравенства Рао — Крамера* и объясним, почему дисперсия $\mathbf{D}\theta^*(\vec{X}_n) = \mathbf{D}\,X_{(1)}$ несмещенной оценки $\theta^*(\vec{X}_n)$ параметра $\theta=\alpha$ меньше величины $\dfrac{1}{nI(\theta)}$.

Имеем

$$I(\alpha)=\int\limits_{\alpha}^{\infty}\Big(\frac{\partial \ln e^{\alpha-x}}{\partial\alpha}\Big)^2 e^{\alpha-x}\,dx=\int\limits_{\alpha}^{\infty}e^{\alpha-x}\,dx=1.$$

Поэтому

$$\frac{1}{nI(\alpha)}=\frac{1}{n}$$

и, следовательно,

$$\mathbf{D}\theta^*(\vec{X}_n)=\frac{1}{n^2}<\frac{1}{n},\quad n>1.$$

Последнее неравенство объясняется тем, что рассматриваемая *параметрическая модель* не является *регулярной*. Действительно, дифференцируя интеграл

$$\int\limits_{-\infty}^{\infty}p(x;\alpha)\,dx=1$$

по параметру α, получаем

$$\frac{d}{d\alpha}\int\limits_{-\infty}^{\infty}p(x;\alpha)\,dx=0.$$

Однако

$$\int\limits_{-\infty}^{\infty}\frac{\partial p(x;\alpha)}{\partial\alpha}\,dx=\int\limits_{\alpha}^{\infty}e^{\alpha-x}\,dx=1.$$

В таких случаях часто можно найти несмещенные оценки, дисперсия которых меньше чем $\frac{1}{nI(\theta)}$. #

Заметим также, что параметрическая модель, рассмотренная в примере 2.3, не является регулярной. При этом, как мы

видим, дисперсия оценок примеров 2.3 и 2.23 является бесконечно малой при $n \to \infty$ более высокого порядка, чем $1/n$. Такие *оценки* называют *сверхэффективными*.

Пример 2.24. Пусть случайная величина X имеет распределение Пуассона

$$\mathbf{P}\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots,$$

где λ — неизвестный параметр. В результате независимых наблюдений получена случайная выборка $(X_1, \ldots, X_n)$. Найдем *методом моментов* точечную оценку $\widehat{\theta}(\vec{X}_n)$ параметра λ и убедимся, что эта оценка является несмещенной и состоятельной.

Так как оценивается один параметр, то для получения оценки нужно составить одно уравнение. Известно [XVI], что $\mathbf{M}X = \lambda$. Следовательно, в качестве точечной оценки параметра λ распределения Пуассона можно взять выборочное среднее, т.е. $\widehat{\theta}(\vec{X}_n) = \overline{X}$. Из теоремы 2.1 следует, что эта оценка является несмещенной и состоятельной.

Пример 2.25. Пусть дана случайная выборка $(X_1, \ldots, X_n)$ объема n из генеральной совокупности X, имеющей равномерный закон распределения

$$p(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b); \\ 0, & x \notin (a, b), \end{cases}$$

с неизвестными параметрами a и b. Найдем методом моментов точечные оценки этих параметров.

Известно [XVI], что для равномерно распределенной случайной величины X

$$\mathbf{M}X = \frac{a+b}{2}, \qquad \mathbf{D}X = \frac{(b-a)^2}{12}.$$

Выборочное среднее $\overline{X}$ и выборочная дисперсия $\widehat{\sigma}^2(\vec{X}_n)$ вычисляются по формулам

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i, \qquad \widehat{\sigma}^2(\vec{X}_n)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

Составляем систему двух уравнений

$$\begin{cases}\dfrac{a+b}{2}=\overline{x},\\ \dfrac{(b-a)^2}{12}=\widehat{\sigma}^2(\vec{x}_n).\end{cases}$$

Решая систему, получаем

$$\widehat{b}=\overline{x}+\sqrt{3}\widehat{\sigma}(\vec{x}_n), \qquad \widehat{a}=\overline{x}-\sqrt{3}\widehat{\sigma}(\vec{x}_n).$$

Окончательно имеем

$$\widehat{b}(\vec{X}_n)=\overline{X}+\sqrt{3}\widehat{\sigma}(\vec{X}_n), \qquad \widehat{a}(\vec{X}_n)=\overline{X}-\sqrt{3}\widehat{\sigma}(\vec{X}_n).$$

Пример 2.26. Пусть дана случайная выборка $(X_1, \ldots, X_n)$ объема n из генеральной совокупности X, распределенной по биномиальному закону

$$P_k(j)=C_k^j\theta^j(1-\theta)^{k-j}, \quad j=\overline{0,k},$$

с неизвестным параметром θ (вероятностью появления события в одном испытании). *Методом максимального правдоподобия* найдем точечную оценку параметра θ.

Функция правдоподобия в этом случае имеет вид

$$L(X_1,\ldots,X_n;\theta)=\prod_{i=1}^{n}p(X_i;\theta),$$

где

$$p(x_i;\theta)=P_k(x_i)=C_k^{x_i}\theta^{x_i}(1-\theta)^{k-x_i}.$$

Отсюда находим

$$\ln L(x_1,\ldots,x_n;\theta) = \ln\left(C_k^{x_1}C_k^{x_2}\ldots C_k^{x_n}\right) + \\ + \sum_{i=1}^{n} x_i \ln\theta + \left(kn - \sum_{i=1}^{n} x_i\right)\ln(1-\theta).$$

Следовательно, *уравнение правдоподобия*

$$\frac{\partial \ln L(\vec{x}_n;\theta)}{\partial\theta} = 0$$

в данном случае сводится к следующему:

$$\frac{1}{\theta}\sum_{i=1}^{n} x_i - \left(kn - \sum_{i=1}^{n} x_i\right)\frac{1}{1-\theta} = 0.$$

Решив уравнение правдоподобия, найдем критическую точку функции правдоподобия

$$\widehat{\theta} = \frac{1}{kn}\sum_{i=1}^{n} x_i.$$

Покажем, что эта точка является точкой максимума выборочного значения функции правдоподобия. Для этого найдем вторую производную по θ:

$$\frac{\partial^2 \ln L(\vec{x}_n;\theta)}{\partial\theta^2} = -\frac{1}{\theta^2}\sum_{i=1}^{n} x_i + \frac{1}{(1-\theta)^2}\left(kn - \sum_{i=1}^{n} x_i\right).$$

Легко убедиться, что

$$\frac{\partial^2 \ln L(\vec{x}_n;\theta)}{\partial\theta^2} < 0, \quad \theta = \widehat{\theta}.$$

Поэтому $\widehat{\theta}$ — точка максимума выборочного значения функции правдоподобия, определяющая *оценку максимального правдоподобия*

$$\widehat{p}(\vec{X}_n) = \frac{1}{kn}\sum_{i=1}^{n} X_i.$$

Пример 2.27. Методом максимального правдоподобия по случайной выборке $(X_1, \ldots, X_n)$ найдем оценку параметра θ ***распределения Парето*** $(\theta > 0, \alpha > 0)$:

$$p(x;\theta) = \begin{cases} \dfrac{\alpha}{\theta}\left(\dfrac{\theta}{x}\right)^{\alpha+1}, & x \geqslant \theta; \\ 0, & x < \theta. \end{cases}$$

Для выборочного значения функции максимального правдоподобия при выполнении условий $\theta \leqslant x_i$, $i = \overline{1,n}$, находим

$$L(x_1,\ldots,x_n;\theta) = \frac{\alpha}{\theta}\left(\frac{\theta}{x_1}\right)^{\alpha+1}\frac{\alpha}{\theta}\left(\frac{\theta}{x_2}\right)^{\alpha+1}\cdots\frac{\alpha}{\theta}\left(\frac{\theta}{x_n}\right)^{\alpha+1} = \\ = \frac{\alpha^n}{x_1\ldots x_n}\left(\frac{\theta}{x_1}\right)^{\alpha}\left(\frac{\theta}{x_2}\right)^{\alpha}\cdots\left(\frac{\theta}{x_n}\right)^{\alpha}.$$

Ясно, что $L(x_1,\ldots,x_n;\theta) = 0$, если $\theta > x_i$ для какого-либо значения индекса i. Из вида функции правдоподобия $L(x_1,\ldots,x_n;\theta)$ заключаем, что она является возрастающей функцией θ при $\theta < X_{(1)}$ и равна нулю при $\theta > X_{(1)}$. Следовательно, $\widehat{\theta}(\vec{X}_n) = X_{(n)}$ — оценка максимального правдоподобия параметра θ.

Пример 2.28. Рассмотрим *параметрическую модель*

$$p(x;\lambda) = \begin{cases} \dfrac{\theta^\lambda}{\Gamma(\lambda)}x^{\lambda-1}e^{-\theta x}, & x > 0; \\ 0, & x \leqslant 0, \end{cases}$$

где $p(x;\lambda)$ — плотность распределения генеральной совокупности X, $\lambda > 0$ — неизвестный параметр, $\theta > 0$, а $\Gamma(x)$ — *гамма-функция*. Для этой модели:

а) найдем оценку $\widehat{\theta}(\vec{X}_n)$ параметра θ методом максимального правдоподобия и покажем, что она смещенная;

б) найдем несмещенную оценку $\theta^*(\vec{X}_n)$ параметра θ;

в) покажем, что для рассматриваемой параметрической модели существует *достаточная статистика*;

г) убедимся, что оценка $\theta^*(\vec{X}_n)$ не эффективная, но *асимптотически эффективная*.

а. Запишем функцию правдоподобия

$$L(X_1,\ldots,X_n;\theta)=\frac{\theta^{n\lambda}}{(\Gamma(\lambda))^n}(X_1X_2\ldots X_n)^{\lambda-1}\exp\Big(-\theta\sum_{i=1}^{n}X_i\Big).$$

Отсюда находим

$$\frac{\partial\ln L(x_1,\ldots,x_n;\theta)}{\partial\theta}=\frac{n\lambda}{\theta}-\sum_{i=1}^{n}x_i.$$

Решая уравнение правдоподобия

$$\frac{n\lambda}{\theta}-\sum_{i=1}^{n}x_i=0,$$

получаем

$$\widehat{\theta}=\frac{\lambda}{\overline{x}}.$$

Покажем, что $\widehat{\theta}(\vec{X}_n)=\lambda/\overline{X}$ — смещенная оценка параметра θ. Плотность распределения случайной величины $Y=1/\overline{X}$ имеет вид

$$p_Y(\theta,\lambda,x)=\frac{(n\theta)^{n\lambda}}{\Gamma(n\lambda)}\Big(\frac{1}{y}\Big)^{n\lambda+1}e^{-\frac{n\theta}{y}}.$$

Этот результат можно получить, зная для данной модели характеристическую функцию [XVI]

$$f(t)=\Big(1-\frac{it}{\theta}\Big)^{-\lambda}.$$

Найдем математическое ожидание случайной величины $Y = 1/\overline{X}$:

$$\mathrm{M}(Y) = \mathrm{M}\Big(\frac{1}{\overline{X}}\Big) = \int\limits_0^\infty \frac{(n\theta)^{n\lambda}}{\Gamma(n\lambda)} y \Big(\frac{1}{y}\Big)^{n\lambda+1} e^{-\frac{n\theta}{y}}\, dy =$$

$$= \frac{(n\theta)^{n\lambda}}{\Gamma(n\lambda)} \int\limits_0^\infty t^{n\lambda-2} e^{-n\theta t}\, dt,$$

где $t = 1/y$. Используя равенство

$$\int\limits_0^\infty x^{\lambda-1} e^{-\alpha x}\, dx = \frac{\Gamma(\lambda)}{\alpha^\lambda}$$

и свойство гамма-функции $\Gamma(\lambda+1) = \lambda\Gamma(\lambda)$, получаем

$$\mathrm{M}\Big(\frac{1}{\overline{X}}\Big) = \frac{(n\theta)^{n\lambda}}{\Gamma(n\lambda)} \int\limits_0^\infty t^{(n\lambda-1)-1} e^{-n\theta t}\, dt =$$

$$= \frac{(n\theta)^{n\lambda}}{\Gamma(n\lambda)} \frac{\Gamma(n\lambda-1)}{(n\theta)^{n\lambda-1}} = \frac{n\theta}{n\lambda-1}.$$

Следовательно,

$$\mathrm{M}\Big(\frac{\lambda}{\overline{X}}\Big) = \frac{n\lambda\theta}{n\lambda-1} \neq \theta.$$

б. Легко заметить, что несмещенной оценкой параметра θ является

$$\theta^*(\vec{X}_n) = \frac{n\lambda-1}{n\overline{X}}.$$

в. Чтобы доказать существование достаточной статистики для рассматриваемой модели, используем критерий (2.7). Для этого функцию правдоподобия представим в виде

$$L(x_1,\ldots,x_n;\theta) = \frac{\theta^{n\lambda}}{(\Gamma(\lambda))^n} \exp\Big(-\theta\sum_{i=1}^n x_i\Big)(x_1\ldots x_n)^{\lambda-1} =$$

$$= g\Big(\sum_{i=1}^n x_i;\theta\Big)\, h(x_1,\ldots,x_n),$$

где

$$g\Big(\sum_{i=1}^{n} x_i\theta\Big) = \frac{\theta^{n\lambda}}{(\Gamma(\lambda))^n}\exp\Big(-\theta\sum_{i=1}^{n} x_i\Big),$$

$$h(x_1,\ldots,x_n) = (x_1\ldots x_n)^{\lambda-1}.$$

Из этого представления, согласно критерию (2.7), вытекает, что

$$T(X_1,\ldots,X_n) = \sum_{i=1}^{n} X_i$$

является достаточной статистикой.

г. Чтобы проверить, является ли несмещенная оценка $\theta^*(\vec{X}_n)$ эффективной, необходимо вычислить ее дисперсию и количество информации по Фишеру $I(\theta)$. Для дисперсии оценки, предполагая, что $n\lambda > 2$, получаем

$$\mathbf{D}\theta^*(\vec{X}_n) = \mathbf{M}\big(\theta^*(\vec{X}_n)\big)^2 - \big(\mathbf{M}\theta^*(\vec{X}_n)\big)^2 =$$

$$= \Big(\frac{n\lambda-1}{n}\Big)^2 \mathbf{M}\Big(\frac{1}{\overline{X}}\Big)^2 - \theta^2 =$$

$$= \Big(\frac{n\lambda-1}{n}\Big)^2 \int_0^\infty \frac{(n\theta)^{n\lambda}}{\Gamma(n\lambda)} y^2 \Big(\frac{1}{y}\Big)^{n\lambda+1} e^{-\frac{n\theta}{y}}\, dy - \theta^2 =$$

$$= \Big(\frac{n\lambda-1}{n}\Big)^2 \frac{n^2}{(n\lambda-1)(n\lambda-2)}\theta^2 - \theta^2 = \frac{\theta^2}{n\lambda-2}.$$

Отметим, что

$$\mathbf{M}X = \frac{\lambda}{\theta}, \qquad \mathbf{D}X = \frac{\lambda}{\theta^2}$$

(см. пример 2.12). Используя эти равенства, вычислим $I(\theta)$:

$$I(\theta) = \mathbf{M}\Big(\frac{\partial \ln p(X;\theta)}{\partial\theta}\Big)^2 = \mathbf{M}\Big(\frac{\lambda}{\theta} - X\Big)^2 = \mathbf{D}X = \frac{\lambda}{\theta^2}.$$

Теперь можно найти показатель эффективности по Рао — Крамеру

$$e(\theta) = \frac{1}{nI(\theta)\,\mathrm{D}\theta^*(\vec{X}_n)} = \frac{n\lambda - 2}{n\lambda}.$$

Поскольку $e(\theta) < 1$, *оценка* $\theta^*(\vec{X}_n)$ не является эффективной по Рао — Крамеру. Но при этом $\lim\limits_{n\to\infty} e(\theta) = 1$, так что $\theta^*(\vec{X}_n)$ — асимптотически эффективная оценка параметра θ. #

В заключение отметим, что для *нормальной модели* $N(\theta_1, \theta_2^2)$ выборочное среднее $\overline{X}$ является несмещенной эффективной оценкой параметра $\theta_1 = \mu$ независимо от того, известен параметр $\theta_2^2 = \sigma^2$ или нет. Для параметра $\theta_2^2 = \sigma^2$ оценка

$$\widehat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

является смещенной (см. теорему 2.2), а оценка

$$S^2(\vec{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad —$$

несмещенной.

При известном параметре $\theta_1 = \mu$ эффективной по Рао — Крамеру является оценка

$$\widetilde{S}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

(см. пример 2.5). При неизвестном параметре $\theta_1 = \mu$ оценка S^2 эффективная, но по Рао — Крамеру она не является эффективной*.

*См.: *Ивченко Г.И., Медведев Ю.И.*

Вопросы и задачи

2.1. Что называют точечной оценкой неизвестного параметра генеральной совокупности?

2.2. Какую точечную оценку называют несмещенной?

2.3. Какую точечную оценку называют состоятельной?

2.4. Какая точечная оценка является несмещенной, состоятельной и эффективной в классе линейных оценок для математического ожидания генеральной совокупности?

2.5. Какая точечная оценка для дисперсии генеральной совокупности является: а) смещенной; б) несмещенной? Являются ли эти оценки состоятельными?

2.6. Какую точечную оценку называют эффективной по Рао — Крамеру?

2.7. Запишите неравенство Рао — Крамера.

2.8. Что называют показателем эффективности по Рао — Крамеру?

2.9. Какую статистику для параметра θ называют достаточной?

2.10. Сформулируйте необходимое и достаточное условие существования достаточной статистики.

2.11. Какая связь существует между достаточными статистиками и эффективными по Рао — Крамеру оценками?

2.12. В чем состоит метод моментов нахождения точечных оценок?

2.13. В чем состоит метод максимального правдоподобия нахождения точечных оценок?

2.14. В условиях задачи 1.16 определить значение несмещенной оценки дисперсии ошибок прибора:

а) если значение измеряемой величины известно и равно 2800;

б) если значение измеряемой величины неизвестно.

Ответ: а) $S^2 = 1287{,}8$; б) $S^2 = 1508{,}5$.

2.15. В условиях задачи 1.18 определите значение несмещенной оценки дисперсии генеральной совокупности X.

Ответ: $S^2 = 1{,}205$.

2.16. Выборка объема n извлечена из равномерно распределенной на отрезке $[a, b]$ генеральной совокупности X. Известна длина этого отрезка $b - a = h$, но не известна середина интервала $c = \frac{a+b}{2}$. В качестве оценки середины интервала предлагается среднее арифметическое крайних членов вариационного ряда выборки. Покажите, что эта оценка несмещенная и состоятельная.

2.17. Из генеральной совокупности, распределенной по биномиальному закону, извлечена выборка объема n. Найдите методом моментов оценку неизвестного параметра p и покажите, что эта оценка будет несмещенной, состоятельной и эффективной по Рао — Крамеру.

2.18. Найдите методом максимального правдоподобия по выборке объема n точечную оценку геометрического распределения

$$\mathrm{P}\{X = x_i\} = p(1-p)^{x_i - 1},$$

где x_i — число испытаний до появления события; p — вероятность появления события в одном испытании.

Ответ: $\widehat{p}(\vec{X}_n) = 1/\overline{X}$.

2.19. Найдите методом максимального правдоподобия по выборке объема n точечную оценку параметра β гамма-распределения (α известно) с плотностью

$$f(x) = \frac{1}{\beta^{\alpha+1}\Gamma(\alpha+1)} x^{\alpha} e^{-x/\beta}, \quad \alpha > -1, \ \beta > 0, \ x \geqslant 0.$$

Ответ: $\widehat{\beta}(\vec{X}_n) = \dfrac{\overline{X}}{\alpha+1}$.

2.20. Имеется выборка объема n из генеральной совокупности X, распределенной по закону χ^2 с плотностью

$$p(x) = \frac{\alpha^k x^{k-1} e^{-\alpha x}}{\Gamma(k)}, \quad x > 0,$$

где α — неизвестный параметр. Найдите с помощью метода максимального правдоподобия оценку параметра α.

Ответ: $\widehat{\alpha}(\vec{X}_n) = k/\overline{X}$.

2.21. Из распределения с плотностью

$$p(x) = \frac{e^{-|x|}}{2(1 - e^{-\theta})}, \quad |x| \leqslant \theta,$$

извлечена выборка объема n. Найдите оценку максимального правдоподобия для параметра θ.

Ответ: $\widehat{\theta}(\vec{X}_n) = \max\limits_{i=\overline{1,n}} |X_i|$.

3. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ И ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

3.1. Понятия интервальной оценки и доверительного интервала

При оценивании неизвестных параметров наряду с рассмотренными выше *точечными оценками* используются также *интервальные оценки*. В отличие от точечной оценки интервальная оценка позволяет получить вероятностную характеристику точности оценивания неизвестного параметра.

Пусть $\vec{X}_n$ — *случайная выборка объема* n из *генеральной совокупности* X с функцией распределения $F(x;\theta)$, зависящей от параметра θ, значение которого неизвестно. Предположим, что для параметра θ построен интервал $\left(\underline{\theta}(\vec{X}_n),\, \overline{\theta}(\vec{X}_n)\right)$, где $\underline{\theta}(\vec{X}_n)$ и $\overline{\theta}(\vec{X}_n)$ являются функциями случайной выборки $\vec{X}_n$, такими, что выполняется равенство

$$\mathbf{P}\left\{\underline{\theta}(\vec{X}_n) < \theta < \overline{\theta}(\vec{X}_n)\right\} = \gamma. \tag{3.1}$$

В этом случае интервал $\left(\underline{\theta}(\vec{X}_n),\, \overline{\theta}(\vec{X}_n)\right)$ называют ***интервальной оценкой*** для параметра θ с ***коэффициентом доверия*** γ (или, сокращенно, ***γ-доверительной интервальной оценкой***), а $\underline{\theta}(\vec{X}_n)$ и $\overline{\theta}(\vec{X}_n)$ соответственно ***нижней*** и ***верхней границами*** интервальной оценки.

Интервальная оценка $\left(\underline{\theta}(\vec{X}_n),\, \overline{\theta}(\vec{X}_n)\right)$ представляет собой интервал со случайными границами, который с заданной вероятностью γ накрывает неизвестное истинное значение параметра θ. Таким образом, для различных *реализаций случайной выборки* $\vec{X}_n$, т.е. для различных элементов *выборочного пространства* $\mathcal{X}_n$, *статистики* $\underline{\theta}(\vec{X}_n)$ и $\overline{\theta}(\vec{X}_n)$ могут принимать различ-

ные значения. Более того, согласно (3.1), существует подмножество $\mathcal{K} \subset \mathcal{X}_n$, такое, что если $\vec{x}_n \in \mathcal{K}$, то $\theta \notin \big(\underline{\theta}(\vec{x}_n), \overline{\theta}(\vec{x}_n)\big)$.

При этом вероятностной характеристикой точности оценивания параметра θ является случайная величина

$$l(\vec{X}_n) = \overline{\theta}(\vec{X}_n) - \underline{\theta}(\vec{X}_n),$$

которая для любой реализации $\vec{x}_n$ случайной выборки $\vec{X}_n$ есть длина интервала $\big(\underline{\theta}(\vec{x}_n), \overline{\theta}(\vec{x}_n)\big)$.

Интервал $\big(\underline{\theta}(\vec{x}_n), \overline{\theta}(\vec{x}_n)\big)$ называют ***доверительным интервалом*** для параметра θ с коэффициентом доверия γ или ***γ-доверительным интервалом.***

Заметим, что наряду с термином „коэффициент доверия" широко используют также термины ***доверительная вероятность*** и ***уровень доверия.*** При этом коэффициент доверия γ чаще всего выбирают равным 0,9, 0,95 или 0,99, т.е. близким к 1.

В некоторых ситуациях (например, при рассмотрении дискретных случайных величин) вместо равенства (3.1) удается обеспечить лишь неравенство

$$\mathbf{P}\big\{\underline{\theta}(\vec{X}_n) < \theta < \overline{\theta}(\vec{X}_n)\big\} \geqslant \gamma,$$

т.е. построить интервальную оценку для параметра θ с коэффициентом доверия, не меньшим γ. Иногда требуется оценить параметр θ только снизу или только сверху. При этом, если

$$\mathbf{P}\big\{\underline{\theta}(\vec{X}_n) < \theta\big\} = \gamma,$$

то статистику $\underline{\theta}(\vec{X}_n)$ называют ***односторонней нижней γ-доверительной границей*** для параметра θ. Аналогично, если

$$\mathbf{P}\big\{\theta < \overline{\theta}(\vec{X}_n)\big\} = \gamma,$$

то статистику $\overline{\theta}(\vec{X}_n)$ называют ***односторонней верхней γ-доверительной границей*** для параметра θ.

Пример 3.1. Пусть θ — среднее значение предела прочности X некоторого материала, которое оценивают независимо друг от друга в каждой из N различных лабораторий по результатам n независимых натурных испытаний. Иначе говоря, среднее значение предела прочности в каждой лаборатории оценивают по „своим“ *экспериментальным данным*, представленным *выборкой объема* n, и в каждой лаборатории получают „свои“ значения верхней и нижней границ γ-доверительного интервала (рис. 3.1).

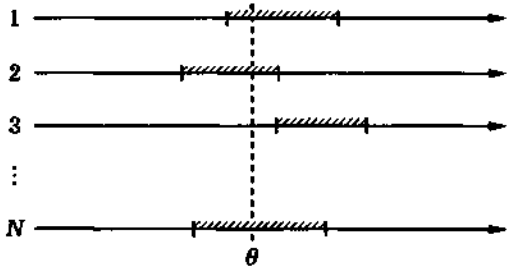


Рис. 3.1

Возможны случаи, когда γ-доверительный интервал для параметра θ не накрывает его истинного значения. Если M — число таких случаев, то при больших значениях N должно выполняться приближенное равенство $\gamma \approx (N-M)/N$. Таким образом, если опыт — получение выборки объема n в лаборатории, то уровень доверия γ — доля тех опытов (при их многократном независимом повторении), в каждом из которых γ-доверительный интервал накрывает истинное значение оцениваемого параметра.

3.2. Построение интервальных оценок

Пусть $\vec{X}_n$ — *случайная выборка объема* n *из генеральной совокупности* X с функцией распределения $F(x;\theta)$, зависящей от параметра θ, значение которого неизвестно. Рассмотрим один

из наиболее распространенных методов построения *интервальных оценок* для θ, связанный с использованием ***центральной статистики*** — любой *статистики* $T(\vec{X}_n, \theta)$, функция распределения которой

$$F_T(t) = \mathbf{P}\{T(\vec{X}_n, \theta) < t\}$$

не зависит от параметра θ. Примеры центральных статистик приведены в **3.3.**

Для упрощения дальнейших рассуждений будем предполагать следующее:

1) функция распределения $F_T(t)$ является непрерывной и возрастающей;

2) заданы такие положительные числа α и β, что *коэффициент доверия* $\gamma = 1 - \alpha - \beta$;

3) для любой выборки $\vec{x}_n$ из генеральной совокупности X функция $T(\vec{x}_n, \theta)$ является непрерывной и возрастающей (убывающей) функцией параметра $\theta \in \Theta$.

Согласно допущению 1, для любого $q \in (0, 1)$ существует единственный корень h_q уравнения $F_T(t) = q$, который называют квантилью уровня q функции распределения $F_T(t)$ случайной величины $T(\vec{X}_n, \theta)$. Таким образом, согласно допущению 2, имеют место равенства

$$\mathbf{P}\{h_\alpha < T(\vec{X}_n, \theta) < h_{1-\beta}\} = \\ = F_T(h_{1-\beta}) - F_T(h_\alpha) = 1 - \beta - \alpha = \gamma, \quad (3.2)$$

которые справедливы для любых возможных значений параметра θ, так как $T(\vec{X}_n, \theta)$ — центральная статистика, и ее функция распределения $F_T(t)$ не зависит от θ. Для преобразования (3.2) в (3.1), т.е. для построения искомой интервальной оценки, воспользуемся следующими соображениями.

Пусть для определенности функция $T(\vec{x}_n, \theta)$ является возрастающей функцией параметра θ. Тогда, согласно допущению 3, для каждой выборки $\vec{x}_n \in \mathcal{X}_n$ уравнения $T(\vec{x}_n, \theta) = h_\alpha$ и

$T(\vec{x}_n,\theta)=h_{1-\beta}$ имеют единственные решения $\underline{\theta}(\vec{x}_n)$ и $\overline{\theta}(\vec{x}_n)$ соответственно. При этом неравенства

$$h_\alpha < T(\vec{x}_n,\theta) < h_{1-\beta}, \qquad \underline{\theta}(\vec{x}_n) < \theta < \overline{\theta}(\vec{x}_n)$$

являются равносильными, т.е. для любой выборки $\vec{x}_n \in \mathcal{X}_n$ они выполняются или не выполняются одновременно. Таким образом,

$$\gamma = \mathbf{P}\left\{h_\alpha < T(\vec{X}_n,\theta) < h_{1-\beta}\right\} = \mathbf{P}\left\{\underline{\theta}(\vec{X}_n) < \theta < \overline{\theta}(\vec{X}_n)\right\}$$

и $\left(\underline{\theta}(\vec{X}_n),\, \overline{\theta}(\vec{X}_n)\right)$ — искомая интервальная оценка.

Завершая рассуждения, заметим, что фактически построение *доверительного интервала* сводится к выполнению следующих действий:

1) построение центральной статистики $T(\vec{X}_n,\theta)$ с известной функцией распределения $F_T(t)$;

2) представление заданного коэффициента доверия γ в виде $\gamma = 1-\alpha-\beta$;

3) нахождение квантилей h_α и $h_{1-\beta}$ уровня α и $1-\beta$ функции распределения $F_T(t)$;

4) нахождение значений *нижней* $\underline{\theta}(\vec{x}_n)$ и *верхней* $\overline{\theta}(\vec{x}_n)$ *границ* искомой интервальной оценки путем решения уравнений

$$T(\vec{x}_n,\underline{\theta}) = h_\alpha, \qquad T(\vec{x}_n,\overline{\theta}) = h_{1-\beta} \tag{3.3}$$

соответственно в случае, когда $T(\vec{x}_n,\theta)$ — возрастающая функция параметра θ. Если же $T(\vec{x}_n,\theta)$ — убывающая функция параметра θ, то $\underline{\theta}(\vec{X}_n)$ и $\overline{\theta}(\vec{X}_n)$ получают путем решения уравнений

$$T(\vec{x}_n,\underline{\theta}) = h_{1-\beta}, \qquad T(\vec{x}_n,\overline{\theta}) = h_\alpha \tag{3.4}$$

соответственно.

3.3. Примеры построения интервальных оценок

Рассмотрим построение *интервальной оценки* для параметров некоторых часто используемых распределений.

Экспоненциальное распределение. Пусть $\vec{X}_n$ — *случайная выборка объема* n *из генеральной совокупности* X *с экспоненциальным законом* распределения, имеющим плотность распределения

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0; \\ 0, & x < 0, \end{cases}$$

где $\lambda > 0$ — неизвестный параметр. Требуется построить интервальную оценку для параметра λ по данным случайной выборки $\vec{X}_n$.

В данном случае $\theta = \lambda$. Рассмотрим *статистику*

$$T(\vec{X}_n, \lambda) = 2\lambda n \overline{X},$$

где $\overline{X}$ — *выборочное среднее* для $\vec{X}_n$. Эта статистика имеет χ^2*-распределение с* $2n$ *степенями свободы* (см. **Д.3.1**), т.е. является *центральной статистикой*. Уравнения (3.3) в данном случае принимают вид

$$2\underline{\lambda} n\overline{x} = \chi^2_{\alpha}(2n), \qquad 2\overline{\lambda} n\overline{x} = \chi^2_{1-\beta}(2n),$$

где $\chi^2_q(2n)$ — квантиль уровня q для χ^2-распределения с $2n$ степенями свободы.

Получаем, что *нижняя* и *верхняя границы* интервальной оценки с *коэффициентом доверия* $\gamma = 1 - \alpha - \beta$ для параметра экспоненциального распределения λ имеют вид

$$\underline{\lambda}(\vec{X}_n) = \frac{\chi^2_{\alpha}(2n)}{2n\overline{X}}, \qquad \overline{\lambda}(\vec{X}_n) = \frac{\chi^2_{1-\beta}(2n)}{2n\overline{X}}.$$

Нормальное распределение. Пусть $\vec{X}_n$ — случайная выборка объема n из генеральной совокупности X, распределенной по нормальному закону с параметрами μ и σ^2. Рассмотрим

некоторые варианты построения интервальных оценок для параметров μ, σ.

Вариант 1 — оценка для математического ожидания при известной дисперсии. В данном случае статистика

$$T(\vec{X}_n, \mu) = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

имеет стандартное нормальное распределение с параметрами $\mu = 0$, $\sigma^2 = 1$, т.е. является центральной статистикой. Функция $T(\vec{X}_n, \mu)$ является убывающей функцией по μ, и система уравнений (3.4) принимает вид

$$\frac{\sqrt{n}\big(\overline{x} - \underline{\mu}(\vec{x}_n)\big)}{\sigma} = u_{1-\beta}, \qquad \frac{\sqrt{n}\big(\overline{x} - \overline{\mu}(\vec{x}_n)\big)}{\sigma} = u_{\alpha},$$

где u_q — квантиль уровня q стандартного нормального распределения. Учитывая, что для нормального закона $u_{1-\alpha} = -u_{\alpha}$, получаем следующие нижнюю и верхнюю границы *γ-доверительного интервала* для параметра μ при $\gamma = 1 - \alpha - \beta$:

$$\underline{\mu}(\vec{x}_n) = \overline{x} - \frac{\sigma}{\sqrt{n}}u_{1-\beta}, \qquad \overline{\mu}(\vec{x}_n) = \overline{x} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha}.$$

Вариант 2 — оценка математического ожидания при неизвестной дисперсии. При неизвестной дисперсии статистика

$$T(\vec{X}_n, \mu) = \frac{\overline{X} - \mu}{S(\vec{X}_n)}\sqrt{n}$$

является центральной, так как имеет *распределение Стьюдента* с $n-1$ степенями свободы (см. **Д.3.1**), которое не зависит от μ и σ^2. Система уравнений (3.4) в данном случае принимает вид

$$\frac{\sqrt{n}\big(\overline{x} - \underline{\mu}(\vec{x}_n)\big)}{S(\vec{x}_n)} = t_{1-\beta}(n-1), \qquad \frac{\sqrt{n}\big(\overline{x} - \overline{\mu}(\vec{x}_n)\big)}{S(\vec{x}_n)} = t_{\alpha}(n-1),$$

где $t_q(n-1)$ — квантиль уровня q распределения Стьюдента с $n-1$ степенями свободы. Поскольку плотность распределения Стьюдента — четная функция, то $t_\alpha(n-1) = -t_{1-\alpha}(n-1)$. Отсюда заключаем, что нижняя и верхняя границы интервальной оценки с коэффициентом доверия $\gamma = 1-\alpha-\beta$ для параметра μ в случае с неизвестной дисперсией можно определить по формулам

$$\underline{\mu}(\vec{X}_n) = \overline{X} - \frac{S(\vec{X}_n)}{\sqrt{n}} t_{1-\beta}(n-1), \quad \overline{\mu}(\vec{X}_n) = \overline{X} + \frac{S(\vec{X}_n)}{\sqrt{n}} t_{1-\alpha}(n-1).$$

Вариант 3 — оценка среднего квадратичного отклонения. Рассмотрим статистику

$$T(\vec{X}_n, \sigma) = \frac{(n-1)S^2(\vec{X}_n)}{\sigma^2}.$$

Эта статистика является центральной, так как имеет χ^2-распределение с $n-1$ степенями свободы (см. **Д.3.1**), которое не зависит от μ и σ^2. При этом $T(\vec{x}_n, \sigma)$ — убывающая функция параметра σ. Исходя из этого, согласно (3.4), находим нижнюю и верхнюю границы интервальной оценки для параметра σ с коэффициентом доверия $\gamma = 1-\alpha-\beta$:

$$\underline{\sigma}(\vec{X}_n) = \frac{S(\vec{X}_n)\sqrt{n-1}}{\sqrt{\chi^2_{1-\beta}(n-1)}}, \qquad \overline{\sigma}(\vec{X}_n) = \frac{S(\vec{X}_n)\sqrt{n-1}}{\sqrt{\chi^2_{\alpha}(n-1)}},$$

где $\chi^2_q(n-1)$ — квантиль уровня q для χ^2-распределения с $n-1$ степенями свободы.

Приближенные интервальные оценки. Сначала рассмотрим два частных случая построения таких оценок.

Пусть требуется найти интервальную оценку для математического ожидания в случае, когда закон распределения генеральной совокупности X неизвестен. Предполагаем, что существуют конечные математическое ожидание $\mu = \mathbf{M}X$ и дисперсия $\sigma^2 = \mathbf{D}X$.

Рассмотрим статистику

$$T(\vec{X}_n) = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}.$$

В соответствии с центральной предельной теоремой эта статистика при больших объемах n случайной выборки $\vec{X}_n$ имеет закон распределения, близкий к стандартному нормальному. Поэтому при достаточно больших n неравенства

$$-u_{1-\beta} \leqslant \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \leqslant u_{1-\alpha}$$

выполняются с вероятностью, близкой к величине $\gamma = 1 - \alpha - \beta$, где u_q — квантиль уровня q стандартного нормального распределения. Приведенные неравенства эквивалентны следующим:

$$\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} \leqslant \mu \leqslant \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\beta}.$$

Эти неравенства не дают еще интервальной оценки для параметра μ, так как их левая и правая части содержат неизвестный параметр σ. Применяя еще одно приближение, а именно: подставляя в указанные неравенства вместо неизвестного точного значения σ его *оценку* $S(\vec{X}_n)$, получаем нижнюю и верхнюю границы (приближенной) интервальной оценки с коэффициентом доверия $\gamma = 1 - \alpha - \beta$ для математического ожидания μ:

$$\underline{\mu}(\vec{X}_n) = \overline{X} - \frac{S(\vec{X}_n)}{\sqrt{n}} u_{1-\beta}, \qquad \overline{\mu}(\vec{X}_n) = \overline{X} + \frac{S(\vec{X}_n)}{\sqrt{n}} u_{1-\alpha}.$$

Пусть проводится серия из n испытаний по схеме Бернулли и X_i, $i = \overline{1, n}$, — исход i-го испытания („успех“ или „отказ“). По данным случайной выборки $\vec{X}_n = (X_1, \ldots, X_n)$ построим доверительный интервал для вероятности p „успеха“ в каждом отдельном испытании.

Рассмотрим суммарное число „успехов“ в серии из n испытаний, т.е. введем случайную величину

$$K(\vec{X}_n) = X_1 + \ldots + X_n,$$

которая имеет биномиальное распределение с параметром p. Для построения доверительного интервала для p воспользуемся статистикой

$$T(\vec{X}_n, p) = \frac{K(\vec{X}_n) - np}{\sqrt{np(1-p)}}.$$

В соответствии с предельной теоремой Муавра — Лапласа статистика $T(\vec{X}_n, p)$ при больших объемах n случайной выборки $\vec{X}_n$ имеет закон распределения, близкий к стандартному нормальному. Тем самым неравенства

$$-u_{1-\beta} \leqslant \frac{K(\vec{X}_n) - np}{\sqrt{np(1-p)}} \leqslant u_{1-\alpha}$$

выполняются с вероятностью, которую при больших n можно считать приближенно равной $\gamma = 1 - \alpha - \beta$. Указанные неравенства могут быть записаны в виде

$$\frac{K(\vec{X}_n)}{n} - \frac{u_{1-\alpha}}{\sqrt{n}}\sqrt{p(1-p)} \leqslant p \leqslant \frac{K(\vec{X}_n)}{n} + \frac{u_{1-\beta}}{\sqrt{n}}\sqrt{p(1-p)}.$$

Эти неравенства еще не дают интервальной оценки параметра p, так как их левая и правая части содержат этот параметр. Поэтому на практике в указанные части неравенств часто подставляют вместо неизвестного точного значения p его оценку $\widehat{p}(\vec{X}_n) = K(\vec{X}_n)/n$. В результате получают следующие верхнюю и нижнюю границы интервальной оценки с коэффи-

циентом доверия $\gamma = 1 - \alpha - \beta$ для параметра p:

$$\underline{p}(\vec{X}_n) = \frac{K(\vec{X}_n)}{n} - \frac{u_{1-\alpha}}{\sqrt{n}}\sqrt{\frac{K(\vec{X}_n)}{n}\left(1 - \frac{K(\vec{X}_n)}{n}\right)},$$

$$\overline{p}(\vec{X}_n) = \frac{K(\vec{X}_n)}{n} + \frac{u_{1-\beta}}{\sqrt{n}}\sqrt{\frac{K(\vec{X}_n)}{n}\left(1 - \frac{K(\vec{X}_n)}{n}\right)}.$$

Подчеркнем, что эти доверительные границы являются приближенными и могут использоваться при достаточно больших объемах наблюдений n.

Приведенный способ построения приближенного доверительного интервала для параметра p биномиального распределения может применяться и в следующей более общей ситуации. Пусть $\widehat{\theta}(\vec{X}_n)$ — *точечная несмещенная оценка* для параметра θ, построенная по данным случайной выборки $\vec{X}_n$. Обозначим через

$$V_n(\theta) = \mathrm{M}\big(\widehat{\theta}(\vec{X}_n) - \theta\big)^2$$

значение дисперсии оценки $\widehat{\theta}(\vec{X}_n)$. Предположим, что оценка $\widehat{\theta}(\vec{X}_n)$ имеет *асимптотически нормальное распределение*. Другими словами, нормированная случайная величина

$$\eta_n = \frac{\widehat{\theta}(\vec{X}_n) - \theta}{\sqrt{V_n(\theta)}}$$

имеет распределение, которое при $n \to \infty$ сходится к стандартному нормальному распределению. В этом случае неравенства

$$-u_{1-\beta} \leqslant \frac{\widehat{\theta}(\vec{X}_n) - \theta}{\sqrt{V_n(\theta)}} \leqslant u_{1-\alpha},$$

где u_q — квантиль уровня q стандартного нормального закона распределения, выполняются с вероятностью, которую при достаточно больших n можно считать приближенно равной

$\gamma = 1 - \alpha - \beta$. Указанные неравенства эквивалентны (см. **3.2**) следующим:

$$\widehat{\theta}(\vec{X}_n) - u_{1-\alpha}\sqrt{V_n(\theta)} \leqslant \theta \leqslant \widehat{\theta}(\vec{X}_n) + u_{1-\beta}\sqrt{V_n(\theta)}.$$

Записанные неравенства еще не дают интервальной оценки для θ, так как их левая и правая части содержат неизвестный параметр θ. Подставляя в левую и правую части указанных неравенств вместо θ оценку $\widehat{\theta}(\vec{X}_n)$, получаем окончательно следующие *нижнюю* и *верхнюю границы* для параметра θ с *коэффициентом доверия* $\gamma = 1 - \alpha - \beta$:

$$\underline{\theta}(\vec{X}_n) = \widehat{\theta}(\vec{X}_n) - u_{1-\alpha}\sqrt{V_n(\widehat{\theta}(\vec{X}_n))},$$

$$\overline{\theta}(\vec{X}_n) = \widehat{\theta}(\vec{X}_n) + u_{1-\beta}\sqrt{V_n(\widehat{\theta}(\vec{X}_n))}.$$

Изложенный метод является приближенным и может применяться при достаточно большом объеме случайной выборки. Заметим, что его использование фактически связано с „двойным приближением", а именно: закон распределения оценки $\widehat{\theta}(\vec{X}_n)$ заменяют нормальным и, кроме того, в приведенных формулах для границ $\underline{\theta}(\vec{X}_n)$, $\overline{\theta}(\vec{X}_n)$ интервальной оценки в дисперсию $V_n(\theta)$ вместо точного значения θ подставляют его оценку $\widehat{\theta}(\vec{X}_n)$. При малых и средних объемах случайной выборки применение указанного метода может приводить к значительным ошибкам. Поэтому использовать его следует с достаточной степенью осторожности и лишь в качестве первого приближения.

Пример 3.2. Рассмотрим построение приближенного доверительного интервала для параметра p биномиального распределения. Пусть проводилось $n = 16$ независимых испытаний с неизвестной вероятностью p „успеха" в каждом испытании, при этом наблюдалось $k = 8$ „успехов". Определим значения границ доверительного интервала для p с коэффициентом доверия $\gamma = 0{,}9$.

Значение точечной оценки параметра p определяется как

$$\widehat{p} = \frac{k}{n},$$

дисперсия этой оценки [XVI]

$$V_n(p) = \frac{p(1-p)}{n}.$$

Применяя приведенные выше формулы, получаем следующие значения для нижней и верхней границ доверительного интервала:

$$\underline{p} = \widehat{p} - u_{0,95}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0{,}294,$$

$$\overline{p} = \widehat{p} + u_{0,95}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0{,}706.$$

3.4. Метод доверительных множеств

Пусть $\vec{X}_n$ — *случайная выборка объема* n *из генеральной совокупности* X, закон распределения которой зависит от r-мерного вектора параметров $\vec{\theta}$. Каждому фиксированному значению вектора параметров $\vec{\theta}$ поставим в соответствие такое множество $H_{\vec{\theta}}$ из *выборочного пространства* $\mathcal{X}_n$, что

$$\mathbf{P}\{\vec{X}_n \in H_{\vec{\theta}}\} \geqslant \gamma,$$

где γ — заданный *коэффициент доверия*.

Как известно (см. **3.1**), *нижняя* и *верхняя границы интервальной оценки* (множества в $\mathbb{R}$) являются случайными величинами, поскольку они функции случайной выборки. Теперь в $\mathbb{R}^r$ рассмотрим такое множество $D_{\vec{X}_n}$ со случайной границей, чтобы при каждом фиксированном значении вектора параметров $\vec{\theta}$ случайные события

$$\{\vec{\theta} \in D_{\vec{X}_n}\}, \qquad \{\vec{X}_n \in H_{\vec{\theta}}\}$$

были эквивалентны, т.е.

$$\mathbf{P}\{\vec{\theta} \in D_{\vec{X}_n}\} \geqslant \gamma.$$

Полученную таким образом совокупность множеств $D_{\vec{X}_n}$ называют ***системой γ-доверительных множеств***, а рассмотренную процедуру — ***методом доверительных множеств*** (методом Неймана) для параметра θ.

Если θ — скаляр, то метод доверительных множеств имеет простую и наглядную графическую интерпретацию (рис. 3.2). Поэтому все дальнейшие рассуждения проведем именно для этого случая, т.е. при $r = 1$.

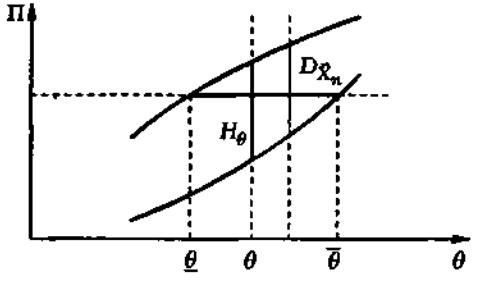


Рис. 3.2

Заметим, что процедура построения доверительных множеств $D_{\vec{X}_n}$ основана на выборе множеств H_θ, а это может быть реализовано различными способами, в том числе и с использованием некоторой *статистики* $\Pi = \Pi(\vec{X})$. Зачастую в качестве статистики $\Pi(\vec{X}_n)$ используют *несмещенную точечную оценку* параметра θ.

Для упрощения дальнейших рассуждений функцию распределения $F_\Pi(t,\theta)$ статистики $\Pi(\vec{X}_n)$ будем предполагать непрерывной, возрастающей по t и убывающей по θ.

Каждому возможному значению параметра θ поставим в соответствие значения $t_1 = t_1(\theta)$, $t_2 = t_2(\theta)$, выбираемые из условий

$$F_\Pi(t_1,\theta) = \alpha, \qquad F_\Pi(t_2,\theta) = 1 - \beta. \tag{3.5}$$

Таким образом, $t_1(\theta)$, $t_2(\theta)$ являются соответственно квантилями уровней α и $1-\beta$ для функции распределения $F_{\Pi}(t,\theta)$ статистики $\Pi(\vec{X}_n)$. При этом выполняется равенство

$$\mathbf{P}\{t_1(\theta) \leqslant \Pi(\vec{X}_n) \leqslant t_2(\theta)\} = \gamma,$$

где $\gamma = 1-\alpha-\beta$. Множество значений статистики $\Pi(\vec{X}_n)$, принадлежащих отрезку $[t_1(\theta), t_2(\theta)]$, обозначим H_θ и назовем ***γ-зоной*** (***гамма-зоной***) для θ (см. рис. 3.2). Как следует из этого определения, для любого возможного значения параметра θ вероятность того, что статистика $\Pi(\vec{X}_n)$ попадет в γ-зону, равна γ.

Далее, каждому значению статистики $\Pi(\vec{X}_n)$ поставим в соответствие *интервал* тех значений θ, для которых данное значение статистики $\Pi(\vec{X}_n)$ попадает в γ-зону (см. рис. 3.2). Значения нижней $\underline{\theta}(\vec{X}_n)$ и верхней $\overline{\theta}(\vec{X}_n)$ границ этого интервала определяются из условий

$$t_2\big(\underline{\theta}(\vec{x}_n)\big) = \Pi(\vec{x}_n), \qquad t_1\big(\overline{\theta}(\vec{x}_n)\big) = \Pi(\vec{x}_n),$$

которые в силу (3.5) эквивалентны следующим:

$$F\big(\Pi(\vec{x}_n), \overline{\theta}\big) = \alpha, \qquad F\big(\Pi(\vec{x}_n), \underline{\theta}\big) = 1-\beta. \tag{3.6}$$

Построенный таким образом интервал является γ-доверительной оценкой для параметра θ. Действительно, при любом возможном, а следовательно, и при неизвестном истинном значении θ интервал $\big(\underline{\theta}(\vec{X}_n), \overline{\theta}(\vec{X}_n)\big)$ накрывает значение θ тогда и только тогда, когда наблюдаемое значение статистики $\Pi(\vec{X}_n)$ попадает в γ-зону H_θ для данного значения θ. Тем самым, согласно определению γ-зоны, выполняется равенство

$$\mathbf{P}\{\underline{\theta}(\vec{X}_n) \leqslant \theta \leqslant \overline{\theta}(\vec{X}_n)\} = \gamma.$$

Если функция распределения $F(t;\theta)$ возрастает по параметру θ, то границы $t_1(\theta)$ и $t_2(\theta)$ γ-зоны убывают по θ. Повторяя

предыдущие рассуждения, заключаем, что в этом случае значения нижней и верхней границ формально (см. **3.2**) определяются из условий

$$F_{\Pi}\big(\Pi(\vec{x}_n),\underline{\theta}(\vec{x}_n)\big)=\alpha,\qquad F_{\Pi}\big(\Pi(\vec{x}_n),\overline{\theta}(\vec{x}_n)\big)=1-\beta. \tag{3.7}$$

Этот метод применяют аналогичным образом и в тех случаях, когда статистика $\Pi(\vec{X}_n)$ является дискретной случайной величиной. Рассмотрим, например, случай, когда статистика $\Pi(\vec{X}_n)$ принимает неотрицательные целые значения 0, 1, 2, ...

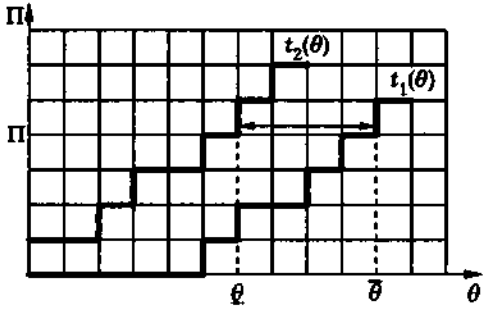


Рис. 3.3

В отличие от непрерывного случая, рассмотренного выше, границы γ-зоны теперь становятся ступенчатыми кривыми (рис. 3.3). При данном фиксированном значении θ границу $t_1(\theta)$ γ-зоны определим как максимальное из чисел k, таких, что выполняется неравенство

$$\mathbf{P}\big\{\Pi(\vec{X}_n)\geqslant k\big\}=1-F\big(k;\theta\big)\geqslant 1-\alpha.$$

Границу $t_2(\theta)$ γ-зоны определим как минимальное из чисел k, удовлетворяющих неравенству

$$\mathbf{P}\big\{\Pi(\vec{X}_n)\leqslant k\big\}=F\big(k+1;\theta\big)\geqslant 1-\beta.$$

Нижнюю $\underline{\theta}(\vec{X}_n)$ и верхнюю $\overline{\theta}(\vec{X}_n)$ границы интервальной оценки параметра θ с коэффициентом доверия не меньше $\gamma=$

$= 1 - \alpha - \beta$ определим как минимальное и максимальное значения среди всех θ, удовлетворяющих неравенствам

$$t_1(\theta) \leqslant \Pi(\vec{X}_n) \leqslant t_2(\theta), \tag{3.8}$$

т.е. среди всех θ, принадлежащих γ-зоне при данном значении статистики $\Pi(\vec{X}_n)$, полученном в результате эксперимента. Неравенства (3.8) эквивалентны неравенствам

$$\begin{cases} F_\Pi\big(\Pi(\vec{x}_n)+1, \theta\big) \geqslant \alpha, \\ F_\Pi\big(\Pi(\vec{x}_n), \theta\big) \leqslant 1-\beta. \end{cases}$$

Отсюда получаем, что если функция распределения статистики $\Pi(\vec{X}_n)$ убывает по θ, то нижнюю и верхнюю границы γ-интервальной оценки для θ формально (см. **3.2**) можно определить из уравнений

$$\begin{cases} F_\Pi\big(\Pi(\vec{X}_n), \underline{\theta}(\vec{X}_n)\big) = 1-\beta, \\ F_\Pi\big(\Pi(\vec{X}_n)+1, \overline{\theta}(\vec{X}_n)\big) = \alpha. \end{cases} \tag{3.9}$$

Аналогично, если функция распределения статистики $\Pi(\vec{X}_n)$ возрастает по θ, то границы γ-интервальной оценки для θ формально (см. **3.2**) можно найти из уравнений

$$\begin{cases} F_\Pi\big(\Pi(\vec{X}_n)+1, \underline{\theta}(\vec{X}_n)\big) = \alpha, \\ F_\Pi\big(\vec{X}_n, \overline{\theta}(\vec{X}_n)\big) = 1-\beta, \end{cases} \tag{3.10}$$

где коэффициент доверия $\gamma = 1 - \alpha - \beta$.

Рассмотрим далее в качестве примеров построение интервальных оценок для параметров биномиального распределения и распределения Пуассона.

Интервальная оценка Клоппера — Пирсона для параметра биномиального распределения. Пусть дискретная случайная величина X_i, $i = \overline{1,n}$, характеризует исход i-го испытания в серии из n испытаний, проводимых по схеме Бернулли. Тогда случайная величина $K = X_1 + \ldots + X_n$ — число

успехов в n испытаниях. При этом $K = K(\vec{X}_n)$ — функция случайной выборки $\vec{X}_n = (X_1, \ldots, X_n)$. В рассматриваемом случае $\Pi(\vec{X}_n) = K(\vec{X}_n)$.

Функция распределения статистики $K(\vec{X}_n)$ имеет вид

$$F(x;p) = \begin{cases} \sum\limits_{j<x} C_n^j p^j (1-p)^{n-j}, & x > 0; \\ 0, & x \leqslant 0. \end{cases}$$

Эта функция убывающая по p. Применяя общую формулу (3.9), получаем, что нижняя и верхняя границы интервальной оценки с коэффициентом доверия $\gamma = 1 - \alpha - \beta$ для параметра p (см. **3.2**) определяются из следующих уравнений:

$$\sum_{j=0}^{K(\vec{X}_n)-1} C_n^j \underline{p}^j(\vec{X}_n)(1 - \underline{p}(\vec{X}_n))^{n-j} = 1 - \beta \quad \text{при} \quad K(\vec{X}_n) \geqslant 1,$$

$$\sum_{j=0}^{K(\vec{X}_n)} C_n^j \overline{p}^j(\vec{X}_n)(1 - \overline{p}(\vec{X}_n))^{n-j} = \alpha \quad \text{при} \quad K(\vec{X}_n) \leqslant n-1.$$

Эти уравнения называются ***уравнениями Клоппера — Пирсона***. При $K(\vec{X}_n) = 0$ нижняя граница $\underline{p}(\vec{X}_n) = 0$. При $K(\vec{X}_n) = n$ верхняя граница $\overline{p}(\vec{X}_n) = 1$.

Заметим, что приведенные уравнения Клоппера — Пирсона могут быть также выражены через неполную бета-функцию (см. **Д.3.1**):

$$\mathrm{B}_{\underline{p}(\vec{X}_n)}(K(\vec{X}_n), n - K(\vec{X}_n) + 1) = \beta,$$

$$\mathrm{B}_{\overline{p}(\vec{X}_n)}(K(\vec{X}_n) + 1, n - K(\vec{X}_n)) = 1 - \alpha.$$

Пример 3.3. Пусть число испытаний $n = 16$, а число наблюдаемых „успехов“ $K = 8$, коэффициент доверия $\gamma = 0{,}95$.

Полагая $\alpha = \beta = 0{,}025$, получаем

$$\underline{p} = 0{,}247, \qquad \overline{p} = 0{,}753.$$

Доверительный интервал для параметра распределения Пуассона. Пусть X — дискретная случайная величина, имеющая распределение Пуассона с неизвестным параметром λ. Требуется построить доверительный интервал для параметра λ на основе наблюдаемого значения d_* случайной величины X.

Согласно предположению, функция распределения случайной величины X имеет вид

$$F(x;\lambda) = \begin{cases} \displaystyle\sum_{j<x} \frac{\lambda^j}{j!} e^{-\lambda}, & x > 0; \\ 0, & x \leqslant 0. \end{cases}$$

Это функция, убывающая по λ. Применяя снова формулы (3.9), получаем уравнения

$$e^{-\underline{\lambda}} \sum_{j=0}^{d_*-1} \frac{\underline{\lambda}^j}{j!} = 1 - \beta, \qquad e^{-\overline{\lambda}} \sum_{j=0}^{d_*} \frac{\overline{\lambda}^j}{j!} = \alpha,$$

решая которые находим значения нижней и верхней границ доверительного интервала для λ с коэффициентом доверия $\gamma = 1 - \alpha - \beta$. При $d_* = 0$ значение нижней границы $\underline{\lambda} = 0$.

3.5. Решение типовых примеров

Пример 3.4. При помощи вольтметра, точность которого характеризуется средним квадратичным отклонением 0,2 В, проведено 10 измерений напряжения бортовой батареи. Найдем *доверительный интервал* для истинного значения напряжения батареи с *коэффициентом доверия* $\gamma = 0{,}95$, если среднее арифметическое результатов наблюдений $\overline{x} = 50{,}2$ В. Контролируемый признак имеет нормальный закон распределения.

Для нахождения доверительного интервала (см. **3.3**)

$$\left(\overline{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

где $u_{1-\alpha/2}$ — квантиль нормального распределения уровня $1-\alpha/2$, а $\alpha = 1-\gamma$, обратимся к таблице квантилей нормального распределения (см. табл. П.1). По этой таблице находим

$$u_{1-\alpha/2} = u_{0,975} = 1{,}96.$$

Поскольку

$$u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 1{,}96\frac{0{,}2}{\sqrt{10}} \approx 0{,}1,$$

доверительный интервал имеет вид $(50{,}2-0{,}1,\ 50{,}2+0{,}1)$, или $(50{,}1,\ 50{,}3)$.

Пример 3.5. Из большой партии электроламп было отобрано случайным образом 400 шт. для определения средней продолжительности горения. Выборочная средняя продолжительность горения ламп оказалась равной 1220 ч. Найдем с коэффициентом доверия $\gamma = 0{,}997$ доверительный интервал для средней продолжительности горения электролампы по всей партии, если среднее квадратичное отклонение продолжительности горения равно 35 ч.

Независимо от закона распределения *генеральной совокупности* X (продолжительности горения электролампы) *статистика*

$$\frac{\overline{X}-\mu}{\sigma}\sqrt{n},$$

где

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

имеет *асимптотически нормальное распределение* с параметрами $(0, 1)$, что следует из центральной предельной теоремы.

Поскольку *объем выборки* большой ($n = 400$), то границы доверительного интервала находим так же, как и в примере 3.4.

Для $\alpha = 1 - \gamma = 0{,}0028$ находим квантиль нормального распределения $u_{1-\alpha/2} = u_{0{,}9986} = 2{,}98$. В силу соотношений

$$u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = \frac{2{,}98 \cdot 35}{\sqrt{400}} \approx 5{,}52$$

доверительный интервал имеет вид $(1220 - 5{,}52,\ 1220 + 5{,}52)$, или $(1214{,}48,\ 1225{,}52)$.

Пример 3.6. В результате пусков 10 ракет получены (в условных единицах) значения боковых отклонений точек попадания от точек прицеливания (табл. 3.1).

Таблица 3.1

Номер ракеты	1	2	3	4	5	6	7	8	9	10
Отклонение	1,0	2,0	1,0	−0,1	−0,5	5,0	−1,0	3,0	0,5	1,0

Полагая, что случайная величина X (случайное боковое отклонение точек попадания от точек прицеливания) имеет нормальное распределение, построим доверительный интервал для ее математического ожидания с коэффициентом доверия $\gamma = 0{,}99$.

Для нахождения доверительного интервала воспользуемся статистикой

$$\frac{\overline{X} - \mu}{\widehat{\sigma}(\vec{X}_n)}\sqrt{n-1},$$

которая имеет *распределение Стьюдента* с $n - 1$ *степенью свободы*. *Выборочное среднее* имеет значение

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\Big(1 + 0{,}2 + 1 - 0{,}1 - 0{,}5 + 5 - \\ - 1 + 3 + 0{,}5 + 1\Big) = 1{,}01,$$

а *выборочная дисперсия* — значение

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{10}\Big((-0{,}01)^2 + 0{,}99^2 + (-0{,}01)^2 + \\ + (-1{,}11)^2 + (-1{,}51)^2 + 3{,}99^2 + (-2{,}01)^2 + 1{,}99^2 + \\ + (-0{,}51)^2 + (-0{,}01)^2\Big) = 2{,}8673.$$

Значение *выборочного среднего квадратичного отклонения* равно $\widehat{\sigma} = \sqrt{2{,}749} \approx 1{,}69$. По таблице квантилей распределения Стьюдента (см. табл. П.2) для $n - 1 = 9$ находим квантиль $t_{1-\alpha/2}(n-1)$ уровня $1 - \alpha/2$. По условию задачи

$$\alpha = 1 - \gamma = 1 - 0{,}99 = 0{,}01.$$

Следовательно, $t_{1-\alpha/2}(n-1) = t_{0{,}995}(9) = 3{,}25$. Вычислив

$$t_{1-\alpha/2}(n-1)\frac{\widehat{\sigma}}{\sqrt{n-1}} = t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}} = 3{,}25 \cdot \frac{1{,}65}{3} \approx 1{,}79,$$

получаем доверительный интервал $(1{,}01 - 1{,}79,\ 1{,}01 + 1{,}79)$, или $(-0{,}78,\ 2{,}80)$.

Пример 3.7. Из партии однотипных высокоомных сопротивлений отобрано 10 штук. У каждого из них измерены отклонения сопротивления от номинального значения (табл. 3.2).

Таблица 3.2

Номер изделия	1	2	3	4	5	6	7	8	9	10
Отклонение	1	3	−2	2	4	2	5	3	−2	4

Предполагая, что контролируемый признак имеет нормальный закон распределения, найдем выборочное среднее $\overline{x}$, *исправленную выборочную дисперсию* S^2 и доверительный интервал для дисперсии с коэффициентом доверия $\gamma = 0{,}96$.

Находим выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1+3-2+2+4+2+5+3-2+4}{10} = 2$$

и исправленную выборочную дисперсию

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1+1+16+4+9+1+1+16+4}{9} \approx 5{,}88.$$

Чтобы построить доверительный интервал для дисперсии, воспользуемся *статистикой*

$$\frac{(n-1)S^2(\vec{X}_n)}{\sigma^2} = \frac{n\widehat{\sigma}^2(\vec{X}_n)}{\sigma^2},$$

имеющей *распределение* χ^2 с $n-1$ степенью свободы. По таблице квантилей распределения χ^2 (см. табл. П.3) находим квантили $\chi^2_{\alpha/2}(n-1)$ и $\chi^2_{1-\alpha/2}(n-1)$ уровней $\alpha/2$ и $1-\alpha/2$. В данном случае

$$\alpha = 1 - \gamma = 1 - 0{,}96 = 0{,}04$$

и распределение имеет девять степеней свободы. Следовательно,

$$\chi^2_{\alpha/2}(9) = \chi^2_{0{,}02}(9) = 2{,}09; \qquad \chi^2_{1-\alpha/2}(9) = \chi^2_{0{,}98}(9) = 21{,}07.$$

Для границ доверительного интервала получаем

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} = \frac{5{,}88\cdot 9}{21{,}7} = 2{,}44; \qquad \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} = \frac{5{,}78\cdot 9}{2{,}09} = 24{,}89.$$

Отсюда находим доверительный интервал для дисперсии с коэффициентом доверия 0,96: (2,4, 24,9).

Пример 3.8. Найдем доверительный интервал для вероятности попадания снаряда в цель с коэффициентом доверия $\gamma = 0{,}9$, если после 220 выстрелов в цель попало 75 снарядов.

Используя таблицу квантилей нормального распределения (см. табл. П.2), находим квантиль $u_{1-\alpha/2}$ для $\alpha = 1-\gamma$. Поскольку $\alpha = 1-\gamma = 1-0{,}9 = 0{,}1$, то $u_{1-\alpha/2} = u_{0{,}95} = 1{,}645$. Границы доверительного интервала (см. **3.3**) имеют вид

$$\underline{p} = \frac{m}{n} - \frac{u_{1-\alpha/2}}{\sqrt{n}}\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)} = \frac{75}{220} - \frac{1{,}645}{\sqrt{220}}\sqrt{\frac{75\cdot 145}{220^2}} \approx 0{,}289,$$

$$\overline{p} = \frac{m}{n} + \frac{u_{1-\alpha/2}}{\sqrt{n}}\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)} = \frac{75}{220} + \frac{1{,}645}{\sqrt{220}}\sqrt{\frac{75\cdot 145}{220^2}} \approx 0{,}393.$$

Значит, доверительный интервал для вероятности попадания снаряда в цель следующий: (0,289, 0,393).

Пример 3.9. По выборке $(x_1, \ldots, x_n)$ объема n из генеральной совокупности X, равномерно распределенной на отрезке $[0, \theta]$, построим доверительный интервал для неизвестного параметра θ, если $n = 500$, $\gamma = 0{,}95$ и задан *статистический ряд* (табл. 3.3).

Таблица 3.3

x_i	0,5	1,5	2,5	3,5	4,5	5,5	6,5	7,5	8,5	9,5	10,5	11,5
n_i	41	34	54	39	40	45	41	33	37	41	47	39

Для построения доверительного интервала воспользуемся статистикой

$$T(X_1, \ldots, X_n) = \frac{X_{(n)}}{\theta},$$

где $X_{(n)} = \max\limits_{i=\overline{1,n}} X_i$ — *крайний член вариационного ряда*. Эта статистика имеет распределение

$$F_{X_{(n)}}(x) = \begin{cases} 0, & x < 0; \\ x^n, & 0 \leqslant x \leqslant 1; \\ 1, & x > 1. \end{cases}$$

Для $\alpha = 1 - \gamma$ находим квантили уровня $\alpha/2$ и $1 - \alpha/2$ данного распределения. Поскольку

$$F(t_{\alpha/2}) = t_{\alpha/2}^n = \frac{\alpha}{2}, \qquad F(t_{1-\alpha/2}) = t_{1-\alpha/2}^n = 1 - \frac{\alpha}{2},$$

то

$$t_{\alpha/2} = \left(\frac{\alpha}{2}\right)^{1/n}, \qquad t_{1-\alpha/2} = \left(\frac{2-\alpha}{2}\right)^{1/n}.$$

Таким образом, значение статистики $X_{(n)}/\theta$ с вероятностью γ попадает в интервал с границами $\left(\frac{\alpha}{2}\right)^{1/n}$ и $\left(\frac{2-\alpha}{2}\right)^{1/n}$. Значит, *интервальная оценка* для θ следующая:

$$\left(\left(\frac{2}{2-\alpha}\right)^{1/n} X_{(n)},\ \left(\frac{2}{\alpha}\right)^{1/n} X_{(n)}\right).$$

Согласно условиям примера, $\alpha = 0{,}05$, а статистика $X_{(n)}$ принимает значение $x_{(500)} = 11{,}5$. Поэтому доверительный интервал имеет вид

$$\left(\left(\frac{2}{1{,}95}\right)^{\frac{1}{500}} \cdot 11{,}5,\ \left(\frac{2}{0{,}05}\right)^{\frac{1}{500}} \cdot 11{,}5\right),$$

или (11,5, 11,6).

Пример 3.10. Построим интервальную оценку для разности математических ожиданий двух генеральных совокупностей, распределенных по нормальному закону с параметрами (μ_1, σ) и (μ_2, σ) с неизвестной дисперсией σ по двум случайным независимым выборкам $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$. Предполагая, что $n = m = 5$, $\widehat{\sigma}_1^2 = 3{,}37$, $\widehat{\sigma}_2^2 = 0{,}46$, $\gamma = 0{,}9$, найдем доверительный интервал.

Для построения интервальной оценки воспользуемся статистикой

$$T(\vec{X}_n) = \frac{T_1(\vec{X}_n)}{\sqrt{T_2(\vec{X}_n)}}\sqrt{n+m-2},$$

где

$$T_1(\vec{X}_n)=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}},\qquad T_2(\vec{X}_n)=\frac{n\widehat{\sigma}_1^2}{\sigma^2}+\frac{m\widehat{\sigma}_2^2}{\sigma^2}.$$

Покажем, что статистика $T_1(\vec{X}_n)$ имеет распределение Стьюдента с $m+n-2$ степенями свободы. Для этого достаточно убедиться, что статистика $T_1(\vec{X}_n)$ имеет нормальный закон распределения с параметрами $(0,1)$, а статистика $T_2(\vec{X}_n)$ — распределение χ^2 с $m+n-2$ степенями свободы.

Действительно, статистика $\overline{X}-\overline{Y}$ имеет нормальное распределение с параметрами $\left(\mu_1-\mu_2,\ \sigma^2(1/n+1/m)\right)$, так как *свертка* нормальных *законов распределения* есть нормальный закон распределения [XVI]. Следовательно,

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

имеет нормальный закон распределения с параметрами $(0,1)$. Статистика $T_2(\vec{X}_n)$ есть сумма независимых случайных величин $n\widehat{\sigma}_1^2(\vec{X}_n)/\sigma^2$ и $m\widehat{\sigma}_2^2(\vec{X}_n)/\sigma^2$, имеющих *распределение* χ^2 с $n-1$ и $m-1$ степенями свободы соответственно, т.е. распределение $T_2(\vec{X}_n)$ есть композиция двух χ^2-распределений, а потому имеет χ^2-распределение с числом степеней свободы $(n-1)+(m-1)=n+m-2$.

Для заданного коэффициента доверия γ по таблице квантилей распределения Стьюдента (см. табл. П.4) находим квантиль $t_\beta(n+m-2)$ уровня $\beta=(1+\gamma)/2$. Соотношение

$$\mathbf{P}\left\{\left|\frac{T_1(\vec{X}_n)}{\sqrt{T_2(\vec{X}_n)}}\sqrt{n+m-2}\right|\leqslant t_\beta(n+m-2)\right\}=\gamma$$

означает, что с вероятностью γ выполняется неравенство

$$\left|\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{n\widehat{\sigma}_1^2(\vec{X}_n)+m\widehat{\sigma}_2^2(\vec{X}_n)}}\sqrt{\frac{mn(n+m-2)}{n+m}}\right| \leqslant t_\beta(n+m-2).$$

Отсюда заключаем, что границы интервальной оценки параметра $\mu_1-\mu_2$ имеют вид

$$\overline{X}-\overline{Y}\pm t_\beta(n+m-2)\sqrt{\frac{\big(n\widehat{\sigma}_1^2(\vec{X}_n)+m\widehat{\sigma}_2^2(\vec{X}_n)\big)(m+n)}{mn(m+n-2)}}.$$

Для $\gamma=0{,}9$, $n=m=5$ находим квантиль $t_{0{,}95}(8)=1{,}86$ распределения Стьюдента (см. табл. П.4). В результате, учитывая, что $\widehat{\sigma}_1^2=3{,}37$, $\widehat{\sigma}_2^2=0{,}46$, получаем доверительный интервал вида

$$\left(-0{,}48-1{,}86\sqrt{\frac{5(3{,}37+0{,}46)\cdot 10}{25\cdot 8}},\ -0{,}48+1{,}86\sqrt{\frac{5(3{,}37+0{,}46)\cdot 10}{25\cdot 8}}\right),$$

или $(-2{,}28,\ 1{,}32)$.

Пример 3.11. Пусть $(X_1,\ \ldots,\ X_n)$ и $(Y_1,\ \ldots,\ Y_m)$ — независимые случайные выборки из двух генеральных совокупностей, распределенных по нормальному закону с параметрами (μ_1,σ_1^2) и (μ_2,σ_2^2) соответственно. Построим интервальную оценку для отношения дисперсий σ_2^2/σ_1^2 с доверительной вероятностью γ. Значения границ доверительного интервала найдем при $n=25$, $m=16$, $\widehat{\sigma}_1^2=1{,}44$, $\widehat{\sigma}_2^2=1{,}21$, $\gamma=0{,}9$.

Для построения интервальной оценки воспользуемся статистикой $T=\frac{T_1(m-1)}{T_2(n-1)}$, где статистики $T_1=n\widehat{\sigma}_1^2(\vec{X}_n)/\sigma_1^2$, $T_2=m\widehat{\sigma}_2^2(\vec{X}_n)/\sigma_2^2$ независимы и имеют χ^2-распределения с $n-1$ и $m-1$ степенями свободы соответственно. Следовательно, статистика T имеет *распределение Фишера* со степенями свободы $m-1$ и $n-1$.

По таблице квантилей распределения Фишера (см. табл. П.5) находим квантили $f_{\alpha/2}(n-1,m-1)$ и $f_{1-\alpha/2}(n-1,m-1)$, где

$\alpha = 1 - \gamma$. В силу соотношения

$$\mathbf{P}\left\{f_{\frac{\alpha}{2}}(n-1, m-1) < \frac{T_1(m-1)}{T_2(n-1)} < f_{1-\frac{\alpha}{2}}(n-1, m-1)\right\} = \gamma$$

интервальная оценка имеет вид

$$\left(f_{\frac{\alpha}{2}}(n-1,m-1)\frac{m\widehat{\sigma}_2^2(\vec{X}_n)}{n\widehat{\sigma}_1^2(\vec{X}_n)}\frac{n-1}{m-1},\ f_{1-\frac{\alpha}{2}}(n-1,m-1)\frac{m\widehat{\sigma}_2^2(\vec{X}_n)}{n\widehat{\sigma}_1^2(\vec{X}_n)}\frac{n-1}{m-1}\right).$$

Для заданных значений $\gamma = 0{,}9$, $n = 25$ и $m = 16$ находим $f_{0{,}05}(24, 15) = 0{,}474$ и $f_{0{,}95}(24, 15) = 2{,}29$. Отсюда получаем границы доверительного интервала

$$f_{0{,}05}(24,15)\frac{m\widehat{\sigma}_2^2}{n\widehat{\sigma}_1^2}\frac{n-1}{m-1} = 0{,}474 \cdot \frac{16 \cdot 1{,}21}{25 \cdot 1{,}44} \cdot \frac{24}{15} = 0{,}408$$

и

$$f_{0{,}95}(24,15)\frac{m\widehat{\sigma}_2^2}{n\widehat{\sigma}_1^2}\frac{n-1}{m-1} = 2{,}29 \cdot \frac{16 \cdot 1{,}21}{25 \cdot 1{,}44} \cdot \frac{24}{15} = 1{,}97.$$

Пример 3.12. Предположим, что некоторый элемент испытывается последовательными независимыми циклами. Точное значение вероятности p безотказной работы элемента в каждом цикле неизвестно. Испытания проводятся до первого отказа. Требуется построить доверительный интервал для p в предположении, что первый отказ наблюдался в цикле с номером n.

Рассмотрим случайную величину ν — номер цикла, в котором наблюдался первый отказ. Эта случайная величина имеет ***отрицательное биномиальное распределение***:

$$P(\nu = n) = (1-p)p^{n-1}, \quad n = 1, 2, \ldots$$

Таким образом, задача сводится к построению доверительного интервала для параметра p отрицательного биномиального распределения по значению наблюдаемой случайной величины.

Функция распределения случайной величины ν определяется выражением

$$F_\nu(n,p) = (1-p)(1+p+\ldots+p^{n-2}) = 1-p^{n-1}.$$

Применяя далее общие уравнения (3.9), получаем значения *нижней* и *верхней границ* интервальной оценки для параметра p с коэффициентом доверия $\gamma = 1-\alpha-\beta$:

$$\underline{p} = \sqrt[n-1]{\beta}, \qquad \overline{p} = \sqrt[n]{1-\alpha}.$$

Тем самым нижняя и верхняя границы интервальной оценки для параметра p, найденные из уравнений (3.9), совпадают соответственно с нижней границей в биномиальной схеме испытаний для случая $d=0$ отказов в серии из $n-1$ испытаний и верхней границей для случая $d=1$ отказу в серии из n испытаний, являющихся решением *уравнений Клоппера — Пирсона.*

Пример 3.13. При доверительном оценивании по результатам испытаний показателя надежности (коэффициента готовности) восстанавливаемого элемента возникает следующая задача. Построить интервальную оценку для отношения $\rho = \lambda/\mu$ параметров двух экспоненциальных законов распределений с плотностями

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0; \\ 0, & x < 0, \end{cases} \quad \text{и} \quad g(x) = \begin{cases} \mu e^{-\mu x}, & x \geqslant 0; \\ 0, & x < 0, \end{cases}$$

на основе двух независимых случайных выборок $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ из этих распределений.

Рассмотрим статистику $T = (T_1 m)/(T_2 n)$, где

$$T_1 = 2\lambda \sum_{i=1}^{n} X_i \quad \text{и} \quad T_2 = 2\mu \sum_{j=1}^{m} Y_j.$$

Статистики T_1 и T_2 имеют χ^2-распределения с $2n$ и $2m$ степенями свободы. Отсюда следует, что статистика T является

центральной и имеет распределение Фишера с $2n$ и $2m$ степенями свободы. Применяя общий подход, получаем нижнюю и верхнюю границы доверительного интервала для параметра $\rho = \lambda/\mu$:

$$\underline{\rho}(\vec{x}_n, \vec{y}_m) = f_\alpha(2n, 2m)\frac{n\sum_{j=1}^{m} y_j}{m\sum_{i=1}^{n} x_i}, \quad \overline{\rho}(\vec{x}_n, \vec{y}_m) = f_{1-\beta}(2n, 2m)\frac{n\sum_{j=1}^{m} y_j}{m\sum_{i=1}^{n} x_i}.$$

Дополнение 3.1. Необходимые сведения о некоторых распределениях

Гамма-распределение. Плотность этого распределения

$$p(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0; \\ 0, & x < 0, \end{cases}$$

определяется двумя параметрами $\lambda > 0$ и $\alpha > 0$. Здесь

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt \quad —$$

гамма-функция.

Далее, если случайная величина ξ имеет гамма-распределение с параметрами λ, α, будем использовать сокращенное обозначение $\xi \sim \Gamma(\lambda, \alpha)$.

Теорема 3.1. Если две случайные величины $\xi \sim \Gamma(\lambda, \alpha)$ и $\eta \sim \Gamma(\lambda, \beta)$ независимы, то $\xi + \eta \sim \Gamma(\lambda, \alpha + \beta)$.

◀ В соответствии с известной формулой *свертки плотностей распределения*, плотность распределения $p_{\xi+\eta}(t)$ суммы двух независимых случайных величин ξ и η имеет вид

$$p_{\xi+\eta}(t) = \int_0^t p_\xi(x)\, p_\eta(t-x)\, dx, \quad t > 0,$$

где учтено, что $p_\xi(x)=0$ при $x<0$ и $p_\eta(t-x)=0$ при $x>t$. При $x>0$ имеем

$$p_\xi(x)=C_1x^{\alpha-1}e^{-\lambda x}, \qquad p_\eta(x)=C_2x^{\beta-1}e^{-\lambda x},$$

где C_1, C_2 — нормировочные константы, вычисляемые по формулам

$$C_1=\frac{\lambda^\alpha}{\Gamma(\alpha)}, \qquad C_2=\frac{\lambda^\beta}{\Gamma(\beta)}.$$

После простых преобразований получаем

$$p_{\xi+\eta}(t)=C_1C_2e^{-\lambda t}\int\limits_0^t x^{\alpha-1}(t-x)^{\beta-1}\,dx$$

и после замены $x=ut$ переменного под знаком интеграла приходим к равенству

$$p_{\xi+\eta}(t)=Ct^{\alpha+\beta-1}e^{-\lambda t}, \tag{3.11}$$

где C — нормировочная константа:

$$C=C_1C_2\int\limits_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du.$$

Доказательство утверждения теперь следует непосредственно из (3.11) и формулы Эйлера для бета- и гамма-функций [VI]. ▶

Из теоремы 3.1 легко получить следующее более общее утверждение.

Теорема 3.2. Если случайные величины $\xi_1, \ldots, \xi_n$ независимы, $\xi_i\sim\Gamma(\lambda,\alpha_i)$, $i=\overline{1,n}$, то $\xi_1+\ldots+\xi_n\sim\Gamma(\lambda,\alpha_1+\ldots+\alpha_n)$.

Распределение Релея. Пусть случайная величина ξ имеет нормальное распределение с математическим ожиданием $\mu=0$

и дисперсией σ^2. Тогда случайная величина ξ^2 имеет ***распределение Релея***:

$$p(x) = \begin{cases} \dfrac{1}{\sigma\sqrt{2\pi x}} e^{-\frac{x}{2\sigma^2}}, & x > 0; \\ 0, & x \leqslant 0, \end{cases} \tag{3.12}$$

для которого далее будем использовать сокращенное обозначение $\xi^2 \sim \Gamma(1/2\sigma^2, 1/2)$.

Действительно, для функции распределения случайной величины ξ^2 при $x > 0$ находим

$$F_{\xi^2}(x) = \mathbf{P}\{\xi^2 < x\} = \mathbf{P}\{|\xi| < \sqrt{x}\} = \\ = \mathbf{P}\{-\sqrt{x} < \xi < \sqrt{x}\} = F_\xi(\sqrt{x}) - F_\xi(-\sqrt{x}),$$

где $F_\xi(t) = \Phi(t/\sigma)$ — функция распределения случайной величины ξ, записанная через функцию

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{u} e^{-z^2/2} dz$$

стандартного нормального распределения. Тем самым

$$F_{\xi^2}(x) = \Phi\Big(\frac{\sqrt{x}}{\sigma}\Big) - \Phi\Big(-\frac{\sqrt{x}}{\sigma}\Big) = 2\Phi\Big(\frac{\sqrt{x}}{\sigma}\Big) - 1, \quad x > 0,$$

откуда после дифференцирования получаем формулу (3.12) в случае $x > 0$. Поскольку $\xi^2 \geqslant 0$, имеем $F_{\xi^2}(x) = 0$ при $x \leqslant 0$.

Распределение χ^2. Пусть $\xi_1, \xi_2, \ldots, \xi_m$ — независимые случайные величины, каждая из которых имеет стандартное нормальное распределение. Тогда из теоремы 3.2 и распределения Релея следует, что

$$\xi_1^2 + \xi_2^2 + \ldots + \xi_m^2 \sim \Gamma\Big(\frac{1}{2}, \frac{m}{2}\Big),$$

т.е. плотность распределения случайной величины $\xi_1^2+\xi_2^2+\ldots$ $\ldots+\xi_m^2$ имеет вид

$$p(x)=\begin{cases}\dfrac{1}{2^{m/2}\Gamma(m/2)}x^{m/2-1}e^{-x/2}, & x>0;\\ 0, & x\leqslant 0.\end{cases}$$

Это распределение называют ***распределением*** χ^2 (хи-квадрат) или χ^2***-распределением*** с m степенями свободы. Для случайной величины ξ с χ^2-распределением с m степенями свободы будем использовать сокращенное обозначение $\xi\sim\chi^2(m)$.

Из теоремы 3.2 и установленной связи между распределением χ^2 и гамма-распределением вытекает следующее утверждение.

Следствие 3.1. Если случайные величины ξ_1, ..., ξ_n независимы, $\xi_i\sim\chi^2(m_i)$, $i=\overline{1,n}$, то сумма этих случайных величин также имеет распределение χ^2:

$$\xi_1+\xi_2+\ldots+\xi_n\sim\chi^2(m_1+m_2+\ldots+m_n).$$

Экспоненциальное распределение. Частным случаем гамма-распределения при $\alpha=1$ является ***экспоненциальное распределение***, его плотность имеет вид

$$p(x)=\begin{cases}\lambda e^{-\lambda x}, & x\geqslant 0;\\ 0, & x<0.\end{cases}\tag{3.13}$$

Экспоненциальное распределение часто используется в математической теории надежности, теории массового обслуживания и других приложениях.

Распределение Эрланга. Пусть $\xi_1,\xi_2,\ldots,\xi_n$ — независимые случайные величины, каждая из которых имеет экспоненциальный закон распределения (3.13). Из теоремы 3.2 следует, что сумма этих случайных величин имеет гамма-распределение:

$$\xi_1+\xi_2+\ldots+\xi_n\sim\Gamma(\lambda,n)\tag{3.14}$$

с плотностью

$$p(x)=\begin{cases}\dfrac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x}, & x\geqslant 0;\\ 0, & x<0,\end{cases}$$

которое называют ***распределением Эрланга*** порядка n.

Замечание 3.1. Исходя из (3.14) нетрудно показать, что случайная величина $\lambda(\xi_1+\ldots+\xi_n)$, где $\xi_1,\ldots,\xi_n$ — независимые случайные величины, каждая из которых имеет экспоненциальный закон распределения 3.13, также имеет гамма-распределение:

$$\lambda(\xi_1+\ldots+\xi_n)\sim\Gamma(1,n).$$

Учитывая указанную выше связь между гамма-распределением и распределением χ^2, можно показать, что случайная величина $2\lambda(\xi_1+\ldots+\xi_n)$ имеет распределение χ^2 с $2n$ степенями свободы:

$$2\lambda(\xi_1+\ldots+\xi_n)\sim\chi^2(2n).$$

Этот факт используют, в частности, при построении доверительных интервалов для параметра λ экспоненциального распределения.

О распределении статистики Стьюдента. При построении *доверительных интервалов* для параметров нормального распределения использовалась статистика

$$T=\frac{\overline{X}-\mu}{S(\vec{X}_n)}\sqrt{n-1},$$

где $\overline{X}$ — *выборочное среднее*, а $S^2(\vec{X}_n)$ — исправленная оценка дисперсии. Покажем, что эта статистика имеет *распределение Стьюдента* с $n-1$ степенями свободы.

Заметим, что выборочное среднее $\overline{X}$ имеет нормальное распределение $N(\mu,\sigma^2/n)$. Отсюда следует, что случайная

величина

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

имеет стандартное нормальное распределение. В то же время, случайная величина

$$V = \frac{(n-1)S^2(\vec{X}_n)}{\sigma^2}$$

имеет распределение χ^2 с $n-1$ степенями свободы, причем случайные величины Z, V независимы*. Статистика T далее может быть представлена в виде

$$T = \frac{Z}{\sqrt{V}}\sqrt{n-1},$$

откуда с учетом определения распределения Стьюдента следует, что статистика T имеет распределение Стьюдента с $n-1$ степенями свободы.

Распределение Фишера. Пусть случайные величины ξ, η независимы и имеют распределение χ^2 с n и m степенями свободы соответственно, т.е. $\xi \sim \chi^2(n)$, $\eta \sim \chi^2(m)$. Тогда случайная величина $\varphi = \frac{m\xi}{n\eta}$ имеет плотность распределения

$$p(x) = \begin{cases} C\dfrac{x^{\frac{n}{2}-1}}{\left(1+\dfrac{nx}{m}\right)^{\frac{n+m}{2}}}, & x \geqslant 0; \\ 0, & x < 0, \end{cases}$$

где C — нормировочная константа, равная

$$C = \frac{\left(\dfrac{n}{m}\right)^{n/2}}{\mathrm{B}\left(\dfrac{n}{2}, \dfrac{m}{2}\right)}.$$

*См., например, *Рао С.Р.*

Это распределение называют ***распределением Фишера*** со степенями свободы n и m.

Бета-распределение. Плотность ***бета-распределения*** (***распределения бета***) с параметрами α, β имеет вид

$$p(x) = \begin{cases} Cx^{\alpha-1}(1-x)^{\beta-1}, & x \in [0,1]; \\ 0, & x \notin [0,1], \end{cases}$$

где C — нормировочная константа, равная $C = 1/\mathrm{B}(\alpha,\beta)$, а $\mathrm{B}(\alpha,\beta)$ — бета-функция. Соответствующая функция распределения при $x \geqslant 0$ имеет вид

$$F(x) = \frac{\mathrm{B}_x(\alpha,\beta)}{\mathrm{B}(\alpha,\beta)},$$

где

$$\mathrm{B}_x(\alpha,\beta) = \int_0^x u^{\alpha-1}(1-u)^{\beta-1}du$$

неполная бета-функция.

Заметим, что для неполной бета-функции справедливо следующее известное равенство:

$$\mathrm{B}_x(m+1, n-m) = 1 - \sum_{j=0}^{m} C_n^j x^j (1-x)^{n-j},$$

где m, n — целые числа. Это равенство, в частности, используется при построении стандартных доверительных границ Клоппера — Пирсона для параметра p биномиального распределения.

Частным случаем бета-распределения при $\alpha = \beta = 1$ является равномерное распределение на отрезке $[0, 1]$.

Вопросы и задачи

3.1. Что называют интервальной оценкой для неизвестного параметра распределения генеральной совокупности?

3.2. Что такое коэффициент доверия (доверительная вероятность), нижняя и верхняя границы интервальной оценки неизвестного параметра?

3.3. Какую статистику называют центральной?

3.4. Какую статистику используют при построении интервальной оценки для параметра экспоненциального распределения?

3.5. Какую статистику используют для построения интервальной оценки для математического ожидания в случае нормальной модели при известной дисперсии? По какому закону статистика распределена?

3.6. Какую статистику используют для построения интервальной оценки для математического ожидания в случае нормальной модели при неизвестной дисперсии? По какому закону статистика распределена?

3.7. Какую статистику используют для построения интервальной оценки для дисперсии нормально распределенной генеральной совокупности? По какому закону она распределена?

3.8. На чем основан метод построения приближенной интервальной оценки для неизвестного параметра генеральной совокупности?

3.9. Какую статистику используют при построении приближенной интервальной оценки: а) для параметра биномиального распределения, б) для математического ожидания случайной величины?

3.10. В чем состоит метод доверительных множеств?

3.11. Что называют γ-зоной для параметра θ?

3.12. Запишите уравнения Клоппера — Пирсона. Как их используют при построении интервальной оценки параметра биномиального распределения?

3.13. Постоянная величина измерена 25 раз с помощью прибора, систематическая ошибка которого равна нулю, а случайные ошибки измерения распределены по нормальному закону со средним квадратичным отклонением $\sigma = 10$ м. Определите значения границ доверительного интервала для измеряемой величины при коэффициенте доверия 0,99, если $\overline{x} = 100$ м.

Ответ: значение нижней границы 94,9 м, верхней — 105,1 м.

3.14. Оценка измеряемой величины определяется формулой

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Результаты отдельных измерений не содержат систематической ошибки и подчинены нормальному закону распределения со средним квадратичным отклонением $\sigma = 2{,}1$. Определите интервальные оценки J_n с доверительной вероятностью 0,9 для значения измеряемой величины при различных объемах случайной выборки $\vec{X}_n$: а) $n = 5$, б) $n = 10$, в) $n = 25$.

Ответ: а) $(\overline{X} - 1{,}55,\ \overline{X} + 1{,}55)$; б) $(\overline{X} - 1{,}09,\ \overline{X} + 1{,}09)$; в) $(\overline{X} - 0{,}69,\ \overline{X} + 0{,}69)$.

3.15. Средняя квадратичная ошибка высотомера $\sigma = 15$ м. Сколько надо иметь таких приборов на самолете, чтобы с достоверностью 0,99 ошибка измерения средней высоты $\overline{X}$ была меньше 30 м? При этом случайные ошибки распределены по нормальному закону, а систематические ошибки отсутствуют.

Ответ: на самолете должно быть не менее двух высотомеров.

3.16. На основании 100 опытов было определено, что в среднем для производства детали требуется $\overline{t} = 5{,}5$ с, а $\widehat{\sigma}_t = 1{,}7$ с. Сделав допущение, что время для производства детали распределено по нормальному закону, определите доверительный

интервал для математического ожидания производства детали с коэффициентом доверия 0,85.

Ответ: (5,25, 5,75).

3.17. По результатам измерений 100 резисторов, случайно отобранных из большой партии однотипных изделий, получена оценка сопротивления $\overline{x} = 10$ кОм. Найдите:

а) вероятность того, что для резисторов всей партии значения сопротивления лежат в пределах $(10 \pm 0{,}1)$ кОм (среднее квадратичное отклонение измерения известно: $\sigma = 1$ кОм);

б) количество измерений, при котором с вероятностью 0,95 можно утверждать, что для всей партии резисторов значения сопротивления лежат в пределах $(10 \pm 0{,}1)$ кОм.

Ответ: а) 0,68; б) $n \geqslant 385$.

3.18. Провели 5 независимых равноточных измерений для определения заряда электрона; получили следующие результаты (в абсолютных электростатических единицах): $4{,}781 \cdot 10^{-10}$; $4{,}792 \cdot 10^{-10}$; $4{,}795 \cdot 10^{-10}$; $4{,}779 \cdot 10^{-10}$; $4{,}769 \cdot 10^{-10}$. Определите значение оценки величины заряда электрона и найти доверительный интервал при коэффициенте доверия 99 %, считая, что ошибки распределены по нормальному закону и измерения не имеют систематических ошибок.

Ответ: $\overline{x} = 4{,}783 \cdot 10^{-10}$; $(4{,}761 \cdot 10^{-10},\ 4{,}805 \cdot 10^{-10})$.

3.19. На контрольных испытаниях 16 осветительных ламп были определены значения оценок математического ожидания и среднего квадратичного отклонения их срока службы, которые оказались равными $\overline{x} = 3000$ ч и $\widehat{\sigma} = 20$ ч соответственно. Считая, что контролируемый признак (срок службы лампы) имеет нормальный закон распределения, определите:

а) доверительный интервал для математического ожидания при доверительной вероятности 0,9;

б) вероятность, с которой можно утверждать, что абсолютная величина ошибки определения m не превысит 10 ч.

Ответ: а) (2991,2, 3008,8); б) 0,93.

3.20. Провели 40 измерений базы длиной L. По результатам опыта получены значения оценок измеряемой величины и среднего квадратичного отклонения: $\overline{x} = 10400$ м и $\widehat{\sigma}_x = 85$ м. Ошибки измерения подчиняются нормальному закону распределения. Найдите вероятность того, что интервал со случайными границами $(0{,}999\overline{X},\ 1{,}001\overline{X})$ накроет неизвестный параметр L.

Ответ: 0,55.

3.21. Из партии валов отобрали $n_1 = 9$ шт. Значения выборочного среднего диаметра вала $\overline{x}_1 = 30$ мм, выборочной дисперсии $\widehat{\sigma}_1^2 = 9\,\text{мм}^2$. Затем осуществили повторный эксперимент, отобрав $n_2 = 16$ шт. и получили значения выборочных оценок $\overline{x}_2 = 29$ мм, $\sigma_2^2 = 4{,}5\,\text{мм}^2$. Используя объединенные выборочные оценки, найдите 99 %-ный доверительный интервал для среднего.

Ответ: (27,98, 30,74).

3.22. По результатам 10 измерений емкости конденсатора прибором, не имеющим систематической ошибки, получили следующие отклонения от номинального значения (пФ):

5,4; −13,9; −11; 7,2; −15,6; 29,2; 1,4; −0,3; 6,6; −9,9.

Найдите 90 %-ный доверительный интервал для дисперсии и среднего квадратичного отклонения, предполагая, что генеральная совокупность имеет нормальное распределение.

Ответ: (96,81, 49,34); (9,84, 22,17).

3.23. По 15 независимым равноточным измерениям были рассчитаны значения оценок математического ожидания и среднего квадратичного отклонения максимальной скорости самолета $\overline{v} = 424{,}7$ м/с и $\widehat{\sigma}_v = 7{,}7$ м/с. Считая, что генеральная совокупность имеет нормальное распределение, определите: а) доверительный интервал для среднего квадратичного отклонения при доверительной вероятности 0,9; б) вероятность того, что абсолютная величина случайной ошибки при определении σ_v по 15 измерениям не превзойдет 2 м/с.

Ответ: а) (6,69, 12,7); б) 0,76.

3.24. Известно, что измерительный прибор не имеет систематических ошибок, а случайные ошибки измерения подчиняются нормальному закону распределения. Сколько надо провести измерений для определения оценки среднего квадратичного отклонения прибора, чтобы с доверительной вероятностью 70 % абсолютная величина ошибки определения этой величины была не более 20 % от $\widehat{\sigma}(\vec{X}_n)$?

Ответ: не менее 15 измерений.

3.25. При проверке 100 деталей из большой партии обнаружено 10 бракованных. Найдите 95 %-ный доверительный интервал для доли бракованных деталей во всей партии.

Ответ: (0,055, 0,174).

3.26. Из большой партии транзисторов одного типа были случайным образом отобраны и проверены 100 шт. Коэффициент усиления 36 транзисторов оказался меньше 10. Найдите 95 %-ный доверительный интервал для доли таких транзисторов во всей партии.

Ответ: (0,266, 0,454).

3.27. С автоматической линии, производящей подшипники, было отобрано 100 шт., причем 10 оказались бракованными. Найдите: а) 90 %-ный доверительный интервал для вероятности того, что произвольно выбранный подшипник окажется бракованным; б) количество подшипников, которые надо проверить, чтобы с вероятностью 0,9973 можно было утверждать, что доля брака отличается от частоты не более чем на 5%.

Ответ: а) (0,12, 0,38); б) $n \geqslant 88$.

3.28. В 10000 сеансах игры с автоматом выигрыш появился 4000 раз. Найдите: а) 95%-ный доверительный интервал для вероятности выигрыша; б) количество сеансов игры, которые следует провести, чтобы с вероятностью 0,99 можно было утверждать, что вероятность p выигрыша отличается от его частоты не более чем на 1 %.

Ответ: а) (0,39, 0,41); б) $n \geqslant 16231$.

3.29. Для экспоненциального распределения со „сдвигом", имеющего плотность

$$f(x;\theta)=\begin{cases} e^{-(x-\theta)}, & x\geqslant\theta;\\ 0, & x<\theta, \end{cases}$$

по выборке объема n постройте интервальную оценку параметра θ с доверительной вероятностью γ.

Указание: В качестве исходной рассмотрите статистику $T=X_{(1)}-\theta$, имеющую функцию распределения

$$F(x)=\begin{cases} 1-e^{-nx}, & x\geqslant 0;\\ 0, & x<0. \end{cases}$$

Ответ: $\left(X_{(1)}+\frac{\ln\beta}{n},\ X_{(1)}+\frac{\ln(1-\alpha)}{n}\right)$, где $\alpha>0$ и $\beta>0$ связаны равенством $1-\alpha-\beta=\gamma$.

3.30. Постройте интервальную оценку для разности $\mu_1-\mu_2$ математических ожиданий двух генеральных совокупностей, распределенных по нормальным законам с параметрами (μ_1,σ_1^2) и (μ_2,σ_2^2) по результатам независимых выборок $(X_1,\ldots,X_n)$ и $(Y_1,\ldots,Y_m)$ в предположении, что дисперсии σ_1^2, σ_2^2 известны.

Указание: В качестве исходной следует взять статистику

$$T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}$$

и убедиться в том, что эта статистика имеет стандартный нормальный закон распределения с параметрами $(0,1)$.

Ответ: $\left(\overline{X}-\overline{Y}-u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}},\ \overline{X}-\overline{Y}+u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right)$.

4. ПРОВЕРКА ГИПОТЕЗ. ПАРАМЕТРИЧЕСКИЕ МОДЕЛИ

В этой главе рассмотрен второй класс задач математической статистики (см. **1.2**), связанных с проверкой *статистических гипотез.*

Выше (см. **2, 3**) были рассмотрены задачи оценивания неизвестного параметра θ по *реализации случайной выборки* из *генеральной совокупности* случайной величины X, закон распределения которой зависит от θ. При этом мы не располагали никакой *априорной информацией* относительно параметра θ.

При проверке статистической гипотезы о параметре θ исследователь заранее на основании той или иной априорной информации выдвигает предположение (гипотезу) о величине θ, например $\theta = \theta_0$, где θ_0 — некоторое заданное значение параметра. После этого он проводит эксперимент, в результате которого получает реализацию $\vec{x}_n$ случайной выборки $\vec{X}_n$ из генеральной совокупности X, распределение которой зависит от параметра θ. По этим данным ему нужно дать ответ на вопрос: согласуется гипотеза $\theta = \theta_0$ с результатами эксперимента или нет? Другими словами, исследователю нужно решить, можно ли принять выдвинутую гипотезу или ее нужно отклонить как противоречащую результатам эксперимента и принять некоторую альтернативную гипотезу (например, $\theta \neq \theta_0$).

4.1. Основные понятия

Пусть имеется *выборка* $\vec{x}_n$, являющаяся *реализацией случайной выборки* $\vec{X}_n$ из *генеральной совокупности* X, плотность распределения которой $p(t;\theta)$ зависит от неизвестного параметра θ.

Статистические гипотезы относительно неизвестного истинного значения параметра θ называют ***параметрическими гипотезами***. При этом если θ — скаляр, то речь идет об ***однопараметрических гипотезах***, а если вектор, — то о ***многопараметрических гипотезах***.

Статистическую гипотезу H называют ***простой***, если она имеет вид

$$H\colon\ \vec{\theta} = \vec{\theta}_0,$$

где $\vec{\theta}_0$ — некоторое заданное значение параметра.

Статистическую гипотезу называют ***сложной***, если она имеет вид

$$H\colon\ \vec{\theta} \in D,$$

где D — некоторое множество значений параметра θ, состоящее более чем из одного элемента.

Пример 4.1. Предположим, проводится серия из n независимых испытаний по схеме Бернулли с неизвестным параметром p, где p — вероятность „успеха“ в одном испытании. Тогда гипотеза $H\colon p = 1/2$ является простой. Примерами сложных гипотез являются следующие: $H_1\colon p \geqslant 1/2$; $H_2\colon p \leqslant 1/2$; $H_3\colon 1/4 \leqslant p \leqslant 3/4$ и т.д.

Пример 4.2. Пусть $\vec{X}_n$ — *случайная выборка объема* n из генеральной совокупности X, распределенной по нормальному закону с неизвестным математическим ожиданием μ и известной дисперсией σ^2. Тогда гипотеза $H\colon \mu = \mu_0$, где μ_0 — некоторое заданное значение параметра μ, является простой.

Гипотезы $H_1\colon \mu \geqslant \mu_0$; $H_2\colon \mu \leqslant \mu_0$; $H\colon \mu_0 \leqslant \mu \leqslant \mu_1$ являются сложными.

Пример 4.3. Пусть в примере 4.2 оба параметра μ и σ неизвестны. В этом случае гипотеза $H\colon \mu = \mu_0$ становится сложной, так как ей соответствует множество значений двумерного вектора $\vec{\theta} = (\mu,\ \sigma)$, для которых $\mu = \mu_0$, $0 < \sigma < \infty$.

4.2. Проверка двух простых гипотез

Рассмотрим сначала случай, когда проверяются две *простые статистические гипотезы* вида

$$H_0\colon\ \theta=\theta_0, \qquad H_1\colon\ \theta=\theta_1,$$

где θ_0, θ_1 — два заданных (различных) значения параметра. Первую ***гипотезу*** H_0 обычно называют ***основной***, а вторую H_1 — ***альтернативной***, или ***конкурирующей***, ***гипотезой***, хотя эта терминология является достаточно условной. Так, например, одна и та же гипотеза может в одних задачах выступать в качестве основной, а в других — в качестве альтернативной. По данным *выборки* $\vec{x}_n$ необходимо принять решение о справедливости одной из указанных гипотез.

Критерием, или ***статистическим критерием***, проверки гипотез называют правило, по которому по данным выборки $\vec{x}_n$ принимается решение о справедливости либо первой, либо второй гипотезы.

Критерий задают с помощью ***критического множества*** W, являющегося подмножеством *выборочного пространства* $\mathcal{X}_n$ *случайной выборки* $\vec{X}_n$. Решение принимают следующим образом:

1) если выборка $\vec{x}_n$ принадлежит критическому множеству W, то отвергают основную гипотезу H_0 и принимают альтернативную гипотезу H_1;

2) если выборка $\vec{x}_n$ не принадлежит критическому множеству W (т.е. принадлежит дополнению $\overline{W}$ множества W до выборочного пространства $\mathcal{X}_n$), то отвергают альтернативную гипотезу H_1 и принимают основную гипотезу H_0.

При использовании любого критерия возможны ошибки следующих видов:

1) принять гипотезу H_1, когда верна H_0 — ***ошибка первого рода***;

2) принять гипотезу H_0, когда верна H_1 — ***ошибка второго рода***.

Вероятности совершения ошибок первого и второго рода обозначают α и β:

$$\alpha = \mathbf{P}\{\vec{X}_n \in W \mid H_0\}, \qquad \beta = \mathbf{P}\{\vec{X}_n \in \overline{W} \mid H_1\},$$

где $\mathbf{P}\{A \mid H_j\}$ — вероятность события A при условии, что справедлива гипотеза H_j, $j = 0, 1$. Указанные вероятности вычисляют с использованием функции плотности распределения случайной выборки $\vec{X}_n$:

$$\alpha = \int\limits_W \dots \int \prod_{k=1}^{n} p(t_k;\theta_0)\, dt_1 \dots dt_n,$$

$$\beta = \int\limits_{\overline{W}} \dots \int \prod_{k=1}^{n} p(t_k;\theta_1)\, dt_1 \dots dt_n.$$

Вероятность совершения ошибки первого рода α называют также ***уровнем значимости критерия***.

Величину $1 - \beta$, равную вероятности отвергнуть основную гипотезу H_0, когда она неверна, называют ***мощностью критерия***.

4.3. Критерий Неймана — Пирсона

При построении *критерия* для проверки *статистических гипотез*, как правило, исходят из необходимости максимизации его *мощности* $1 - \beta$ (минимизации вероятности совершения *ошибки второго рода*) при фиксированном *уровне значимости* α *критерия* (вероятности совершения *ошибки первого рода*). Для упрощения дальнейших рассуждений будем считать, что $\vec{X}_n$ — *случайная выборка объема* n из *генеральной совокупности* непрерывной случайной величины X, плотность распределения вероятностей которой $p(t;\theta)$ зависит от неизвестного

параметра θ, и рассмотрим две *простые гипотезы* H_0: $\theta = \theta_0$ и H_1: $\theta = \theta_1$.

Введем функцию случайной выборки $\vec{X}_n$:

$$\varphi(\vec{X}_n) = \frac{L(\vec{X}_n;\theta_1)}{L(\vec{X}_n;\theta_0)}, \qquad L(\vec{X}_n;\theta) = \prod_{i=1}^{n} p(X_i;\theta).$$

Статистика $\varphi(\vec{X}_n)$ представляет собой отношение *функций правдоподобия* при истинности *альтернативной* и *основной гипотез* соответственно. Ее называют ***отношением правдоподобия***. Для построения ***оптимального**** (наиболее мощного) при заданном уровне значимости α ***критерия Неймана — Пирсона*** в *критическое множество* W включают те элементы $\vec{x}_n$ *выборочного пространства* $\mathcal{X}_n$ случайной выборки $\vec{X}_n$, для которых выполняется неравенство

$$\varphi(\vec{x}_n) \geqslant C_\varphi,$$

где константу C_φ выбирают из условия

$$\mathbf{P}\{\varphi(\vec{X}_n) \geqslant C_\varphi \mid H_0\} = \alpha,$$

которое обеспечивает заданное значение уровня значимости α и может быть записано в виде

$$\int\!\!\dots\!\!\int_{\varphi(t_1,\dots,t_n)\geqslant C_\varphi} L(t_1,\dots,t_n;\theta_0)\,dt_1\dots dt_n = \alpha.$$

При этом вероятность ошибки второго рода не может быть уменьшена при данном значении вероятности ошибки первого рода α.

Рассмотрим примеры построения оптимального критерия Неймана — Пирсона при проверке простых гипотез относительно параметров основных, наиболее часто используемых распределений.

*См.: *Леман Э.*

Пример 4.4. Построение оптимального критерия Неймана — Пирсона для параметра μ нормального закона распределения с известной дисперсией σ^2 проведем для случая двух простых гипотез

$$H_0\colon\ \mu=\mu_0, \qquad H_1\colon\ \mu=\mu_1,$$

где μ_0 и μ_1 — некоторые заданные значения, связанные неравенством $\mu_0<\mu_1$.

В рассматриваемом случае функция правдоподобия имеет вид

$$L(X_1,\ldots,X_n;\mu)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\mu)^2\right),$$

а отношение правдоподобия —

$$\varphi(\vec{X}_n)=\frac{L(X_1,\ldots,X_n;\mu_1)}{L(X_1,\ldots,X_n;\mu_0)}=$$

$$=\exp\left(\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n X_i\right)\exp\left(-\frac{n(\mu_1-\mu_0)^2}{2\sigma^2}\right).$$

В данном случае неравенство

$$\varphi(\vec{x}_n)=\exp\left(\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n x_i\right)\exp\left(-\frac{n(\mu_1-\mu_0)^2}{2\sigma^2}\right)\geqslant C_\varphi$$

равносильно неравенству

$$\sum_{i=1}^n x_i\geqslant C, \tag{4.1}$$

где константу C выбирают из условия обеспечения заданного уровня значимости α:

$$\mathbf{P}\left\{\sum_{i=1}^n X_i\geqslant C\,\middle|\,\mu=\mu_0\right\}=\alpha. \tag{4.2}$$

Действительно,

$$\ln\left(\exp\left(\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n x_i\right)\exp\left(-\frac{n(\mu_1-\mu_0)^2}{2\sigma^2}\right)\right)=$$

$$=\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n x_i-\frac{n(\mu_1-\mu_0)^2}{2\sigma^2}\geqslant \ln C_\varphi,$$

откуда следует, что

$$\sum_{i=1}^n x_i\geqslant\frac{\sigma^2}{\mu_1-\mu_0}\left(\ln C_\varphi-\frac{n(\mu_1-\mu_0)^2}{2\sigma^2}\right)=C.$$

Случайная величина $X_1+\ldots+X_n$ имеет нормальное распределение с математическим ожиданием $n\mu$ и дисперсией $n\sigma^2$ (см. **1.2**). Поэтому условие (4.2) можно записать в виде

$$1-\Phi\left(\frac{C-n\mu_0}{\sigma\sqrt{n}}\right)=\alpha, \tag{4.3}$$

или

$$\frac{C-n\mu_0}{\sigma\sqrt{n}}=u_{1-\alpha}.$$

Таким образом, константа C, задающая критическую область в (4.1), определяется равенством

$$C=n\mu_0+u_{1-\alpha}\sigma\sqrt{n}. \tag{4.4}$$

При этом вероятность совершения ошибки второго рода

$$\beta=\mathbf{P}\left\{\sum_{i=1}^n X_i<C\,\middle|\,\mu=\mu_1\right\}=\Phi\left(\frac{C-n\mu_1}{\sigma\sqrt{n}}\right) \tag{4.5}$$

является минимально возможной при данном значении α.

Пример 4.5. Если в условиях примера 4.4 неравенство $\mu_0 < \mu_1$ заменить неравенством $\mu_1 < \mu_0$, то в этом случае критическое множество W задается неравенством

$$\sum_{i=1}^{n} x_i \leqslant C,$$

где константу C выбирают из условия

$$\mathbf{P}\Big\{\sum_{i=1}^{n} X_i \leqslant C \,\Big|\, \mu = \mu_0\Big\} = \alpha.$$

Таким образом,

$$\Phi\left(\frac{C - n\mu_0}{\sigma\sqrt{n}}\right) = \alpha$$

или, что то же самое,

$$\frac{C - n\mu_0}{\sigma\sqrt{n}} = u_\alpha = -u_{1-\alpha}.$$

Из последнего равенства находим $C = n\mu_0 - u_{1-\alpha}\sigma\sqrt{n}$.

Пример 4.6. Построение оптимального критерия Неймана — Пирсона в случае экспоненциального распределения с параметром λ проведем для двух простых гипотез

$$H_0\colon\ \lambda = \lambda_0, \qquad H_1\colon\ \lambda = \lambda_1,$$

где $\lambda_0 < \lambda_1$. В этом случае функция правдоподобия

$$L(X_1, \ldots, X_n; \lambda) = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} X_i\Big).$$

Таким образом,

$$\varphi(\vec{X}_n) = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\Big(-(\lambda_1 - \lambda_0)\sum_{i=1}^{n} X_i\Big).$$

Отсюда видно, что критическое множество можно задать неравенством

$$\sum_{i=1}^{n} x_i \leqslant C,$$

где константа C выбрана из условия обеспечения заданного уровня значимости α:

$$\mathbf{P}\Big\{\sum_{i=1}^{n} X_i \leqslant C \mid \lambda = \lambda_0\Big\} = \alpha.$$

Случайная величина $2\lambda(X_1 + \ldots + X_n)$ при $\lambda = \lambda_0$ имеет *χ^2-распределение* с $2n$ степенями свободы (см. Д.3.1). Исходя из этого, получаем выражение для константы C:

$$C = \frac{\chi^2_\alpha(2n)}{2}\lambda_0,$$

где $\chi^2_\alpha(2n)$ — квантиль уровня α для χ^2-распределения с $2n$ степенями свободы. При этом вероятность совершения ошибки второго рода равна

$$\beta = \mathbf{P}\Big\{\sum_{i=1}^{n} X_i > C \mid \lambda = \lambda_1\Big\} = \\ = 1 - H_{2n}(2\lambda_1 C) = 1 - H_{2n}\Big(\chi^2_\alpha(2n)\frac{\lambda_1}{\lambda_0}\Big),$$

где $H_{2n}(t)$ — функция распределения случайной величины, имеющей χ^2-распределение с $2n$ степенями свободы.

Пример 4.7. Построение оптимального критерия Неймана — Пирсона для параметра биномиального распределения проведем для случая двух простых гипотез

$$H_0\colon\ p = p_0, \qquad H_1\colon\ p = p_1,$$

где p — вероятность „успеха“ в одном испытании при реализации схемы независимых испытаний Бернулли, а p_0 и p_1 —

заданные значения параметра, удовлетворяющие неравенству $p_0 < p_1$.

Пусть объем испытаний достаточно велик и X_j — результат j-го испытания. Случайная величина X_j принимает значения 0 и 1 с вероятностями $1-p$ и p соответственно. Функция правдоподобия в этом случае имеет вид

$$L(X_1,\ldots,X_n;p) = C_n^{K(\vec{X}_n)} p^{K(\vec{X}_n)} (1-p)^{n-K(\vec{X}_n)},$$

где $K(\vec{X}_n) = X_1 + \ldots + X_n$ — общее число „успехов“ в серии из n испытаний. Отношение правдоподобия определяется равенством

$$\varphi(\vec{X}_n) = \frac{L(X_1,\ldots,X_n;p_1)}{L(X_1,\ldots,X_n;p_0)} = \left(\frac{p_1}{p_0}\right)^{K(\vec{X}_n)} \left(\frac{1-p_0}{1-p_1}\right)^{n-K(\vec{X}_n)}.$$

Значит, критическое множество для оптимального критерия Неймана — Пирсона в данном случае имеет вид

$$K(\vec{x}_n) = \sum_{i=1}^{n} x_i \geqslant C. \tag{4.6}$$

Константу C выбирают исходя из условия

$$\mathbf{P}\{X_1 + \ldots + X_n \geqslant C \mid p = p_0\} = \alpha.$$

Распределение случайной величины $K(\vec{X}_n)$ при достаточно больших n в соответствии с известной интегральной теоремой Муавра — Лапласа имеет асимптотически нормальное распределение с математическим ожиданием $\mu = np$ и дисперсией $\sigma^2 = np(1-p)$. Используя указанное распределение, выберем константу C в (4.6) из условия обеспечения заданного уровня значимости α, т.е. из условия

$$\mathbf{P}\{K(\vec{X}_n) \geqslant C \mid p = p_0\} \approx 1 - \Phi\left(\frac{C - np_0}{\sqrt{np_0(1-p_0)}}\right) = \alpha, \tag{4.7}$$

откуда, используя квантиль $u_{1-\alpha}$ стандартного нормального закона, получаем

$$C = np_0 + u_{1-\alpha}\sqrt{np_0(1-p_0)}.$$

При этом вероятность ошибки второго рода равна

$$\beta = \mathbf{P}\{K(\vec{X}_n) < C \mid p = p_1\} \approx \Phi\left(\frac{C - np_1}{\sqrt{np_1(1-p_1)}}\right) =$$

$$= \Phi\left(\frac{n(p_0 - p_1) + u_{1-\alpha}\sqrt{np_0(1-p_0)}}{\sqrt{np_1(1-p_1)}}\right). \quad (4.8)$$

4.4. Определение объема выборки

Выше (см. 4.3) при построении *оптимального критерия Неймана — Пирсона* с заданным *уровнем значимости* α предполагалось, что *объем* n *случайной выборки* $\vec{X}_n$ известен и фиксирован. Но возможной является ситуация, когда возникает необходимость в определении (заранее, до проведения наблюдений) такого объема n^* случайной выборки, при котором может быть построен *критерий* для проверки двух *простых гипотез* H_0: $\theta = \theta_0$ и H_1: $\theta = \theta_1$ с заданными или меньшими значениями вероятностей α и β совершения *ошибок первого* и *второго рода* соответственно.

В рассматриваемой ситуации величину n^* определяют как минимальное целое значение n, для которого система неравенств

$$\begin{cases} \mathbf{P}\{\varphi(X_1, \ldots, X_n) \geqslant C_\varphi \mid \theta = \theta_0\} \leqslant \alpha, \\ \mathbf{P}\{\varphi(X_1, \ldots, X_n) < C_\varphi \mid \theta = \theta_1\} \leqslant \beta \end{cases} \quad (4.9)$$

может быть выполнена при некотором значении константы $C = C^*$. При этом соответствующий оптимальный критерий Неймана — Пирсона, обеспечивающий заданные значения α, β

будет иметь *критическое множество*, определяемое неравенством

$$\varphi(x_1,\ldots,x_n) \geqslant C_\varphi^*.$$

Пример 4.8. Определим объем выборки для случая нормальной модели.

Для ситуации, рассмотренной в примере 4.4, из выражений (4.3), (4.5) получаем, что система неравенств (4.9) в этом случае имеет вид

$$1-\Phi\left(\frac{C-n\mu_0}{\sigma\sqrt{n}}\right)\leqslant\alpha,\qquad \Phi\left(\frac{C-n\mu_1}{\sigma\sqrt{n}}\right)\leqslant\beta.$$

Следовательно, для обеспечения заданных значений α, β вероятностей совершения ошибок первого и второго рода минимально необходимый объем n^* выборки и соответствующую константу C^* можно определить из системы уравнений

$$1-\Phi\left(\frac{C-n\mu_0}{\sigma\sqrt{n}}\right)=\alpha,\qquad \Phi\left(\frac{C-n\mu_1}{\sigma\sqrt{n}}\right)=\beta.$$

Используя квантили стандартного нормального распределения, запишем эти уравнения в виде

$$\frac{C-n\mu_0}{\sigma\sqrt{n}}=u_{1-\alpha},\qquad \frac{C-n\mu_1}{\sigma\sqrt{n}}=u_\beta=-u_{1-\beta}. \tag{4.10}$$

Исключая из уравнений константу C, находим необходимый объем выборки

$$n^*=\frac{\sigma^2\left(u_{1-\alpha}+u_{1-\beta}\right)^2}{\left(\mu_1-\mu_0\right)^2}. \tag{4.11}$$

Пусть, например, требуется проверить гипотезы

$$H_0\colon\ \mu=\mu_0=3{,}5,\qquad H_1\colon\ \mu=\mu_1=3{,}8$$

при $\sigma=0{,}8$ и заданных значениях вероятностей $\alpha=0{,}05$, $\beta=0{,}1$. Применяя формулу (4.11) и учитывая, что $u_{1-\alpha}=u_{0{,}95}=1{,}64$,

$u_{1-\beta} = u_{0,9} = 1{,}28$, получаем необходимый в этом случае объем выборки $n^* = 61$.

Пример 4.9. Определим объем выборки для схемы испытаний Бернулли.

Для задачи проверки гипотез, рассмотренной в примере 4.7, вновь используем возможность аппроксимации биномиального распределения нормальным распределением с параметрами $\mu = np$ и $\sigma^2 = np(1-p)$. После этого, согласно (4.7), (4.8), приходим к системе уравнений для определения n^* и C^*:

$$1 - \Phi\left(\frac{C - np_0}{\sqrt{np_0(1-p_0)}}\right) = \alpha, \qquad \Phi\left(\frac{C - np_1}{\sqrt{np_1(1-p_1)}}\right) = \beta,$$

которая может быть представлена в следующем виде:

$$\frac{C - np_0}{\sqrt{np_0(1-p_0)}} = u_{1-\alpha}, \qquad \frac{C - np_1}{\sqrt{np_1(1-p_1)}} = u_\beta = -u_{1-\beta}.$$

Решая эту систему, находим

$$n^* = \frac{\left(u_{1-\alpha}\sqrt{p_0(1-p_0)} + u_{1-\beta}\sqrt{p_1(1-p_1)}\right)^2}{(p_1 - p_0)^2}. \tag{4.12}$$

Равенство (4.12) (приближенно) определяет минимально необходимый объем выборки, позволяющий обеспечить заданные значения вероятностей совершения ошибок первого и второго рода при проверке простых гипотез вида $H_0 : p = p_0$, $H_1 : p = p_1$ в схеме Бернулли. Поскольку в (4.12) величина n^* не обязательно целая, то на практике в качестве объема выборки берут наименьшее целое число, большее или равное n^*. #

Минимально необходимый объем наблюдений, определенный с использованием оптимального критерия Неймана — Пирсона, не может быть улучшен (уменьшен) в ситуации, когда объем выборки фиксируется и задается заранее, до наблюдений. Тем не менее средний объем наблюдений может быть

уменьшен при тех же значениях вероятностей совершения ошибок первого и второго рода в последовательной схеме наблюдений, когда решение об остановке наблюдений принимается по ходу процесса наблюдений, в зависимости от получаемых данных (см. 4.6).

4.5. Сложные параметрические гипотезы

Предположим, что требуется проверить две *сложные гипотезы*

$$H_0\colon\ \theta \in \Theta_0, \qquad H_1\colon\ \theta \in \Theta_1, \tag{4.13}$$

где Θ_0, Θ_1 — некоторые непересекающиеся области значений параметра θ. Например, области Θ_0, Θ_1 могут быть заданы неравенствами $\theta \leqslant \theta_0$ и $\theta \geqslant \theta_1$, где θ_0 и θ_1 — некоторые фиксированные значения параметра, удовлетворяющие неравенству $\theta_0 < \theta_1$.

Критерий проверки сложных гипотез (4.13) по-прежнему задается с помощью *критического множества* W *реализаций случайной выборки* $\vec{X}_n$, на основе которого решение принимают следующим образом:

– если реализация $\vec{x}_n$ случайной выборки $\vec{X}_n$ принадлежит критическому множеству W, тогда *основную гипотезу* H_0 отвергают и принимают *альтернативную гипотезу* H_1;

– если реализация $\vec{x}_n$ случайной выборки $\vec{X}_n$ не принадлежит критическому множеству W, тогда отвергают альтернативную гипотезу H_1 и принимают основную гипотезу H_0.

Вероятности совершения *ошибок первого* и *второго рода* в случае сложных гипотез имеют прежний смысл и определяются выражениями

$$\alpha(\theta) = \mathbf{P}\{(X_1, \ldots, X_n) \in W \mid \theta\}, \quad \theta \in \Theta_0;$$

$$\beta(\theta) = \mathbf{P}\{(X_1, \ldots, X_n) \in \overline{W} \mid \theta\}, \quad \theta \in \Theta_1.$$

В отличие от случая *простых гипотез*, величины $\alpha(\theta)$, $\beta(\theta)$ являются некоторыми функциями от параметра θ.

Максимально возможное значение вероятности совершения ошибки первого рода

$$\alpha = \max_{\theta \in \Theta_0} \alpha(\theta)$$

называют ***размером критерия.***

Функцию

$$M(\theta) = \mathbf{P}\{(X_1, \ldots, X_n) \in W \mid \theta\},$$

определяющую значение вероятности отклонения основной гипотезы H_0 в зависимости от истинного значения параметра θ, называют ***функцией мощности критерия.*** Если существует критерий, который при данном фиксированном размере α максимизирует функцию мощности $M(\theta)$ по всем возможным критериям одновременно при всех θ из множества Θ_1, то такой ***критерий*** называют ***равномерно наиболее мощным.*** Равномерно наиболее мощные критерии существуют лишь в некоторых частных случаях при проверке гипотез относительно одномерных параметров (см. примеры 4.10–4.12).

Вероятности совершения ошибок первого и второго рода связаны с функцией мощности следующими соотношениями:

$$\alpha(\theta) = M(\theta), \quad \theta \in \Theta_0; \tag{4.14}$$

$$\beta(\theta) = 1 - M(\theta), \quad \theta \in \Theta_1. \tag{4.15}$$

Тем самым равномерно наиболее мощный критерий, если он существует, минимизирует вероятность совершения ошибки второго рода $\beta(\theta)$ (при фиксированном размере α) одновременно при всех $\theta \in \Theta_1$.

Замечание 4.1. Формально равенства (4.14), (4.15) справедливы при всех возможных значениях θ, но при значениях θ, отличных от указанных в (4.14), (4.15), величины $\alpha(\theta)$, $\beta(\theta)$

теряют свой смысл — вероятностей совершения соответствующих ошибок. #

Иногда наряду с функцией мощности используется также ***оперативная характеристика критерия***

$$s(\theta) = \mathbf{P}\{(X_1, \ldots, X_n) \in \overline{W} \mid \theta\},$$

представляющая собой вероятность принятия основной гипотезы H_0 при условии, что истинное значение параметра равно θ. Нетрудно увидеть, что оперативная характеристика и функция мощности связаны соотношением $s(\theta) = 1 - M(\theta)$.

Построение критериев для проверки сложных параметрических гипотез проиллюстрируем далее для случая нормальной модели.

Пример 4.10. Рассмотрим проверку простой гипотезы H_0: $\mu = \mu_0$ против сложной гипотезы H_1: $\mu > \mu_0$ относительно параметра — среднего μ нормального распределения при известной дисперсии σ^2.

При любом $\mu_1 > \mu_0$ критическая область оптимального наиболее мощного критерия Неймана — Пирсона размера α для простых гипотез $\mu = \mu_0$ против $\mu = \mu_1$ имеет вид (4.1), где константу C выбирают из условия (4.2) или (4.3). Поэтому она не зависит от μ_1. Это означает, что построенный уже выше для указанных простых гипотез критерий с критическим множеством, задаваемым неравенством (4.1)

$$\sum_{i=1}^{n} x_i \geqslant C = n\mu_0 + u_{1-\alpha}\sigma\sqrt{n}, \tag{4.16}$$

является равномерно наиболее мощным критерием размера α для данной задачи со сложной альтернативной гипотезой H_1: $\mu > \mu_0$.

Пример 4.11. В условиях предыдущего примера рассмотрим проверку простой гипотезы H_0: $\mu = \mu_0$ против сложной гипотезы H_1: $\mu < \mu_0$.

В этом случае, используя результаты, полученные при рассмотрении примера 4.5, приходим к выводу, что равномерно наиболее мощный критерий размера α для данной задачи задается критическим множеством, определяемым неравенством

$$\sum_{i=1}^{n} x_i \leqslant C = n\mu_0 - u_{1-\alpha}\sigma\sqrt{n}.$$

Пример 4.12. В условиях примера 4.10 рассмотрим проверку двух сложных гипотез вида

$$H_0\colon\ \mu \leqslant \mu_0, \qquad H_1\colon\ \mu \geqslant \mu_1, \tag{4.17}$$

где $\mu_0 < \mu_1$.

Заметим, что для критерия с критическим множеством (4.16) вероятность совершения ошибки первого рода

$$\alpha(\mu) = \mathbf{P}\Big\{\sum_{i=1}^{n} X_i \geqslant C \,\Big|\, \mu\Big\} = 1 - \Phi\Big(u_{1-\alpha} + (\mu_0 - \mu)\frac{\sqrt{n}}{\sigma}\Big)$$

есть возрастающая функция переменного μ. Тем самым максимальное значение вероятности совершения ошибки первого рода, определяемое как

$$\alpha = \max_{\mu \leqslant \mu_0} \alpha(\mu),$$

достигается в точке $\mu = \mu_0$, откуда следует, что данный критерий, применяемый к сложным гипотезам (4.17), имеет размер $\alpha = \alpha(\mu_0)$.

Рассуждая далее так же, как в примере 4.10, получаем, что указанный критерий с критической областью (4.16) является равномерно наиболее мощным критерием для данной задачи со сложными гипотезами.

Пример 4.13. Рассмотрим проверку гипотез относительно параметра нормального распределения μ следующего вида:

$$H_0\colon\ \mu = \mu_0, \qquad H_1\colon\ \mu \neq \mu_0$$

(по-прежнему предполагаем, что дисперсия σ^2 известна).

В этом случае основная гипотеза H_0 является простой, а альтернативная гипотеза H_1 является сложной. При $\mu = \mu_0$ рассмотрим *статистику*

$$\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n},$$

которая имеет *стандартное нормальное распределение*. Критическое множество для проверки указанных гипотез H_0, H_1 определим следующим образом:

$$\frac{|\overline{x} - \mu_0|}{\sigma}\sqrt{n} \geqslant u_{1-\alpha/2}.$$

Соответствующий критерий по построению имеет вероятность совершения ошибки первого рода α.

Пример 4.14. Рассмотрим проверку двух сложных гипотез

$$H_0\colon\ \mu = \mu_0, \qquad H_1\colon\ \mu > \mu_0 \tag{4.18}$$

относительно параметра μ нормального закона распределения в случае, когда дисперсия σ^2 неизвестна.

В отличие от примера 4.10 гипотеза H_0 также является сложной. При $\mu = \mu_0$ статистика

$$\frac{\overline{X} - \mu_0}{S(\vec{X}_n)}\sqrt{n} \tag{4.19}$$

имеет *распределение Стьюдента* с $n-1$ *степенями свободы* (см. **Д.3.1**). Исходя из этого получаем, что критерий с уровнем значимости α для гипотез (4.18) задается критическим множеством

$$\frac{\overline{x} - \mu_0}{S(\vec{x}_n)}\sqrt{n} \geqslant t_{1-\alpha}(n-1),$$

где $t_{1-\alpha}(n-1)$ — квантиль уровня $1-\alpha$ распределения Стьюдента с $n-1$ степенями свободы.

Аналогично на основе статистики (4.19) строят критерий для проверки сложных гипотез

$$H_0\colon\ \mu=\mu_0, \qquad H_1\colon\ \mu<\mu_0 \tag{4.20}$$

или

$$H_0\colon\ \mu=\mu_0, \qquad H_1\colon\ \mu\neq\mu_0. \tag{4.21}$$

Для гипотез (4.20) критерий размера α задается критическим множеством, определяемым неравенством

$$\frac{\overline{x}-\mu_0}{S(\vec{x}_n)}\sqrt{n}\leqslant -t_{1-\alpha}(n-1).$$

Для гипотез вида (4.21) критерий размера α задают критическим множеством, определяемым неравенством

$$\frac{|\overline{x}-\mu_0|}{S(\vec{x}_n)}\sqrt{n}\geqslant t_{1-\alpha/2}(n-1).$$

Пример 4.15. Рассмотрим проверку гипотез о равенстве математических ожиданий для двух различных нормальных распределений.

Пусть определены две случайные выборки $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ объемов n и m из *генеральных совокупностей* независимых случайных величин $X\sim N(\mu_1,\sigma_1^2)$ и $Y\sim N(\mu_2,\sigma_2^2)$ соответственно. Рассмотрим следующие задачи проверки сложных гипотез относительно параметров μ_1, μ_2 в случае, когда дисперсии σ_1^2, σ_2^2 известны:

$$H_0\colon\ \mu_1=\mu_2, \qquad H_1\colon\ \mu_1>\mu_2; \tag{4.22}$$

$$H_0\colon\ \mu_1=\mu_2, \qquad H_1\colon\ \mu_1<\mu_2; \tag{4.23}$$

$$H_0\colon\ \mu_1=\mu_2, \qquad H_1\colon\ \mu_1\neq\mu_2. \tag{4.24}$$

Разность *выборочных средних* $\overline{X}-\overline{Y}$ имеет нормальное распределение с математическим ожиданием $\mu_1-\mu_2$ и дисперсией

$\sigma_1^2/n + \sigma_2^2/m$. Отсюда следует, что при справедливости основной гипотезы, т.е. при $\mu_1 = \mu_2$, статистика

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}} \tag{4.25}$$

имеет стандартное нормальное распределение. Исходя из этого, заключаем, что критерии размера α для указанных задач задаются критическими множествами

$$\frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}} \geqslant u_{1-\alpha}; \quad \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}} \leqslant -u_{1-\alpha}; \quad \frac{|\overline{x} - \overline{y}|}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}} \geqslant u_{1-\alpha/2}. \quad \#$$

Рассмотрим также задачу проверки гипотез (4.22)–(4.23) о равенстве средних двух нормальных распределений в предположении, что их дисперсии не известны, но равны между собой: $\sigma_1 = \sigma_2 = \sigma$. Обозначим через

$$S_1^2(\vec{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S_2^2(\vec{Y}_m) = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2$$

соответствующие исправленные оценки дисперсии. Статистики $(n-1)S_1^2(\vec{X}_n)/\sigma^2$ и $(m-1)S_2^2(\vec{Y}_m)/\sigma^2$ имеют *χ^2-распределения* с $n-1$ и $m-1$ степенями свободы. Тем самым статистика

$$\frac{(n-1)S_1^2(\vec{X}_n)}{\sigma^2} + \frac{(m-1)S_2^2(\vec{Y}_m)}{\sigma^2}$$

имеет также χ^2-распределение с $n+m-2$ степенями свободы (см. **Д.3.1**). Учитывая, что случайная величина (4.25) при $\mu_1 = \mu_2$ имеет стандартное нормальное распределение, получаем, что статистика

$$\widetilde{T}(\vec{X}_n, \vec{Y}_m) = \frac{(\overline{X} - \overline{Y})\sqrt{n+m-2}}{\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}\sqrt{(n-1)S_1^2(\vec{X}_n) + (m-1)S_2^2(\vec{Y}_m)}}$$

имеет распределение Стьюдента с $n+m-2$ степенями свободы (см. **Д.3.1**). Поэтому критерии размера α для проверки гипотез (4.22)–(4.23) задаются с помощью критических множеств, определяемых следующими неравенствами:

$$\widetilde{T}(\vec{x}_n, \vec{y}_m) \geqslant t_{1-\alpha}(n+m-2),$$
$$\widetilde{T}(\vec{x}_n, \vec{y}_m) \leqslant -t_{1-\alpha}(n+m-2),$$
$$\left|\widetilde{T}(\vec{x}_n, \vec{y}_m)\right| \geqslant t_{1-\alpha/2}(n+m-2).$$

4.6. Последовательный критерий отношения правдоподобия

Во многих случаях на практике наблюдения проводят последовательно. При этом статистическая информация поступает не один раз, а последовательными порциями данных. Предположим, что наблюдается последовательность независимых одинаково распределенных непрерывных случайных величин $X_1, \ldots, X_n, \ldots$, каждая из которых имеет плотность распределения $p(x;\theta)$, где θ — некоторый параметр, значение которого неизвестно. На основе результатов наблюдений $x_1, \ldots, x_n, \ldots$ нужно проверить две *простые гипотезы* H_0: $\theta = \theta_0$ и H_1: $\theta = \theta_1$, где θ_0, θ_1 — некоторые заданные значения параметра.

В рассматриваемой ситуации количество наблюдаемых случайных величин (и тем самым *объем выборки*) не фиксируется заранее, а определяется по ходу наблюдений, в зависимости от получаемых данных. ***Последовательный критерий отношения правдоподобия*** (***критерий Вальда****) строят следующим образом. На очередном n-м шаге наблюдений, исходя из полученных результатов наблюдений $x_1, \ldots, x_n$, вычисляют

$$\varphi_n(x_1, \ldots, x_n) = \frac{p(x_1;\theta_1)\, p(x_2;\theta_1) \ldots p(x_n;\theta_1)}{p(x_1;\theta_0)\, p(x_2;\theta_0) \ldots p(x_n;\theta_0)}.$$

*См.: *Вальд А.*

Эта величина имеет смысл значения *отношения правдоподобия* для гипотез H_0 и H_1 на n-м шаге наблюдений. На каждом n-м шаге проверяют следующие два неравенства:

$$B < \varphi_n(x_1, \ldots, x_n) < A, \tag{4.26}$$

где B и A — некоторые заданные константы, удовлетворяющие условию $0 < B < 1 < A$. Если оба неравенства (4.26) выполняются, то наблюдения продолжают, т.е. осуществляют наблюдение следующей случайной величины X_{n+1}. Другими словами, неравенства (4.26) задают „область продолжения наблюдений" для критерия Вальда.

Наблюдения прекращают при первом нарушении хотя бы одного неравенства (4.26). При нарушении левого неравенства принимают гипотезу H_0. При нарушении правого неравенства принимают гипотезу H_1. Таким образом, номер ν шага, на котором прекращают наблюдения для критерия Вальда, определяют из равенства

$$\nu = \min\{n\colon \varphi_n(x_1, \ldots, x_n) \notin (B, A)\}. \tag{4.27}$$

Вектор результатов наблюдений для любого последовательного критерия, и в том числе для критерия Вальда, имеет вид $(\nu, x_1, \ldots, x_\nu)$, где ν — номер шага, на котором прекращены наблюдения, $x_1, \ldots, x_\nu$ — совокупность всех результатов наблюдений.

Для критерия Вальда правило принятия решения по результатам испытаний $x_1, \ldots, x_n$ имеет следующий вид:

– если $\varphi_\nu(x_1, \ldots, x_\nu) \leqslant B$, то принять гипотезу H_0;

– если $\varphi_\nu(x_1, \ldots, x_\nu) \geqslant A$, то принять гипотезу H_1.

Вероятности совершения *ошибок первого и второго рода* (***риски первого и второго рода***) для этого критерия равны соответственно

$$\alpha = \mathrm{P}\{\varphi_\nu(X_1, \ldots, X_\nu) \geqslant A \mid H_0\}, \quad \beta = \mathrm{P}\{\varphi_\nu(X_1, \ldots, X_\nu) \leqslant B \mid H_1\}.$$

Значения рисков α, β для критерия Вальда могут быть приближенно оценены с помощью следующих известных соотношений*:

$$\begin{cases} \dfrac{\beta}{1-\alpha} = \mathrm{M}\big(\varphi_\nu(X_1,\ldots,X_\nu)\,|\,H_0\big), \\ \dfrac{1-\beta}{\alpha} = \mathrm{M}\big(\varphi_\nu(X_1,\ldots,X_\nu)\,|\,H_1\big), \end{cases} \tag{4.28}$$

где $\mathrm{M}\big(\varphi_\nu(X_1,\ldots,X_\nu)\,|\,H_j\big)$ — условное математическое ожидание случайной величины $\varphi_\nu(X_1,\ldots,X_n)$ при условии, что на шаге ν (шаг, на котором прекращаются наблюдения) по результатам наблюдений $x_1,\ldots,x_\nu$ принято решение о справедливости гипотезы H_j, $j=0,1$.

Из формул (4.28) можно получить соответствующие неравенства и приближенные оценки для значений рисков α, β критерия Вальда. Действительно, для критерия Вальда справедливы неравенства:

а) $\varphi_\nu(x_1,\ldots,x_\nu) \leqslant B$, если принята гипотеза H_0;

б) $\varphi_\nu(x_1,\ldots,x_\nu) \geqslant A$, если принята гипотеза H_1.

Поскольку для условных математических ожиданий справедливы аналогичные неравенства

$$\mathrm{M}\big(\varphi_\nu(X_1,\ldots,X_\nu)\,|\,H_0\big) \leqslant B, \qquad \mathrm{M}\big(\varphi_\nu(X_1,\ldots,X_\nu)\,|\,H_1\big) \geqslant A,$$

то с учетом (4.28) получаем известные неравенства для точных значений рисков α, β критерия Вальда

$$\frac{\beta}{1-\alpha} \leqslant B, \qquad \frac{1-\beta}{\alpha} \geqslant A. \tag{4.29}$$

Множество точек плоскости (α,β), координаты которых удовлетворяют неравенствам (4.29), показано на рис. 4.1 штриховкой.

*См.: *Вальд А.*

Из неравенств (4.29) следуют более грубые неравенства

$$\beta \leqslant B, \qquad \alpha \leqslant \frac{1}{A}, \tag{4.30}$$

которые также иногда используют при оценке рисков α, β.

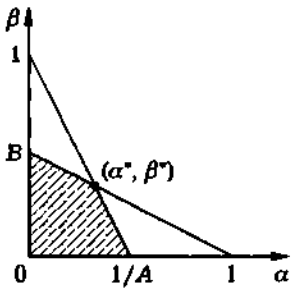


Рис. 4.1

Заметим, что для критерия Вальда наблюдения прекращают на шаге $\nu = n$, на котором впервые происходит выход значения $\varphi_n(x_1, \ldots, x_n)$ из интервала (B, A), или, другими словами, „перескок" значения $\varphi_n(x_1, \ldots, x_n)$ через уровень A (снизу вверх) или через уровень B (сверху вниз). Пренебрежем указанным „перескоком", т.е. будем считать, что на шаге прекращения наблюдений выполняется одно из двух приближенных равенств

$$\varphi_\nu(x_1, \ldots, x_\nu) \approx B, \text{ если принята гипотеза } H_0,$$

$$\varphi_\nu(x_1, \ldots, x_\nu) \approx A, \text{ если принята гипотеза } H_1.$$

Тогда из точных равенств (4.28) получим известные приближенные (с точностью до указанного „перескока") ***равенства Вальда****:

$$\frac{\beta}{1-\alpha} \approx B, \qquad \frac{1-\beta}{\alpha} \approx A. \tag{4.31}$$

Эти приближенные равенства часто используют на практике для оценки значений рисков α, β. Поэтому, согласно (4.31), будем считать, что точные значения рисков α, β удовлетворяют приближенным равенствам

$$\alpha \approx \alpha^*, \qquad \beta \approx \beta^*,$$

где приближенные значения рисков α^*, β^* находятся из равенств

$$\frac{\beta^*}{1-\alpha^*} = B, \qquad \frac{1-\beta^*}{\alpha^*} = A. \tag{4.32}$$

*См.: *Вальд А.*

Таким образом,

$$\alpha^* = \frac{1-B}{A-B}, \qquad \beta^* = \frac{b(A-1)}{A-B},$$

точка (α^*, β^*) в плоскости (α, β) находится на пересечении прямых линий $\beta = B - B\alpha$ и $\beta = 1 - A\alpha$ (см. рис. 4.1).

Как следует из рис. 4.1, точные значения рисков α, β критерия Вальда (они находятся внутри заштрихованной области) всегда удовлетворяют неравенствам

$$\alpha + \beta \leqslant \alpha^* + \beta^*,$$

где α^*, β^* — указанные выше приближенные значения рисков. Кроме того, неравенства (4.30) с учетом (4.32) могут быть записаны в виде

$$\beta \leqslant \frac{\beta^*}{1-\alpha^*}, \qquad \alpha \leqslant \frac{\alpha^*}{1-\beta^*}.$$

Средний объем испытаний. Рассмотрим также вычисление среднего объема испытаний для критерия Вальда. В соответствии с (4.27) номер шага ν прекращения наблюдений может быть представлен в виде дискретной случайной величины:

$$\nu = \min\{n\colon Z_n \notin (-b, a)\},$$

где $a = \ln A > 0$, $b = -\ln B > 0$, Z_n — логарифм отношения функций правдоподобия на n-м шаге:

$$Z_n = \ln \varphi_n(X_1, \ldots, X_n) = \ln \frac{p(X_1;\theta_1)\, p(X_2;\theta_1) \ldots p(X_n;\theta_1)}{p(X_1;\theta_0)\, p(X_2;\theta_0) \ldots p(X_n;\theta_0)}. \quad (4.33)$$

Введем случайные величины

$$Z(X) = \ln \frac{p(X;\theta_1)}{p(X;\theta_0)}, \quad (4.34)$$

и

$$Z(X_n) = \ln \frac{p(X_n;\theta_1)}{p(X_n;\theta_0)}, \quad n = 1, 2, \ldots$$

В этом случае, согласно равенствам (4.32)–(4.34), случайная величина Z_n представляет собой сумму n независимых одинаково распределенных случайных величин $Z(X_1), \ldots, Z(X_n)$, т.е.

$$Z_n = Z(X_1) + \ldots + Z(X_n). \tag{4.35}$$

Обозначим через $\mathrm{M}_0\nu$ математическое ожидание объема испытаний (номера шага, на котором прекращаются наблюдения), если справедлива гипотеза H_0: $\theta = \theta_0$.

В соответствии с (4.35) справедливо равенство

$$\mathrm{M}_0 Z_\nu = \mathrm{M}_0(Z(X_1) + \ldots + Z(X_\nu)), \tag{4.36}$$

где M_0 — математическое ожидание при $\theta = \theta_0$. Для математического ожидания суммы случайного числа независимых одинаково распределенных случайных величин*

$$\mathrm{M}_0\big(Z(X_1) + \ldots + Z(X_\nu)\big) = \mathrm{M}_0\nu\,\mathrm{M}_0 Z. \tag{4.37}$$

Для левой части равенства (4.36) имеем

$$\mathrm{M}_0 Z_\nu = \mathrm{P}_0\{H_0\}\,\mathrm{M}_0(Z_\nu \mid H_0) + \mathrm{P}_0\{H_1\}\,\mathrm{M}_0(Z_\nu \mid H_1), \tag{4.38}$$

где $\mathrm{P}_0\{H_j\}$ — вероятность принятия гипотезы H_j, $j = 0, 1$, при условии, что истинной является гипотеза H_0. Таким образом, согласно (4.32)–(4.38) и определению ошибки первого рода,

$$\mathrm{M}_0 Z_\nu = (1-\alpha)\,\mathrm{M}_0\big(\ln \varphi_\nu(X_1, \ldots, X_\nu) \mid H_0\big) + \\ + \alpha\,\mathrm{M}_0\big(\ln \varphi_\nu(X_1, \ldots, X_\nu) \mid H_1\big). \tag{4.39}$$

*См.: *Вальд А.*, а также: *Ширяев А.Н.*

Из равенства (4.39), пренебрегая снова указанным выше „перескоком“ и используя приближенные равенства (4.31), получаем

$$\mathrm{M}_0 Z_\nu \approx (1-\alpha)\ln\frac{\beta}{1-\alpha} + \alpha\ln\frac{1-\beta}{\alpha}. \tag{4.40}$$

В результате из (4.37) и (4.40) получаем приближенное значение среднего объема испытаний при $\theta = \theta_0$:

$$\mathrm{M}_0\nu \approx \frac{w(\alpha,\beta)}{\mathrm{M}_0(-Z)}, \tag{4.41}$$

где

$$w(\alpha,\beta) = (1-\alpha)\ln\frac{1-\alpha}{\beta} + \alpha\ln\frac{\alpha}{1-\beta}.$$

Аналогично можно получить приближенное значение среднего объема $\mathrm{M}_1\nu$ испытаний при справедливости гипотезы H_1, т.е. при $\theta = \theta_1$:

$$\mathrm{M}_1\nu \approx \frac{w(\beta,\alpha)}{\mathrm{M}_1(Z)}. \tag{4.42}$$

С учетом (4.34) формулы для среднего объема испытаний могут быть также представлены в следующем виде:

$$\mathrm{M}_0\nu \approx \frac{w(\alpha,\beta)}{\rho(\theta_0,\theta_1)}, \qquad \mathrm{M}_1\nu \approx \frac{w(\beta,\alpha)}{\rho(\theta_1,\theta_0)}, \tag{4.43}$$

где

$$\rho(\theta_1,\theta_0) = \mathrm{M}_1 Z = \mathrm{M}_1\ln\frac{p(x;\theta_1)}{p(x;\theta_0)}. \tag{4.44}$$

Формулы (4.41)–(4.43) для среднего объема испытаний являются приближенными с учетом „перескока“.

Нижняя граница для среднего объема испытаний. Из равенства (4.39) далее нетрудно получить также нижнюю границу среднего объема испытаний при данных фиксированных значениях рисков α, β. Учитывая, что функция $\ln u$ выпукла вверх [II], и применяя неравенство Йенсена [XVIII] для математического ожидания от выпуклой функции, получаем

$$\mathrm{M}_0\big(\ln \varphi_\nu(X_1,\ldots,X_\nu)\,|\,H_j\big) \leqslant$$

$$\leqslant \ln \mathrm{M}_0\big(\varphi_\nu(X_1,\ldots,X_\nu)\,|\,H_j\big), \quad j=0,1. \quad (4.45)$$

Из неравенства (4.45) с учетом (4.39), (4.28) находим

$$\mathrm{M}_0 Z_\nu \leqslant (1-\alpha)\ln\frac{\beta}{1-\alpha}+\alpha\ln\frac{1-\beta}{\alpha},$$

откуда с учетом (4.36), (4.37) получаем неравенство

$$\mathrm{M}_0\nu \geqslant \frac{w(\alpha,\beta)}{\mathrm{M}_0(-Z)}. \quad (4.46)$$

Аналогично можно получить неравенство для среднего объема испытаний при $\theta=\theta_1$:

$$\mathrm{M}_1\nu \geqslant \frac{w(\beta,\alpha)}{\mathrm{M}_1 Z}. \quad (4.47)$$

Неравенства (4.46), (4.47) определяют нижние границы для ***среднего объема испытаний*** при $\theta=\theta_0$ и $\theta=\theta_1$ при заданных значениях рисков α, β. Как следует из приближенных равенств (4.41), (4.42), средний объем испытаний для критерия Вальда достигает нижней границы, указанной в (4.46), (4.47), по крайней мере приближенно (с учетом указанного выше „перескока“).

Пример 4.16. Обратимся к биномиальной схеме испытаний и рассмотрим последовательность независимых случайных величин

$$\delta_1, \quad \delta_2, \quad \ldots, \quad \delta_n, \quad \ldots, \quad (4.48)$$

имеющих биномиальное распределение. Пусть δ_n — индикатор отказа элемента на n-м шаге (исход n-го испытания), принимающий значения 0 или 1 с вероятностями $\mathbf{P}\{\delta_n = 0\} = 1 - q$ и $\mathbf{P}\{\delta_n = 1\} = q$, где q — вероятность отказа. Требуется по результатам наблюдений проверить гипотезы

$$H_0\colon\ q = q_0, \qquad H_1\colon\ q = q_1,$$

где q_0 и q_1 — заданные критические уровни показателя q, удовлетворяющие условию $q_0 < q_1$.

В данном случае параметром является $\theta = q$, наблюдаемой на n-м шаге случайной величиной — $X_n = \delta_n$, а закон распределения имеет вид $p(\delta; q) = q^{\delta}(1-q)^{1-\delta}$, где δ принимает два значения 0 и 1. Отношение правдоподобия на n-м шаге испытаний имеет вид

$$\varphi_n(\delta_1, \ldots, \delta_n) = \frac{p(\delta_1; q_1)\, p(\delta_2; q_1) \ldots p(\delta_n; q_1)}{p(\delta_1; q_0)\, p(\delta_2; q_0) \ldots p(\delta_n; q_0)} = \\ = \left(\frac{q_1}{q_0}\right)^{D_n} \left(\frac{1-q_1}{1-q_0}\right)^{n-D_n},$$

где случайная величина $D_n = \delta_1 + \delta_2 + \ldots + \delta_n$ есть суммарное число отказов за n шагов. Пусть d_n — значение случайной величины D_n. Область продолжения испытаний (4.26) для критерия Вальда задается неравенствами

$$B < \left(\frac{q_1}{q_0}\right)^{d_n} \left(\frac{1-q_1}{1-q_0}\right)^{n-d_n} < A,$$

которые после простых преобразований сводятся к неравенствам

$$C_1 n - C_2 b < d_n < C_1 n + C_2 a, \tag{4.49}$$

где $a = \ln A > 0$, $b = -\ln B > 0$,

$$C_2 = \frac{1}{\ln \dfrac{q_1(1-q_0)}{q_0(1-q_1)}}, \qquad C_1 = C_2 \ln \frac{1-q_0}{1-q_1}.$$

Испытания продолжаются, если выполняются оба неравенства (4.49) и прекращаются на том шаге $\nu = n$, на котором впервые нарушается хотя бы одно из этих неравенств. При нарушении левого неравенства принимается решение о справедливости гипотезы H_0: $q = q_0$. При нарушении правого неравенства принимается решение о справедливости гипотезы H_1: $q = q_1$. Таким образом, границы области прекращения наблюдений имеют вид прямых линий на плоскости (n, d_n) (рис. 4.2), которые определяются уравнениями:

а) $d_n = C_1 n - C_2 b$ — граница „области принятия" гипотезы H_0;

б) $d_n = C_1 n + C_2 a$ — граница „области принятия" гипотезы H_1.

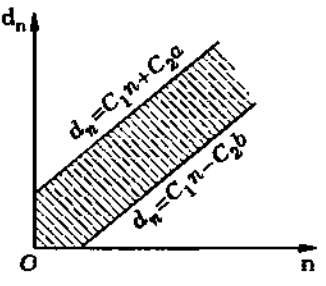


Рис. 4.2

В биномиальной схеме испытаний случайная величина (4.34) имеет вид

$$Z(\delta) = \ln \frac{p(\delta; q_1)}{p(\delta; q_0)} = \delta \ln \frac{q_1}{q_0} + (1 - \delta) \ln \frac{1 - q_1}{1 - q_0},$$

откуда, учитывая, что $\mathrm{M}_0\,\delta = q_0$, $\mathrm{M}_1\,\delta = q_1$, получаем формулы (4.41)–(4.43) для среднего объема испытаний при $q = q_0$ и $q = q_1$:

$$\mathrm{M}_0\,\nu \approx \frac{w(\alpha, \beta)}{\rho(q_0, q_1)}, \qquad \mathrm{M}_1\,\nu \approx \frac{w(\beta, \alpha)}{\rho(q_1, q_0)}, \tag{4.50}$$

где

$$\rho(q_0, q_1) = q_0 \ln \frac{q_0}{q_1} + (1 - q_0) \ln \frac{1 - q_0}{1 - q_1}.$$

Пример 4.17. В эксперименте наблюдают последовательность независимых случайных величин

$$X_1, \quad X_2, \quad \ldots, \quad X_n, \quad \ldots, \tag{4.51}$$

имеющих *экспоненциальный закон распределения* с параметром λ. Требуется на основе наблюдений проверить следующие гипотезы относительно параметра λ:

$$H_0\colon\ \lambda=\lambda_0, \qquad H_1\colon\ \lambda=\lambda_1, \tag{4.52}$$

где $\lambda_0<\lambda_1$.

В этом случае отношение правдоподобия на n-м шаге испытаний имеет вид

$$\varphi_n(X_1,\dots,X_n)=\frac{p(X_1;\lambda_1)\,p(X_2;\lambda_1)\dots p(X_n;\lambda_1)}{p(X_1;\lambda_0)\,p(X_2;\lambda_0)\dots p(X_n;\lambda_0)}= \\ =\left(\frac{\lambda_1}{\lambda_0}\right)^n e^{-(\lambda_1-\lambda_0)S_n},$$

где случайная величина $S_n=X_1+\dots+X_n$ представляет собой сумму результатов наблюдений за n шагов, а s_n — ее реализация. Область продолжения испытаний для критерия Вальда в данном случае задается неравенствами

$$B<\left(\frac{\lambda_1}{\lambda_0}\right)^n e^{-(\lambda_1-\lambda_0)s_n}<A,$$

или

$$C_1n-C_2a<s_n<C_1n+C_2b, \tag{4.53}$$

где $a=\ln A$, $b=-\ln B$,

$$C_2=\frac{1}{\lambda_1-\lambda_0}, \qquad C_1=C_2\ln\frac{\lambda_1}{\lambda_0}. \tag{4.54}$$

Испытания продолжают, если выполняются оба неравенства (4.53), и прекращают при первом нарушении хотя бы одного из этих неравенств. При нарушении правого неравенства в (4.53) принимают решение о справедливости гипотезы H_0, а при нарушении левого — решение о справедливо-

сти гипотезы H_1. Границы области прекращения наблюдений, как и в предыдущем примере, также имеют вид прямых линий на плоскости $(n,\ s_n)$ (рис. 4.3), эти линии задаются уравнениями $s_n = C_1 n + C_2 b$ (граница „области принятия" гипотезы H_0) и $s_n = C_1 n - C_2 a$ (граница „области принятия" гипотезы H_1).

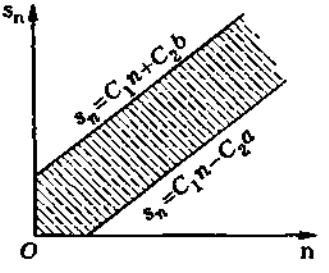


Рис. 4.3

Оценим средний объем испытаний. Случайная величина (4.34) в данном случае имеет вид

$$Z(X) = \ln \frac{p(X;\lambda_1)}{p(X;\lambda_0)} = \ln \frac{\lambda_1}{\lambda_0} - X(\lambda_1 - \lambda_0).$$

Учитывая, что $\mathrm{M}_0 X = 1/\lambda_0$, $\mathrm{M}_1 X = 1/\lambda_1$, получаем формулы для среднего объема испытаний (4.41)–(4.43) при $\lambda = \lambda_0$ и $\lambda = \lambda_1$:

$$\mathrm{M}_0 \nu \approx \frac{w(\alpha,\beta)}{\rho(\lambda_0,\lambda_1)}, \qquad \mathrm{M}_1 \nu \approx \frac{w(\beta,\alpha)}{\rho(\lambda_1,\lambda_0)}, \tag{4.55}$$

где

$$\rho(\lambda_0,\lambda_1) = \frac{\lambda_1 - \lambda_0}{\lambda_0} - \ln \frac{\lambda_1}{\lambda_0}. \quad \#$$

Приведем далее численный пример, иллюстрирующий выигрыш в среднем объеме испытаний, который дает последовательный критерий Вальда по сравнению с *оптимальным критерием Неймана — Пирсона* с детерминированным объемом испытаний.

Пример 4.18. Пусть в условиях примера 4.14 требуется проверить гипотезы (4.52) относительно параметра интенсивности отказов, где критические уровни равны $\lambda_0 = 0{,}1$ и $\lambda_1 = 0{,}2$. Заданные значения рисков первого и второго рода равны $\alpha = 0{,}1$ и $\beta = 0{,}1$.

Если номер шага прекращения наблюдений (объем испытаний) определен заранее и является детерминированной величиной: $\nu = n$, то наилучший для этого случая критерий Неймана — Пирсона имеет следующий вид:

– если $s_n < C$, то принимают гипотезу H_1;

– если $s_n \geqslant C$, то принимают гипотезу H_0.

Здесь s_n — значение случайной величины $S_n = X_1 + \ldots + X_n$. Риски первого и второго рода для этого критерия равны соответственно следующим вероятностям: $\mathbf{P}_0\{S_n < C\}$ и $\mathbf{P}_1\{S_n \geqslant C\}$. Тем самым объем испытаний $n^* = n^*(\alpha, \beta)$, необходимый для того, чтобы обеспечить заданные значения рисков α, β, определяется как минимальное целое число n, удовлетворяющее двум неравенствам

$$\mathbf{P}_0\left\{S_n(\vec{X}_n) < C\right\} \leqslant \alpha, \qquad \mathbf{P}_1\left\{S_n(\vec{X}_n) \geqslant C\right\} \leqslant \beta.$$

Эти неравенства могут быть записаны в виде

$$\mathbf{P}_0\{2\lambda_0 S_n < 2\lambda_0 C\} \leqslant \alpha, \qquad \mathbf{P}_1\{2\lambda_1 S_n \geqslant 2\lambda_1 C\} \leqslant \beta,$$

или, учитывая, что случайная величина $2\lambda S_n(\vec{X}_n)$ имеет χ^2-распределение с $2n$ степенями свободы (см. **Д.3.1**), в виде

$$\chi^2(2\lambda_0 C, 2n) \leqslant \alpha, \qquad 1 - \chi^2(2\lambda_1 C, 2n) \leqslant \beta,$$

где $\chi^2(x, 2n)$ — плотность χ^2-распределения с $2n$ степенями свободы. Отсюда после преобразований получаем, что необходимый объем испытаний n^* является минимальным целым числом n, удовлетворяющим неравенству

$$\chi^2_{1-\beta}(2n) \leqslant \frac{\lambda_1}{\lambda_0}\chi^2_{\alpha}(2n).$$

По таблице квантилей χ^2-распределения (см. табл. П.3) находим

$$n^* = \min\left\{n\colon \chi^2_{0,9}(2n) \leqslant 2\chi^2_{0,1}(2n)\right\} = 15.$$

Применяя далее формулы (4.55), находим средний объем испытаний при $\lambda = \lambda_0$ и $\lambda = \lambda_1$ для последовательного критерия Вальда (при тех же значениях λ_0, λ_1, α, β):

$$\mathrm{M}_0\nu \approx \frac{(1-\alpha)\ln\frac{1-\alpha}{\beta} + \alpha\ln\frac{\alpha}{1-\beta}}{\frac{\lambda_1}{\lambda_0} - 1 - \ln\frac{\lambda_1}{\lambda_0}} = \frac{0{,}9\ln 9 - 0{,}1\ln 9}{1-\ln 2} = 5{,}7,$$

$$\mathrm{M}_1\nu \approx \frac{(1-\beta)\ln\frac{1-\beta}{\alpha} + \beta\ln\frac{\beta}{1-\alpha}}{\frac{\lambda_0}{\lambda_1} - 1 - \ln\frac{\lambda_0}{\lambda_1}} = \frac{0{,}9\ln 9 - 0{,}1\ln 9}{\ln 2 - 0{,}5} = 9{,}2.$$

Таким образом, при $\lambda = \lambda_0$ и $\lambda = \lambda_1$ выигрыш в среднем объеме испытаний, который дает последовательный критерий Вальда по сравнению с детерминированным объемом испытаний n^*, равен соответственно

$$\frac{n^*}{\mathrm{M}_0\nu} \approx \frac{15}{5{,}7} = 2{,}63, \qquad \frac{n^*}{\mathrm{M}_1\nu} \approx \frac{15}{9{,}2} = 1{,}63.$$

4.7. Решение типовых примеров

Пример 4.19. Для *выборки объема* $n = 9$ построить *оптимальный критерий Неймана — Пирсона* для проверки двух *простых гипотез* относительно параметра μ нормального распределения

$$H_0\colon\ \mu = \mu_0 = 53, \qquad H_1\colon\ \mu = \mu_1 = 54$$

с заданным *уровнем значимости* (вероятностью *ошибки первого рода*) $\alpha = 0{,}1$ при известной дисперсии $\sigma^2 = 16$. Для построенного критерия найти вероятность *ошибки второго рода* β и *мощность критерия*.

Решение. В соответствии с результатами примера 4.4 *критическое множество* задается неравенством

$$\sum_{i=1}^{n} X_i \geqslant C, \tag{4.56}$$

где константа C выбирается из условия обеспечения заданного уровня $\alpha = 0{,}1$:

$$C = n\mu_0 + u_{1-\alpha}\sigma\sqrt{n} = 9 \cdot 53 + 1{,}28 \cdot 4 \cdot 3 = 492{,}4.$$

Для построенного критерия с критическим множеством (4.56) вероятность ошибки второго рода равна

$$\beta = \Phi\Big(\frac{C - n\mu_1}{\sigma\sqrt{n}}\Big) = \Phi\Big(\frac{492{,}4 - 9 \cdot 54}{4 \cdot 3}\Big) = 0{,}76.$$

Мощность критерия равна $1 - \beta = 0{,}24$. Значение мощности невелико, что объясняется в данном случае относительно малым объемом выборки $n = 9$.

Пример 4.20. В предыдущей задаче найти минимально необходимый объем выборки n^*, позволяющий обеспечить заданные значения вероятностей ошибок $\alpha = 0{,}1$, $\beta = 0{,}1$. Построить соответствующий оптимальный критерий Неймана — Пирсона в этой ситуации.

Р е ш е н и е. В соответствии с результатами примера 4.8

$$n^* = \frac{\sigma^2(u_{1-\alpha} + u_{1-\beta})^2}{(\mu_1 - \mu_0)^2} = \frac{16 \cdot (1{,}28 + 1{,}28)^2}{1^2} = 105.$$

Оптимальный критерий Неймана — Пирсона в этом случае задается с помощью критического множества

$$\sum_{i=1}^{n} x_i \geqslant C,$$

где константа C определяется из равенства (4.4):

$$C = n^*\mu_0 + u_{1-\alpha}\sigma\sqrt{n^*} = 570{,}4.$$

Пример 4.21. Партии волокна испытываются на прочность, при этом предел прочности X распределен по нормальному закону с дисперсией $\sigma^2 = 9$. Партия считается удовлетворительной, если среднее значение предела прочности входящих

в партию образцов $\mu = \mathbf{M}X \geqslant 14$, и неудовлетворительной, если $\mu \leqslant 10$. Из каждой партии на испытание ставится n образцов, для которых измеряют значения их прочности $x_1, \ldots, x_n$. Требуется проверить по результатам испытаний две *сложные гипотезы* H_0: $\mu \geqslant 14$, H_1: $\mu \leqslant 10$ с заданными максимальными вероятностями ошибок $\alpha = 0{,}1$, $\beta = 0{,}05$. Для этой ситуации необходимо решить следующие задачи:

– найти необходимый объем выборки n^*, при котором могут быть обеспечены данные значения α, β;

– построить *равномерно наиболее мощный критерий* при найденном объеме выборки;

– для построенного критерия найти *функцию мощности* и *оперативную характеристику*.

Решение. Для решения поставленных задач используем результаты **4.3**–**4.5**. Необходимый объем выборки находим по формуле (4.11):

$$n^* = \frac{\sigma^2(u_{1-\alpha} + u_{1-\beta})^2}{(\mu_1 - \mu_0)^2} = 5.$$

Равномерно наиболее мощный критерий совпадает с оптимальным критерием Неймана — Пирсона для двух простых гипотез H_0: $\mu = 14$, H_1: $\mu = 10$. Соответствующее критическое множество задается неравенством (4.16):

$$\sum_{i=1}^{n^*} x_i \leqslant C,$$

где $C = n^*\mu_0 - u_{1-\alpha}\sigma\sqrt{n^*}$, откуда $C = 61{,}4$.

Функция мощности (вероятность отвергнуть гипотезу H_0) в данном случае имеет вид

$$M(\mu) = \mathbf{P}\Big\{\sum_{i=1}^{n^*} X_i \leqslant C \,\Big|\, \mu\Big\} = \Phi\Big(\frac{C - n^*\mu}{\sigma\sqrt{n^*}}\Big).$$

Оперативная характеристика критерия $S(\mu) = 1 - M(\mu)$.

Пример 4.22. В условиях примера 4.7 найдем минимальный объем выборки, если H_0: $p = p_0 = 0{,}1$, H_1: $p = p_1 = 0{,}2$, $\alpha = 0{,}01$ и $\beta = 0{,}05$.

Решение. По таблице квантилей нормального распределения (см. табл. П.2) находим $u_{1-\alpha} = u_{0{,}99} = 2{,}33$, $u_\beta = u_{0{,}05} = -u_{0{,}95} = -1{,}65$. Далее, используя (4.12), получаем

$$n^* \geqslant \frac{\left(2{,}33\sqrt{0{,}1 \cdot 0{,}9} + 1{,}65\sqrt{0{,}2 \cdot 0{,}8}\right)^2}{(0{,}2 - 0{,}1)^2} \approx 185.$$

Пример 4.23. В цехе завода выпускают валы электродвигателей. Из продукции одного станка произвольно выбирают 50 изделий, измеряют их диаметры и вычисляют значение *выборочного среднего* $\overline{x} = 42{,}972$ мм. По техническим условиям станок настраивается на номинальный размер 43 мм. Можно ли на основании полученных результатов сделать вывод о том, что станок обеспечивает заданный номинальный размер, или полученные данные свидетельствуют о неудовлетворительной наладке технологического оборудования. Контролируемый признак имеет нормальное распределение, $\sigma^2 = 0{,}01$ мм^2.

Решение. Для оценки правильности настройки оборудования необходимо проверить гипотезу H_0: $\mu = \mu_0 = 43$ мм о математическом ожидании нормально распределенной *генеральной совокупности* X (σ^2 известна) при альтернативной гипотезе H_1: $\mu \neq 43$ мм, выбор который объясняется тем, что станок можно настроить на размер как выше, так и ниже номинального.

Выбираем уровень значимости $\alpha = 0{,}05$. Для рассматриваемых гипотез при $\alpha = 0{,}05$ критическое множество имеет вид (см. пример 4.13)

$$\left|\frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}\right| \geqslant 1{,}96,$$

где 1,96 — квантиль $u_{1-\alpha/2} = u_{0{,}975}$ стандартного нормального распределения (см. табл. П.2). Находим *выборочное значение*

статистики $Z = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$:

$$Z_в = \frac{42{,}972 - 43}{0{,}1}\sqrt{50} = -1{,}98.$$

Поскольку полученное значение принадлежит критическому множеству ($1{,}98 > 1{,}96$), то гипотезу H_0 отклоняем.

Пример 4.24. В условиях примера 4.23 проверим гипотезу H_0: $\mu = \mu_0 = 43$ мм при альтернативной гипотезе H_1: $\mu \neq 43$ мм, если σ^2 неизвестна. Рассчитанное по результатам выборочное среднее квадратичное отклонение $S = 0{,}1$ мм.

Решение. Выбираем уровень значимости $\alpha = 0{,}05$. По таблице квантилей *распределения Стьюдента* (см. табл. П.4) находим квантиль $t_{1-\alpha/2} = 2{,}01$ с числом степеней свободы 49. Критическое множество для рассматриваемых гипотез (см. пример 4.13) имеет вид

$$\left|\frac{\overline{x} - \mu_0}{S}\sqrt{n}\right| \geqslant 2{,}01.$$

Вычисляя выборочное значение статистики

$$t = \frac{\overline{X} - \mu_0}{S}\sqrt{n}(\vec{X}_n),$$

получаем

$$t_в = \frac{42{,}972 - 43}{0{,}1}\sqrt{50} = -1{,}98.$$

Полученное значение не принадлежит критическому множеству, поэтому гипотезу H_0 принимаем.

Пример 4.25. Ведутся наблюдения за состоянием технологического процесса. Разладка оборудования приводит к изменению номинального значения контролируемого признака X, имеющего нормальное распределение с дисперсией $\sigma^2 = 0{,}069\,\text{мм}^2$. Для проверки стабильности технологического про-

цесса через каждые три смены изучают выборку объема $n = 50$. По результатам двух выборок рассчитывают $\overline{x}_1 = 3{,}038$ мм и $\overline{x}_2 = 2{,}981$ мм. Проверим стабильность технологического процесса.

Решение. Для проверки стабильности технологического процесса необходимо проверить гипотезу о равенстве математических ожиданий H_1: $\mu_1 = \mu_2$ (σ^2 известна). В качестве конкурирующей гипотезы выбираем H_1: $\mu_1 > \mu_2$, так как номинальное значение контролируемого признака уменьшается с течением времени.

Выбираем уровень значимости, например $\alpha = 0{,}0027$. По таблице квантилей нормального распределения (см. табл. П.2) находим квантиль уровня $1 - \alpha = 0{,}9973$: $u_{1-\alpha} = 2{,}78$. Критическое множество имеет вид (см. пример 4.15)

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \geqslant 2{,}78.$$

Поскольку значение

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{3{,}038 - 2{,}981}{\sqrt{0{,}069\left(\frac{1}{50} + \frac{1}{50}\right)}} = 1{,}085$$

не принадлежит критическому множеству, гипотезу H_0 принимаем, т.е. делаем вывод, что технологический процесс на момент проверки можно считать стабильным.

Пример 4.26. Давление в камере измерялось дважды двумя манометрами. По результатам 10 замеров получены следующие данные (в единицах шкалы приборов): $\overline{x} = 1573$, $\overline{y} = 1671$, $S_1^2 = 0{,}72$, $S_2^2 = 0{,}75$. Выясним, есть ли основание считать, что давление в камере не изменилось, если ошибки измерения распределены по нормальному закону.

Решение. Проверяем гипотезу H_0: $\mu_1 = \mu_2$ при альтернативной H_1: $\mu_1 \neq \mu_2$, предполагая, что дисперсии не известны,

но одинаковы. Задаем уровень значимости $\alpha = 0{,}01$. Для построения критического множества используем статистику (см. пример 4.15)

$$\widetilde{T} = \frac{(\overline{X} - \overline{Y})\sqrt{(m+n-2)nm}}{\sqrt{((n-1)S_1^2(\vec{X}_n) + (m-1)S_2^2(\vec{X}_n))(n+m)}}.$$

По таблице квантилей распределения Стьюдента (см. табл. П.4) находим квантиль уровня $\alpha = 0{,}995$ с числом степеней свободы $n+m-2 = 18$: $t_{1-\alpha/2} = t_{0{,}995} = 2{,}88$. Рассчитываем выборочное значение статистики $\widetilde{T}$:

$$\frac{1573 - 1671}{\sqrt{10 \cdot 0{,}147}}\sqrt{\frac{10 \cdot 10 \cdot 18}{20}} \approx -243.$$

Гипотезу H_0 отвергаем, так как значение -243 принадлежит критическому множеству: $|-243| > 2{,}88$.

Пример 4.27. Цех выпускает болты. Из партии болтов взята выборка объема $n = 20$ и измерена длина каждого болта, по которым рассчитаны выборочное среднее $\overline{X} = 18$ мм и выборочная дисперсия $S^2 = 784\,\text{мм}^2$. Выясним, можно ли считать, что станок обеспечивает допустимый для данной партии разброс, или же расчетное значение S^2 указывает на несоответствие точности изготовления деталей предъявляемым требованиям, согласно которым $\sigma_0^2 = 400\,\text{мм}^2$. Контролируемый признак распределен по нормальному закону.

Решение. Для ответа на поставленный вопрос проверим гипотезу о величине дисперсии H_0: $\sigma_0^2 = 400\,\text{мм}^2$, выбрав в качестве альтернативной гипотезу H_1: $\sigma^2 > \sigma_0^2$. Назначаем уровень значимости $\alpha = 0{,}05$. Для того чтобы построить критическое множество, воспользуемся статистикой

$$V = \frac{(n-1)S^2(\vec{X}_n)}{\sigma^2}$$

(см. **Д.3.1**), которая имеет *распределение* χ^2 с числом степеней свободы $n-1$. По таблице квантилей этого распределения (см. табл. П.3) $\chi^2_{1-\alpha}(19) = 30{,}1$. Критическое множество имеет вид $\frac{(n-1)S^2}{\sigma^2} > 30{,}1$. Вычисляем выборочное значение статистики V:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{19 \cdot 784}{400} = 37{,}24.$$

Гипотезу H_0 отклоняем, так как $37{,}24 > 30{,}1$.

Пример 4.28. До наладки станка была проверена точность изготовления 10 изделий и найдена оценка дисперсии контролируемого признака $S_1^2 = 9{,}6$. После наладки измерено еще 15 изделий и получена оценка дисперсии $S_2^2 = 5{,}7$. Можно ли считать, что точность изготовления изделий после наладки повысилась? Контролируемый признак имеет нормальное распределение.

Решение. Для ответа на поставленный вопрос проверим гипотезу о равенстве дисперсий H_0: $\sigma_1^2 = \sigma_2^2$ при альтернативной гипотезе H_2: $\sigma_1^2 > \sigma_2^2$. Назначаем уровень значимости $\alpha = 0{,}05$.

Для построения критического множества используем статистику $F = S_1^2(\vec{X}_n)/S_2^2(\vec{X}_n)$, которая имеет *распределение Фишера* со степенями свободы $\nu_1 = 9$ и $\nu_2 = 14$ (см. **Д.3.1**). По таблице квантилей распределения Фишера (см. табл. П.5) находим $f_{1-\alpha}(9,14) = f_{0{,}95}(9,14) = 2{,}65$. Критическое множество имеет вид $S_1^2(\vec{x}_n)/S_2^2(\vec{x}_n) > 2{,}65$. Вычисляем значение статистики F: $9{,}6/5{,}7 = 1{,}68$. Гипотезу H_0 принимаем, так как $1{,}68 < 2{,}65$.

Пример 4.29. В условиях примера 4.15 найдем *границы областей принятия гипотез* H_0, H_1 и средний объем испытаний (нижнюю границу), если H_0: $q = q_0 = 1/3$, H_1: $q = q_1 = 1/2$, $\alpha = 0{,}05$, $\beta = 0{,}1$.

Решение. Для определения границ областей принятия гипотез H_0 и H_1 (см. (4.49)) находим $B \approx 0{,}1/0{,}95 = 2/19$,

$A = 0{,}9/0{,}05 = 18$. Тогда $b = -\ln(2/19) = 2{,}25$, $a = \ln 18 = 2{,}89$. Далее получаем

$$C_2 = \ln^{-1}\frac{0{,}1\cdot 0{,}95}{0{,}05\cdot 0{,}9} = \ln^{-1}(19/9) = 1{,}34, \qquad C_1 = \ln\frac{19/18}{19/9} = 0{,}072.$$

Итак, $d_n = 0{,}072n - 2{,}915$ — граница области принятия гипотезы H_0, а $d_n = 0{,}072n + 3{,}873$ — граница области принятия гипотезы H_1. Наконец, находим средний объем испытаний для двух рассматриваемых гипотез:

$$\mathrm{M}_0\nu = \frac{(1-0{,}05)\ln\frac{0{,}95}{0{,}1} + 0{,}05\ln\frac{0{,}05}{0{,}9}}{\frac{1}{3}\ln\frac{2}{3} + \frac{2}{3}\ln\frac{4}{3}} \approx 35{,}2;$$

$$\mathrm{M}_1\nu = \frac{0{,}9\ln\frac{0{,}9}{0{,}05} + 0{,}1\ln\frac{0{,}1}{0{,}95}}{\frac{1}{2}\ln\frac{3}{2} + \frac{1}{2}\ln\frac{3}{4}} \approx 40{,}3.$$

Вопросы и задачи

4.1. Что такое статистическая гипотеза (гипотеза)?

4.2. Какую статистическую гипотезу называют параметрической, однопараметрической, многопараметрической?

4.3. Какую гипотезу называют основной, альтернативной, простой, сложной?

4.4. Что такое статистический критерий?

4.5. Что такое уровень значимости критерия для проверки статистической гипотезы?

4.6. Какое множество называют критическим для проверки статистических гипотез?

4.7. В чем состоит ошибка первого рода, второго рода?

4.8. Что называют мощностью критерия?

4.9. Какой критерий называют оптимальным (наиболее мощным) при заданном уровне значимости?

4.10. Что называют размером критерия?

4.11. Какую функцию называют функцией мощности критерия, оперативной характеристикой критерия?

4.12. Какой критерий называют равномерно наиболее мощным?

4.13. В чем состоит общий метод отношения правдоподобия для сложных параметрических гипотез?

4.14. В чем состоит последовательный критерий отношения правдоподобия (критерий Вальда)?

4.15. Чему равны средний объем испытаний, нижняя граница среднего объема испытаний для критерия Вальда?

4.16. Генеральная совокупность имеет нормальный закон распределения, $\sigma^2 = 1$. Укажите объем выборки, при котором может быть построен критерий для проверки двух простых гипотез H_0: $\mu = \mu_0 = 4{,}6$, H_1: $\mu = \mu_1 = 5$. Заданы вероятности $\alpha = 0{,}01$ — ошибка первого рода и $\beta = 0{,}05$ — ошибка второго рода.

Указание: используйте решение примера 4.8.

Ответ: $n^* \geqslant 99$.

4.17. Из продукции автомата, производящего некоторые детали с номинальным значением контролируемого размера $\mu_0 = 40$мм, была взята выборка объема $n = 36$. Значение выборочного среднего контролируемого размера $\overline{x} = 40{,}2$мм. Есть основание предполагать, что фактические размеры образуют нормальную генеральную совокупность с дисперсией $\sigma^2 = 1\,\text{мм}^2$. Выясните:

а) можно ли по результатам проведенного выборочного обследования утверждать, что контролируемый размер не больше номинального (принять $\alpha = 0{,}01$);

б) каково критическое множество в этом случае.

Ответ: а) да; б) $\overline{x} > 40{,}32$.

4.18. В соответствии с техническими условиями среднее время безотказной работы приборов из большой партии должно составлять не менее 1000 ч со средним квадратичным отклонением 100 ч. Значение выборочного среднего времени безотказной работы для случайно отобранных 25 приборов оказалось равным 970 ч. Предположим, что среднее квадратичное времени безотказной работы для приборов в выборке совпадает со средним квадратичным во всей партии, а контролируемая характеристика имеет нормальное распределение. Выясните, можно ли считать, что вся партия приборов не удовлетворяет техническим условиям, если: а) $\alpha = 0{,}1$; б) $\alpha = 0{,}01$.

Ответ: а) да; б) нет.

4.19. Решите предыдущую задачу при условии, что среднее квадратичное отклонение времени безотказной работы, вычисленное по выборке, равно 115 ч.

Ответ: а) нет; б) да.

4.20. Утверждается, что шарики, изготовленные станком-автоматом, имеют средний диаметр $d_0 = 10$ мм. Используя односторонний критерий при $\alpha = 0{,}05$, проверьте эту гипотезу, если в выборке из $n = 16$ шариков средний диаметр оказался равным 10,3 мм, считая, что: а) дисперсия σ^2 известна и равна $\sigma^2 = 1\,\text{мм}^2$; б) значение оценки дисперсии, определенное по выборке, составляет $S^2 = 1{,}21\,\text{мм}^2$. Контролируемый размер имеет нормальное распределение.

Ответ: а) гипотеза принимается; б) гипотеза принимается.

4.21. Для проверки внутреннего диаметра кольца была взята выборка объема $n = 25$ и найдены отклонения от размера (погрешность изготовления) 100 мм. По результатам измерений подсчитано значение выборочного среднего $\overline{x} = 31{,}52$ мм и оценка среднего квадратичного отклонения $S = 6$ мм. Требуется проверить, существенно ли превышает рассчитанное по выборке среднее значение (31,52 мм) номинальный размер

(30 мм). В производстве недопустимы большие положительные отклонения. Погрешность изготовления имеет нормальное распределение. Уровень значимости $\alpha = 0{,}05$.

Ответ: номинальный размер согласуется с опытными данными.

4.22. Из большой партии резисторов одного типа и номинала случайным образом отобраны 37 шт. Значение выборочного среднего величины сопротивления при этом оказалось равным 9,3 кОм. Используя двусторонний критерий при $\alpha = 0{,}05$, проверьте гипотезу о том, что выборка взята из партии с номинальным значением 10 кОм при альтернативной гипотезе, согласно которой номинальное значение не равно 10 кОм, если: а) дисперсия рассматриваемой случайной величины известна и равна 4 кОм^2; б) дисперсия значения сопротивления неизвестна, а значение выборочной дисперсии равно 6,25 кОм. Распределение контролируемого признака нормальное.

Ответ: а) гипотеза отклоняется; б) гипотеза принимается.

4.23. Установка имеет среднюю производительность 1000 кг вещества в сутки со средним квадратичным отклонением, равным 80 кг^2. При изменении технологии производительность возрастает до 1100 кг вещества в сутки с тем же средним квадратичным отклонением. Можно ли считать, что новая технология обеспечивает повышение производительности, если: а) $\alpha = 0{,}05$; б) $\alpha = 0{,}1$? Контролируемый признак имеет нормальное распределение.

Ответ: а) да; б) да.

4.24. Ожидается, что при добавлении специальных веществ жесткость воды уменьшается. По оценкам жесткости воды до и после добавления специальных веществ по 40 и 50 пробам соответственно получили средние значения жесткости (в стандартных единицах), равные 4,0 и 3,8. Дисперсия измерений в обоих случаях предполагается равной 0,25. Подтверждают ли

эти результаты ожидаемый эффект? Принять $\alpha = 0{,}05$. Контролируемый признак имеет нормальное распределение.

Ответ: да.

4.25. Два штурмана определяли пеленг маяка по нескольким замерам, используя различные пеленгаторы. Результаты замеров: $\overline{x} = 70{,}2°$ при $n_1 = 4$ и $\overline{y} = 70{,}5°$ при $n_2 = 9$. С помощью двустороннего критерия проверьте при $\alpha = 0{,}05$ гипотезу о том, что различие результатов вызвано только случайными ошибками, если средние квадратичные отклонения для обоих пеленгаторов известны и равны $\sigma_1 = \sigma_2 = 0{,}5°$.

Ответ: гипотеза принимается.

4.26. Заводы A и B выпускают приборы одного типа. По выборке из 50 приборов завода A установили среднюю продолжительность работы прибора 1288 ч со средним квадратичным отклонением 80 ч, а также по выборке того же объема с завода B — 1208 ч со средним квадратичным отклонением 94 ч. На уровне значимости $\alpha = 0{,}05$ проверьте гипотезу о том, что средний срок службы приборов с обоих заводов одинаков. Считать, что продолжительность работы одного прибора распределена приближенно по нормальному закону.

Ответ: гипотеза отклоняется.

4.27. При обработке втулок на станке-автомате ведутся наблюдения за режимом его работы. Для проверки стабильности работы станка через определенные промежутки времени изучают выборки объема $n = 10$. По результатам двух выборок (табл. 4.1) проверьте стабильность работы станка. Распределение контролируемого признака предполагается нормальным. Также предполагается, что дисперсии генеральных совокупностей, из которых получены выборки, равны. Уровень значимости $\alpha = 0{,}05$.

Ответ: гипотезу о стабильности работы станка следует отклонить.

Таблица 4.1

Номер изделия	1	2	3	4	5	6	7	8	9	10
x_i	2,060	2,063	2,068	2,060	2,067	2,063	2,059	2,062	2,062	2,060
y_i	2,063	2,060	2,057	2,056	2,059	2,058	2,062	2,059	2,059	2,057

4.28. Точность наладки станка-автомата, производящего некоторые детали, характеризуется дисперсией длины деталей. Если эта величина будет больше 400 мкм^2, станок останавливается для наладки. Значение выборочной дисперсии, найденное по 15 случайно отобранным деталям из продукции станка, оказалось равным $\sigma^2 = 680$ мкм^2. Определите, нужна ли наладка станка, если: а) $\alpha = 0{,}01$; б) $\alpha = 0{,}1$. Контролируемый признак имеет нормальное распределение.

Ответ: а) нет; б) да.

4.29. При изменении определенной процедуры проверки коэффициента трения установлено, что дисперсия результатов измерений этого коэффициента составляет 0,1. Значение выборочной дисперсии, вычисленное по результатам 26 измерений коэффициента трения, оказалось равным 0,2. При уровне значимости $\alpha = 0{,}1$ проверьте гипотезу о том, что дисперсия результатов измерений коэффициента трения равна 0,1. Предполагается, что контролируемый признак имеет нормальное распределение.

Ответ: гипотеза отклоняется.

4.30. На двух токарных автоматах изготавливают детали по одному чертежу. Из продукции первого станка было отобрано $n_1 = 9$ деталей, а из продукции второго $n_2 = 11$ деталей. Оценки выборочных дисперсий контрольного размера, определенные по этим выборкам, равны $\widehat{\sigma}_1^2 = 5{,}9$ мкм^2 и $\widehat{\sigma}_2^2 = 23{,}3$ мкм^2 соответственно. Проверьте гипотезу о равенстве дисперсий при $\alpha = 0{,}05$, если альтернативная гипотеза утверждает следу-

ющее: а) дисперсии не равны; б) дисперсия размера для второго станка больше, чем для первого.

Ответ: а) гипотеза принимается; б) гипотеза отклоняется.

4.31. Давление в камере контролируется по двум манометрам. Для сравнения точности этих приборов одновременно фиксируют их показания. По результатам 10 замеров значения оценок (в единицах шкалы приборов) оказались следующими: $\overline{x} = 1573$, $\overline{y} = 1671$, $S_x^2 = 0{,}72$, $S_y^2 = 0{,}15$. При $\alpha = 0{,}1$ проверьте гипотезу о равенстве дисперсий.

Ответ: гипотеза принимается.

4.32. На двух станках A и B производят одну и ту же продукцию, контролируемую по внутреннему диаметру изделия. Из продукции станка A была взята выборка из 16 изделий, а из продукции станка B — выборка из 25 изделий и получены значения $\overline{x}_A = 36{,}5$ мм, $S_A^2 = 1{,}21$ мм^2, $\overline{x}_B = 36{,}8$ мм, $S_B^2 = 1{,}44$ мм^2. Проверьте гипотезу о равенстве математических ожиданий контролируемых размеров в продукции обоих станков при двусторонней альтернативной гипотезе, если: а) $\alpha = 0{,}05$, б) $\alpha = 0{,}1$. Предполагается, что распределение контролируемых размеров нормальное и $\sigma_A^2 = \sigma_B^2$.

Ответ: а) гипотеза принимается; б) гипотеза принимается.

4.33. Сравниваются прочностные характеристики сталей марок A и B. Для этого испытаны на предел прочности 145 образцов марки A и 200 образцов марки B. В результате получили $\overline{x}_1 = 31{,}40$, $S_1^2 = 3{,}36$, $\overline{x}_2^2 = 28{,}84$, $S_2^2 = 3{,}51$. Можно ли на уровне значимости $\alpha = 0{,}1$ считать, что стали имеют разные прочностные характеристики? Предварительно следует убедиться, что дисперсии равны. Контролируемый признак имеет нормальное распределение.

Ответ: да.

4.34. При 50 подбрасываниях монеты „герб“ появился 20 раз. Можно ли считать, что процент появления „герба“ не равен 50? Принять $\alpha = 0{,}10$.

Ответ: гипотеза принимается.

4.35. При 120 бросаниях игральной кости „шестерка“ выпала 40 раз. Согласуется ли этот результат с утверждением, что кость „правильная“?

Ответ: нет.

4.36. В условиях примера 4.15 найдите границы областей принятия гипотез H_0, H_1 и средний объем испытаний, если H_0: $q = q_0 = 0{,}4$; H_1: $q = q_1 = 0{,}5$; $\alpha = 0{,}05$; $\beta = 0{,}05$.

Ответ: $d_n = 0{,}45n - 8{,}26$; $d_n = 0{,}45n + 8{,}26$; $\mathrm{M}_0\nu = 127$; $\mathrm{M}_1\nu = 148$.

4.37. В условиях примера 4.16 найдите границы областей принятия гипотез H_0, H_1 и средний объем испытаний при $\lambda = \lambda_0 = 0{,}1$; $\lambda = \lambda_1 = 0{,}3$.

Ответ: $S_n \simeq 5{,}49n + 11{,}25$; $S_n = 5{,}49n - 14{,}45$; $\mathrm{M}_0\nu = 3$; $\mathrm{M}_1\nu = 6$.

5. ПРОВЕРКА НЕПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ

Статистические методы, изложенные в 2–4, опираются на различные априорные допущения о виде исследуемой *статистической модели.* Например, *метод максимального правдоподобия* применяют при известном (с точностью до вектора параметров) законе распределения *генеральной совокупности.* Основные методы построения *доверительных интервалов* и проверки *статистических гипотез* основаны на предположении о нормальном законе распределения генеральной совокупности. Все эти методы предполагают, что результаты наблюдений являются реализациями независимых случайных величин.

Оказывается, что многие предположения о виде статистической модели, в том числе все перечисленные выше, можно сформулировать как статистические гипотезы и проверить при помощи *статистических критериев* на основании *статистических данных.* Наиболее важные из этих критериев рассмотрены в этой главе.

5.1. Критерии согласия. Простая гипотеза

Критериями согласия называют статистические критерии, предназначенные для обнаружения расхождений между гипотетической статистической моделью и реальными данными, которые эта модель призвана описать. Другими словами, они выясняют, насколько предположения о распределении случайных величин соответствуют экспериментальным данным, т.е. не вступает ли принятая статистическая модель в противоречие с имеющимися данными.

Критерий Колмогорова. Пусть $\vec{X}_n$ — *случайная выборка объема* n *из генеральной совокупности* X. Рассмотрим задачу проверки *простой статистической гипотезы* H_0 о том, что функция распределения $F(t)$ случайной величины X совпадает с некоторой известной функцией $F_0(t)$:

$$H_0\colon\ F(t) = F_0(t),\quad t \in \mathbb{R}. \tag{5.1}$$

Предположим, что случайная величина X непрерывна. Проверка *основной гипотезы* H_0 против *альтернативной гипотезы*

$$H_1\colon\ F(t) \neq F_0(t)\ \text{ для некоторых } t \in \mathbb{R} \tag{5.2}$$

основана на *статистике* $D(\vec{X}_n)$, реализации $D(\vec{x}_n)$ которой определяют по формуле

$$D(\vec{x}_n) = \sup_t |F_n(t) - F_0(t)|, \tag{5.3}$$

где $F_n(t)$ — *эмпирическая функция распределения*, построенная по *реализации* $\vec{x}_n$ *случайной выборки* $\vec{X}_n$.

При заданной вероятности α совершения *ошибки первого рода* ***критерий Колмогорова**** отклоняет гипотезу H_0 в пользу H_1 на *уровне значимости* α, если

$$D(\vec{x}_n) > D_{1-\alpha},$$

где $D_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения случайной величины $D(\vec{X}_n)$ при условии истинности основной гипотезы H_0.

Если же

$$D(\vec{x}_n) \leqslant D_{1-\alpha},$$

то делается вывод о непротиворечивости (согласии) статистических данных гипотезе H_0.

*А.Н. Колмогоров (1903–1987) — крупнейший советский математик, один из создателей теории вероятностей.

Разобраться в сути этого формального определения можно при помощи следующего нестрогого рассуждения. Согласно теореме 1.1, случайная величина $\widehat{F}(t;\vec{X}_n) - F_0(t)$, где $\widehat{F}(t;\vec{X}_n)$ — *выборочная функция распределения*, для любого t в случае истинности основной гипотезы H_0 стремится к нулю при $n \to \infty$, а в случае истинности альтернативной гипотезы H_1 — к величине $F(t) - F_0(t)$, которая для некоторых значений t может быть отлична от нуля. Поэтому при $n \to \infty$ случайная величина $D(\vec{X}_n)$ стремится к неслучайной величине $\sup\limits_t |F(t) - F_0(t)|$, которая в случае истинности основной гипотезы H_0 равна нулю, а в случае истинности альтернативной гипотезы H_1 является положительной величиной. Следовательно, если для *статистических данных*, представленных *выборкой* $\vec{x}_n$, случайная величина $D(\vec{X}_n)$ приняла „достаточно большое" значение, то гипотезу H_0 естественно отклонить в пользу гипотезы H_1, а если $D(\vec{X}_n)$ приняла значение, „близкое к нулю", то гипотезу H_0 следует принять.

Оказывается, что при истинности основной гипотезы H_0 распределение случайной величины $D(\vec{X}_n)$ не зависит от $F_0(t)$ (хотя зависит от *объема выборки* n), что чрезвычайно важно для вычисления квантилей случайной величины $D(\vec{X}_n)$, поскольку не нужно составлять отдельные таблицы значений функции распределения статистики $D(\vec{X}_n)$ для каждой функции $F_0(t)$, а можно обойтись всего лишь одной таблицей. Это свойство вытекает из приводимой без доказательства следующей теоремы*.

Теорема 5.1. Пусть $\widehat{R}(t,\vec{X}_n)$ — выборочная функция распределения, построенная по случайной выборке $\vec{X}_n$ объема n из генеральной совокупности с равномерным законом распределения на отрезке $[0,1]$. Тогда при истинности H_0 функция распределения случайной величины $D(\vec{X}_n)$ совпадает с функ-

*См.: *Ивченко Г.И., Медведев Ю.И.*

цией распределения случайной величины

$$\sup_{0\leqslant t\leqslant 1}\left|\hat{R}(t,\vec{X}_n)-t\right|. \quad \#$$

Из теоремы 5.1 следует, что для проверки гипотезы о виде распределения достаточно составить таблицы значений функции распределения статистики $D(\vec{X}_n)$ только для случайной выборки $\vec{X}_n$ из генеральной совокупности X с равномерным законом распределения. Для $n\leqslant 100$ такие таблицы существуют*. При больших n для вычисления квантилей $D_{1-\alpha}$ уровня $1-\alpha$ следует использовать приближенную формулу, которая основана на доказанном А.Н. Колмогоровым предельном соотношении

$$\lim_{n\to\infty}\mathbf{P}\left\{\sqrt{n}D(\vec{X}_n)<t\right\}=K(t),\quad t>0,$$

где

$$K(t)=\sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2t^2}. \tag{5.4}$$

Это соотношение справедливо при истинности основной гипотезы H_0. Из него следует, что если n достаточно велико, то

$$D_{1-\alpha}\approx\frac{t_{1-\alpha}}{\sqrt{n}},$$

где величина $t_{1-\alpha}$ определяется уравнением

$$K(t_{1-\alpha})=1-\alpha.$$

Подробные таблицы значений функции $K(t)$ приведены в литературе**. Как показывает практика, приближением с помощью функции $K(t)$ можно пользоваться уже при $n\geqslant 20$. Для вычисления значений $D(\vec{x}_n)$ статистики $D(\vec{X}_n)$ удобна формула

$$D(\vec{x}_n)=\max_{1\leqslant i\leqslant n}\left\{\frac{i}{n}-F_0(x_{(i)}),F_0(x_{(i)})-\frac{i-1}{n}\right\}, \tag{5.5}$$

*См.: *Большев Л.Н., Смирнов Н.В.*

**См. там же.

которую можно также записать в виде

$$D(\vec{x}_n) = \max_{1 \leqslant i \leqslant n} \left(\left| F_0(x_{(i)}) - \frac{2i-1}{2n} \right| + \frac{1}{2n} \right).$$

Здесь $x_{(i)}$, $i = \overline{1, n}$, — *члены вариационного ряда*, построенного по выборке $x_1, \ldots, x_n$.

Пример 5.1. Для выборки $\vec{x}_{10}$ объема 10 с элементами

−0,29;	1,06;	0,16;	−0,12;	−1,20;
1,09;	−0,91;	1,22;	−1,15;	1,29

на уровне значимости $\alpha = 0{,}1$ проверим гипотезу H_0 о том, что эта выборка является реализацией случайной выборки $\vec{X}_n$ из генеральной совокупности X, имеющей стандартное нормальное распределение. Это распределение, согласно (5.1), имеет функцию распределения

$$F_0(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{s^2}{2}}\, ds.$$

В качестве альтернативной возьмем гипотезу (5.2).

Вариационный ряд $x_{(1)}, \ldots, x_{(10)}$ выборки $\vec{x}_{10}$ будет иметь вид

−1,20;	−1,15;	−0,91;	−0,29;	−0,12;
0,16;	1,06;	1,09;	1,22;	1,29.

Значения функции распределения $F_0(t)$ в этих точках равны

0,115;	0,125;	0,181;	0,386;	0,452;
0,564;	0,855;	0,862;	0,899;	0,901.

Вычисляем значения функции $\frac{i}{n} - F_0(x_{(i)})$ при $i = \overline{1, 10}$ и $n = 10$:

−0,015;	0,075;	0,119;	0,014;	0,048;
0,036;	−0,155;	−0,062;	0,001;	0,099,

и значения $F_0(x_{(i)}) - \frac{i-1}{n}$ при тех же i и n:

0,115;	0,025;	−0,019;	0,086;	0,052;
0,064;	0,255;	0,162;	0,089;	0,001.

Наибольшим из этих чисел будет 0,255. Значит, $D(\vec{x}_{(10)}) = 0{,}255$. По таблице квантилей статистики* $D(\vec{X}_n)$ для $n = 10$ и $\alpha = 0{,}1$ находим $D_{1-\alpha} = D_{0,9} = 0{,}369$. Так как $D(\vec{x}_{10}) < D_{0,9}$, то оснований отклонить гипотезу H_0 нет.

Критерий ω^2. Из равенства (5.3), задающего статистику $D(\vec{X}_n)$, следует, что критерий Колмогорова „хорошо различает" функции распределения $F(t)$ и $F_0(t)$, отличающиеся друг от друга достаточно сильно пусть даже на небольшом интервале. Если же число $\sup\limits_t |F(t) - F_0(t)|$ невелико, но $F(t) \neq F_0(t)$ на достаточно большом промежутке, то можно показать, что для проверки гипотезы (5.1) при альтернативе (5.2) целесообразно использовать так называемый ***критерий*** ω^2 (омега-квадрат), использующий статистику

$$\omega^2(\vec{X}_n) = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left(F_0(X_{(i)}) - \frac{2i-1}{2n}\right)^2, \qquad (5.6)$$

где $X_{(i)}$, $i = \overline{1, n}$, — элементы вариационного ряда случайной выборки $X_1, \ldots, X_n$. Основная гипотеза (5.1) отклоняется в пользу альтернативной гипотезы (5.2) на уровне значимости α, если

$$\omega^2(\vec{x}_n) > \omega^2_{1-\alpha},$$

где $\omega^2_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения статистики $\omega^2(\vec{X}_n)$ при условии истинности гипотезы H_0.

Так же как и для критерия Колмогорова, можно доказать, что распределение статистики $\omega^2(\vec{X}_n)$ при истинности основной

*См.: *Большев Л.Н., Смирнов Н.В.*

гипотезы H_0 не зависит от F_0. Для малых n существуют таблицы* квантилей статистики $\omega^2(\vec{X}_n)$. При больших n нужно пользоваться предельным распределением статистики $n\omega^2(\vec{X}_n)$, для которого также составлены таблицы.

Пример 5.2. Вернемся к задаче, рассмотренной в примере 5.1, но для ее решения используем критерий ω^2.

С помощью формулы (5.6) найдем значение $\omega^2(\vec{x}_{10})$ статистики $\omega^2(\vec{X}_n)$, используя значения $F_0(x_{(i)})$, $i=\overline{1,10}$, вычисленные в примере 5.1:

$$\omega^2(\vec{x}_{10})=\frac{1}{12\cdot 10^2}+\frac{1}{10}\Big(0{,}065^2+0{,}125^2+0{,}169^2+0{,}064^2+0{,}098^2+$$

$$+0{,}086^2+0{,}105^2+0{,}012^2+0{,}061^2+0{,}149^2\Big)\approx 0{,}0114.$$

По таблицам распределения статистики $\omega^2(\vec{X}_n)$ для $n=10$ находим

$$\omega^2_{1-\alpha}=\omega^2_{0{,}95}\approx 0{,}046.$$

Так как $\omega^2(\vec{x}_{10})<\omega^2_{0{,}95}$, то гипотеза H_0 на уровне значимости $\alpha=0{,}05$ не отклоняется.

Критерий согласия χ^2. При анализе критериев Колмогорова и ω^2 предполагалось, что $\vec{X}_n$ — случайная выборка объема n из генеральной совокупности непрерывной случайной величины X. Пусть теперь наблюдается дискретная случайная величина X, принимающая r различных значений $u_1, \ldots, u_r$ с положительными вероятностями $p_1, \ldots, p_r$:

$$\mathbf{P}\{X=u_k\}=p_k,\quad k=\overline{1,r},\quad \sum_{k=1}^{r}p_k=1.$$

Допустим, что в выборке $\vec{x}_n=(x_1, \ldots, x_n)$ число u_k встретилось $n_k(\vec{x}_n)$ раз, $k=\overline{1,r}$. Отметим, что $\sum\limits_{k=1}^{r} n_k(\vec{x}_n)=n$, т.е.

*См.: *Большев Л.Н., Смирнов Н.В.*

случайные величины $n_1(\vec{X}_n)$, ..., $n_r(\vec{X}_n)$ зависимы. При этих условиях справедлива следующая теорема*.

Теорема 5.2 (теорема Пирсона). Распределение случайной величины

$$\sum_{k=1}^{r} \frac{(n_k(\vec{X}_n) - np_k)^2}{np_k}$$

при $n \to \infty$ слабо сходится к χ^2-*распределению* с $r-1$ степенями свободы. #

Этой теоремой можно воспользоваться для проверки *простой гипотезы*

$$H_0\colon\ p_1 = p_{10},\ \ \ldots,\ \ p_r = p_{r0}, \tag{5.7}$$

где $p_{10}, \ldots, p_{r0}$ — известные величины, против альтернативной гипотезы

$$H_1\colon \text{ существуют такие } k, \text{ что } p_k \neq p_{k0},\ k = \overline{1, r}. \tag{5.8}$$

Если истинной является гипотеза H_0, то по теореме 5.2 при $n \to \infty$ распределение случайной величины

$$\chi^2(\vec{X}_n) = \sum_{k=1}^{r} \frac{(n_k(\vec{X}_n) - np_{k0})^2}{np_{k0}} = n \sum_{k=1}^{r} \frac{\left(\frac{n_k(\vec{X}_n)}{n} - p_{k0}\right)^2}{p_{k0}} \tag{5.9}$$

стремится к распределению χ^2 с $r-1$ степенями свободы.

Если основная гипотеза H_0 не является истинной, то в этом случае по закону больших чисел при $n \to \infty$

$$\frac{n_k(\vec{X}_n)}{n} \to p_k, \quad k = \overline{1, r}.$$

*Доказательство теоремы см.: *Крамер Г.*

Поэтому при $n \to \infty$

$$\frac{n_k(\vec{X}_n)}{n} - p_{k0} = \left(\frac{n_k(\vec{X}_n)}{n} - p_k\right) + (p_k - p_{k0}) \to p_k - p_{k0}.$$

Следовательно, если $p_k - p_{k0} \neq 0$ для некоторых $k = \overline{1,r}$, то статистика $\chi^2(\vec{X}_n)$ принимает бóльшие значения, чем в случае истинности основной гипотезы H_0.

Таким образом, становится естественным следующее определение ***критерия согласия*** χ^2 (хи-квадрат). Этот критерий при больших n на уровне значимости α отклоняет гипотезу H_0 в пользу альтернативной гипотезы H_1, если

$$\chi^2(\vec{x}_n) > \chi^2_{1-\alpha}(r-1),$$

где $\chi^2_{1-\alpha}(r-1)$ — квантиль уровня $1-\alpha$ χ^2-распределения с $r-1$ степенями свободы, а $\chi^2(\vec{x}_n)$ — реализация случайной величины (5.9).

Если же

$$\chi^2(\vec{x}_n) \leqslant \chi^2_{1-\alpha}(r-1),$$

то делается вывод о том, что гипотеза H_0 не противоречит статистическим данным и ее следует принять.

В отличие от критериев Колмогорова и ω^2 критерием χ^2 при небольших объемах выборки n пользоваться нельзя. Более того, для удовлетворительной аппроксимации распределения случайной величины $\chi^2(\vec{X}_n)$ распределением χ^2 необходимо, чтобы не только n было велико, но и все величины np_k, $k = \overline{1,r}$, также были немалыми. На практике при небольших r необходимо, чтобы выполнялись условия $np_k \geqslant 10$, $k = \overline{1,r}$, а если r велико ($r \geqslant 20$), достаточно, чтобы было $np_k \geqslant 5$, $k = \overline{1,r}$. Поскольку теорема Пирсона носит асимптотический характер, то ***критерий*** χ^2 является ***асимптотически непараметрическим.***

Критерий χ^2 можно использовать и тогда, когда случайная величина X непрерывна или дискретна, но принимает счетное множество значений с положительными вероятностями.

В этом случае множество M возможных значений X разбивают на r непересекающихся подмножеств M_k, $k=\overline{1,r}$, таким образом, чтобы вероятность p_k, $k=\overline{1,r}$, попадания случайной величины X в k-е подмножество M_k удовлетворяла условию $np_k \geqslant 5$ или $np_k \geqslant 10$, $k=\overline{1,r}$. Если X — непрерывная случайная величина, то в качестве M_k, $k=\overline{1,r}$, обычно берут множества вида

$$(-\infty, s_1), \quad [s_1, s_2), \quad \ldots, \quad [s_{r-2}, s_{r-1}), \quad [s_{r-1}, \infty),$$

где $s_1 < s_2 < \ldots < s_{r-1}$, $s_k \in \mathbb{R}$, $k=\overline{1,r-1}$.

Определим дискретную случайную величину X', принимающую значение k тогда и только тогда, когда $X \in M_k$, $k=\overline{1,r}$. В этом случае исходная задача проверки статистических гипотез сводится к проверке основной гипотезы (5.7) при альтернативной гипотезе (5.8), где в случае непрерывности случайной величины X

$$p_{k0} = \int\limits_{M_k} dF_0(t) = \int\limits_{M_k} p_0(t)\,dt \quad —$$

вероятность попадания случайной величины X в множество M_k в предположении, что функция распределения случайной величины X есть $F_0(t)$, а плотность — $p_0(t)$. Если X — дискретная случайная величина, имеющая счетное множество возможных значений $z_1, z_2, \ldots$ и $\mathrm{P}\{X = z_j\} = q_j > 0$, $j = 1, 2, \ldots$, то вместо проверки гипотезы

$$H_0\colon\ q_j = q_{j0}, \quad j = 1, 2, \ldots,$$

где q_{j0}, $j = 1, 2, \ldots$, — известные числа, при альтернативной гипотезе

$$H_1\colon \text{ существуют такие } j, \text{ что } q_j \neq q_{j0}, \quad j = 1, 2, \ldots,$$

проверяют гипотезу (5.7) при альтернативной гипотезе (5.8), где вероятности p_{k0}, $k=\overline{1,r}$, вычисляют по формулам

$$p_{k0}=\sum_{z_j\in M_k} q_{j0},\quad k=\overline{1,r}.$$

Далее для выборки $\vec{x}_n$ находят число $n_k(\vec{x}_n)$ ее элементов, принадлежащих множеству M_k, $k=\overline{1,r}$. Затем, подставляя $\vec{x}_n$ вместо $\vec{X}_n$ в формулу (5.9), определяют реализацию $\chi^2(\vec{x}_n)$ случайной величины $\chi^2(\vec{X}_n)$. Гипотеза H_0 отклоняется в пользу гипотезы H_1, если $\chi^2(\vec{x}_n)>\chi^2_{1-\alpha}(r-1)$ и принимается в противном случае.

Недостатком использования критерия χ^2 для случайных величин, принимающих бесконечное множество значений, является некоторая потеря информации при переходе от X к случайной величине X' с конечным числом значений.

Пример 5.3. Среди *элементов выборки* $\vec{x}_{1000}$ дискретной случайной величины X значение 0 встретилось 343 раза, значение 1 — 372 раза, значение 2 — 201 раз, значение 3 — 68 раз, а значения, бóльшие или равные 4, встретились 16 раз. Проверим на уровне значимости $\alpha=0{,}05$ гипотезу H_0 о том, что наблюдаемая случайная величина имеет распределение Пуассона с параметром $\lambda=1$, т.е.

$$\mathbf{P}\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\ldots$$

Предполагая истинность основной гипотезы H_0, находим

$$p_{01}=\mathbf{P}\{X=0\}=0{,}368,\qquad p_{02}=\mathbf{P}\{X=1\}=0{,}368,$$
$$p_{03}=\mathbf{P}\{X=2\}=0{,}184,\qquad p_{04}=\mathbf{P}\{X=3\}=0{,}061,$$
$$p_{05}=\mathbf{P}\{X\geqslant 4\}=0{,}019.$$

Заменим случайную величину X, принимающую бесконечное число значений, случайной величиной X', принимающей

только пять различных значений 0, 1, 2, 3 и 4 с положительными вероятностями $p_{10}=0{,}368$, $p_{20}=0{,}368$, $p_{30}=0{,}184$, $p_{40}=0{,}061$ и $p_{50}=0{,}019$ соответственно. По формуле (5.9) для $r=5$, $n=1000$ получаем

$$\chi^2(\vec{x}_n)=\frac{(343-0{,}368\cdot 1000)^2}{0{,}368\cdot 1000}+\frac{(372-0{,}368\cdot 1000)^2}{0{,}368\cdot 1000}+$$

$$+\frac{(201-0{,}184\cdot 1000)^2}{0{,}184\cdot 1000}+\frac{(68-0{,}061\cdot 1000)^2}{0{,}061\cdot 1000}+\frac{(16-0{,}019\cdot 1000)^2}{0{,}019\cdot 1000}=$$

$$=1{,}6984+0{,}043+1{,}5706+0{,}8033+0{,}4737=4{,}58.$$

По таблице квантелей χ^2-распределения (см. табл. П.3) находим $\chi^2_{0{,}95}(4)\approx 9{,}49$. Так как $4{,}58<9{,}49$, то гипотеза H_0 принимается.

5.2. Критерии согласия. Сложная гипотеза

Критерии Колмогорова и ω^2 для сложной гипотезы. Задача проверки *простой гипотезы* о виде закона распределения случайной величины X на практике встречается довольно редко. Гораздо чаще бывает необходимо проверить по *случайной выборке* $\vec{X}_n$ из *генеральной совокупности* X *сложную гипотезу* о принадлежности функции распределения $F(t)$ случайной величины X заданному *параметрическому множеству* распределений $\{F(t;\theta),\ \theta\in\Theta\}$, $\Theta\subset\mathbb{R}^d$:

$$H_0:\ F(t)=F_0(t;\theta),\quad \theta\in\Theta.$$

Кажется естественным сначала каким-то образом построить оценку $\widehat{\theta}(\vec{X}_n)$ параметра θ, а затем применить *критерии Колмогорова* и ω^2 для проверки *гипотезы*

$$H_0:\ F(t)=F_0(t;\widehat{\theta}(\vec{x}_n)),$$

где $\widehat{\theta}(\vec{x}_n)$ — значение оценки $\widehat{\theta}(\vec{X}_n)$ по данным *выборки* $\vec{x}_n$. К сожалению, при таком подходе эти *критерии* уже не будут

непараметрическими — при гипотезе H_0 распределение модифицированных *статистик* $\widehat{D}(\vec{X}_n)$ и $\widehat{\omega}^2(\vec{X}_n)$, где

$$\widehat{D}(\vec{x}_n) = \sup_t |F_n(t) - F_0(t;\widehat{\theta}(\vec{x}_n))|,$$

$$\widehat{\omega}^2(\vec{x}_n) = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\Big(F_0\big(x_{(i)};\widehat{\theta}(\vec{x}_n)\big) - \frac{2i-1}{2n}\Big)^2,$$

вообще говоря, зависит от F_0 и от метода нахождения оценки $\widehat{\theta}(\vec{X}_n)$, что требует составления большого количества таблиц распределений.

Однако если $\widehat{\theta}(\vec{X}_n)$ — *оценки максимального правдоподобия* параметра θ, а элементы $F(t;\theta)$ параметрического множества $\{F(t;\theta),\ \theta \in \Theta\}$ функций распределений получаются при помощи преобразования сдвига и масштаба какого-нибудь одного своего представителя $F(t;\theta_0)$, т.е.

$$F(t;\theta) = F\Big(\frac{t-a}{b};\theta_0\Big),$$

то для критериев Колмогорова и ω^2 достаточно иметь только одну таблицу для каждого семейства. К таким семействам относятся все важные типы распределений, и, в частности, нормальное. Более того, при небольшой модификации статистик $\widehat{D}(\vec{X}_n)$ и $\widehat{\omega}^2(\vec{X}_n)$ их распределение при $n \geqslant 5$ практически перестает зависеть* от n.

Критерий χ^2 для сложной гипотезы. Пусть функция распределения дискретной случайной величины X, принимающей конечное множество значений $u_1, \ldots, u_r$, зависит от d-мерного вектора параметров θ. Тогда вероятность p_k того, что X примет возможное значение u_k, зависит от θ, т.е. $p_k = p_k(\theta)$, $k = \overline{1,r}$. А так как вероятности $p_1(\theta), \ldots, p_r(\theta)$ полностью определяют функцию распределения случайной величины

*См.: *Тюрин Ю.Н., Макаров А.А.*

X, то в рассматриваемом случае *основная гипотеза* принимает следующий вид:

$$H_0\colon\ Pr X = u_k = p_k(\theta),\quad k=\overline{1,r},\quad \theta\in\Theta\subset\mathbb{R}^d.$$

Эту сложную гипотезу можно проверить при помощи модификации критерия χ^2 Пирсона.

Пусть $\widehat{\theta}(\vec{x}_n)$ — значение оценки $\widehat{\theta}(\vec{X}_n)$ максимального правдоподобия для θ, а $n_k(\vec{x}_n)$ — количество элементов выборки $\vec{x}_n$, равных u_k, $k=\overline{1,r}$. Оценку $\widehat{\theta}(\vec{X}_n)$ получают в результате минимизации логарифма *функции правдоподобия*

$$L(\vec{X}_n;\theta)=\frac{n!}{n_1!\dots n_r!}\prod_{k=1}^{r}p_i^{n_k(\vec{X}_n)}(\theta),\quad \sum_{i=1}^{r}n_i(\vec{X}_n)=n,$$

как (см. (3.2)) решение системы уравнений

$$\sum_{k=1}^{r}\frac{n_k(\vec{X}_n)}{p_k(\theta)}\frac{\partial p_k(\theta)}{\partial\theta_j}=0,\quad j=\overline{1,d}.$$

Можно показать*, что при некоторых предположениях о гладкости функций $p_k(\theta)$, $k=\overline{1,r}$, распределение случайной величины при $n\to\infty$

$$\chi^2(\vec{X}_n)=\sum_{i=1}^{r}\frac{\left(n_i(\vec{X}_n)-np_i(\widehat{\theta}(\vec{X}_n))\right)^2}{np_i(\widehat{\theta}(\vec{X}_n))}$$

слабо сходится к случайной величине, имеющей *χ^2-распределение* с $r-d-1$ степенями свободы.

Если X — непрерывная случайная величина с функцией распределения $F(t)$, то, разбивая множество возможных значений X на конечное число непересекающихся подмножеств и переходя к дискретной случайной величине X', можно проверить сложную гипотезу

$$H_0\colon\ F(t)\in\{F(t;\theta),\,\theta\in\Theta\subset\mathbb{R}^d\}.$$

*См.: *Крамер Г.*

Необходимо только помнить, что оценку максимального правдоподобия $\widehat{\theta}(\vec{X}_n)$ следует строить не по наблюдениям $X_1, \ldots, X_n$ случайной величины X, а по значениям частот $n_1(\vec{x}'_n), \ldots, n_r(\vec{x}'_n)$ случайной величины X', что, как правило, гораздо труднее. Построение такой оценки для наиболее распространенных параметрических семейств распределений (нормального, экспоненциального, пуассоновского и т.д.) можно найти в специальной литературе*.

Двухвыборочная задача. Критерий Смирнова. Пусть $\vec{X}_m = (X_1, \ldots, X_m)$ и $\vec{Y}_n = (Y_1, \ldots, Y_n)$ — случайные выборки из генеральных совокупностей X и Y с функциями распределения $F(t)$ и $G(t)$ соответственно. Рассмотрим задачу проверки сложной гипотезы

$$H_0\colon\ F(t) = G(t), \quad t \in \mathbf{R}, \tag{5.10}$$

против *альтернативной гипотезы*

$$H_1\colon\ F(t) \neq G(t) \quad \text{для некоторых} \quad t \in \mathbf{R}. \tag{5.11}$$

Для непрерывных случайных величин X и Y гипотезу H_0 против альтернативной гипотезы H_1 можно проверить, воспользовавшись статистикой $D(\vec{X}_m, \vec{Y}_n)$, реализация которой определяется формулой

$$D(\vec{x}_m, \vec{y}_n) = \sup_t \bigl|F_m(t) - G_n(t)\bigr|, \tag{5.12}$$

где $F_m(t)$ и $G_n(t)$ — *эмпирические функции распределения*, построенные по *реализациям* $\vec{x}_m$ и $\vec{y}_n$ *случайных выборок* $\vec{X}_m$ и $\vec{Y}_n$ соответственно. Если истинной является основная гипотеза H_0, то, согласно закону больших чисел, для любого $t \in \mathbf{R}$

$$\lim_{m,n\to\infty} \bigl(F_m(t) - G_n(t)\bigr) = F(t) - G(t) = 0.$$

*См.: *Крамер Г.*

Если же истинной является альтернативная гипотеза H_1, то для любого $t \in \mathbb{R}$

$$\lim_{m,n\to\infty} \bigl(F_m(t) - G_n(t)\bigr) = F(t) - G(t) \neq 0.$$

Следовательно, значения $D(\vec{x}_m, \vec{y}_n)$, близкие к нулю, свидетельствуют о том, что, по-видимому, верна гипотеза H_0, а большие значения $D(\vec{x}_m, \vec{y}_n)$ указывают на большую правдоподобность гипотезы H_1. На этом факте и основан ***критерий Смирнова***. А именно критерий Смирнова отклоняет H_0 в пользу H_1 на *уровне значимости* α, если *выборочное значение* $D(\vec{x}_m, \vec{y}_n)$ статистики $D(\vec{X}_m, \vec{Y}_n)$ удовлетворяет неравенству

$$D(\vec{x}_m, \vec{y}_n) > D_{1-\alpha},$$

где $D_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения статистики $D(\vec{X}_m, \vec{Y}_n)$ при истинности гипотезы H_0. Если

$$D(\vec{x}_m, \vec{y}_n) \leqslant D_{1-\alpha},$$

то отклонять гипотезу H_0 нет оснований.

Доказано*, что распределение статистики $D(\vec{X}_m, \vec{Y}_n)$ при истинности H_0 не зависит от $F(t)$ и $G(t)$. Для небольших m и n таблицы квантилей функции распределения случайной величины $D(\vec{X}_m, \vec{Y}_n)$ при истинности H_0 есть в соответствующих справочниках**. Н.В. Смирнов*** показал, что при $m, n \to \infty$

$$\mathbf{P}\left\{\sqrt{\frac{mn}{m+n}}\, D(\vec{X}_m, \vec{Y}_n) < t\right\} \to K(t), \quad t > 0, \tag{5.13}$$

где функция $K(t)$ определена равенством (5.4).

Рассмотрим *вариационный ряд*

$$Z_{(1)}; \quad Z_{(2)}; \quad \ldots; \quad Z_{(N)}, \quad N = m + n, \tag{5.14}$$

*См.: *Смирнов Н.В.*

**См., например: *Большев Л.Н., Смирнов Н.В.*

***См.: *Смирнов Н.В.*

объединенной *случайной выборки* $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. Можно показать, что

$$D(\vec{X}_m, \vec{Y}_n) = \max_{i=\overline{1,N}} \left| F_m(Z_{(i)}) - G_n(Z_{(i)}) \right|.$$

Значение $D(\vec{x}_m, \vec{y}_n)$ статистики $D(\vec{X}_m, \vec{Y}_n)$ удобнее вычислять следующим образом. Пусть

$$\delta_i = \begin{cases} 1, & z_{(i)} \text{ — одно из наблюдений } X; \\ 0, & z_{(i)} \text{ — одно из наблюдений } Y, \end{cases} \tag{5.15}$$

где $z_{(i)}$ — значение случайной величины $Z_{(i)}$, $i = \overline{1, N}$. Положим

$$s_j = \frac{jm}{N} - \sum_{k=1}^{j} \delta_k, \quad j = \overline{1, N}. \tag{5.16}$$

Тогда

$$D(\vec{x}_m, \vec{y}_n) = \frac{N}{mn} \max\{s_1, \ldots, s_N\}. \tag{5.17}$$

Пример 5.4. Пусть X и Y — непрерывные случайные величины с функциями распределения $F(t)$ и $G(t)$ соответственно. Даны выборка $\vec{x}_{10}$ с элементами

$-0{,}15;\ 8{,}60;\ 5{,}00;\ 3{,}71;\ 4{,}29;\ 7{,}74;\ 2{,}48;\ 3{,}25;\ -1{,}15;\ 8{,}38$

и выборка $\vec{y}_{10}$ с элементами

$2{,}55;\ 12{,}07;\ 0{,}46;\ 0{,}35;\ 2{,}69;\ -0{,}94;\ 1{,}73;\ 0{,}73;\ -0{,}35;\ -0{,}37.$

Проверим на уровне значимости $\alpha = 0{,}05$ гипотезу (5.10) против альтернативной гипотезы (5.11).

Выписываем значения объединенного вариационного ряда заданных выборок

$-1{,}15;\ -0{,}94;\ -0{,}37;\ -0{,}35;\ 0{,}46;\ 0{,}73;\ 1{,}73;\ 2{,}48;$
$2{,}55;\ 2{,}69;\ 3{,}25;\ 3{,}71;\ 4{,}29;\ 5{,}00;\ 7{,}74;\ 8{,}38;$
$8{,}60;\ 12{,}07$

и последовательность чисел δ_i, $i = \overline{1,20}$,

1; 0; 0; 0; 1; 0; 0; 0; 0; 1; 0; 0; 1; 1; 1; 1; 1; 1; 1; 0.

Вычислив по формуле (5.16) значения величин s_j, $j = \overline{1,20}$, и подставив их в (5.17), определим, что $D(\vec{x}_{10}, \vec{y}_{10}) = 6$. В таблице квантилей распределения статистики* $D(\vec{X}_m, \vec{Y}_n)$ квантили $D_{1-\alpha} = D_{0,95}$ нет, но есть квантиль $D_{0,9476} = 6$. Поэтому гипотезу (5.10) следует отклонить в пользу альтернативной гипотезы (5.11) на уровне значимости $\alpha = 0{,}0524$.

5.3. Критерии независимости

Критерий Спирмена. Пусть имеется *случайная выборка* $(X_1, Y_1), \ldots, (X_n, Y_n)$ из *генеральной совокупности* двумерной непрерывной случайной величины (X, Y) с функцией распределения $F(t,\tau)$, а $F_X(t)$ и $F_Y(\tau)$ — функции распределения случайных величин X и Y соответственно. Если случайные величины X и Y имеют нормальные распределения, то для проверки *статистической гипотезы* об их независимости

$$H_0\colon\ F(t,\tau) = F_X(t)\,F_Y(\tau) \tag{5.18}$$

можно использовать процедуру, связанную с вычислениями *выборочного коэффициента корреляции* (см. формулу (6.12)).

Если же о распределениях непрерывных случайных величин X и Y ничего не известно, то для проверки *основной гипотезы* (5.18) при *альтернативной гипотезе*

$$H_1\colon\ F(t,\tau) \neq F_X(t)\,F_Y(\tau) \ \text{ для некоторых } (t,\tau) \in \mathbf{R}^2$$

используют ***ранговый критерий Спирмена***, основанный на следующем понятии.

Определение 5.1. ***Рангом*** $R_i(\vec{z}_N)$ ***элемента*** z_i числовой ***последовательности*** $\vec{z}_N = (z_1, \ldots, z_N)$ называют его порядковый номер в *вариационном ряду* $z_{(1)}, \ldots, z_{(N)}$.

*См.: *Большев Л.Н., Смирнов Н.В.*

Согласно определению, $R_i(\vec{z}_N)$ — это число элементов последовательности $z_1, \ldots, z_N$, не бóльших чем z_i, которое можно записать следующим образом:

$$R_i(\vec{z}_N) = 1 + \sum_{k=1}^{N} \eta(z_i - z_k),$$

где $\eta(t)$ — *функция Хевисайда*. Ранг любого элемента последовательности $\vec{z}_N$ — это натуральное число в диапазоне от 1 до N, причем ранг наименьшего элемента последовательности равен 1, а ранг наибольшего — N.

Пример 5.5. Рассмотрим выборку $\vec{z}_4 = (3{,}8,\ 4{,}7,\ -2{,}6,\ 17{,}3)$. Ее вариационный ряд имеет вид $-2{,}6;\ 3{,}8;\ 4{,}7;\ 17{,}3$. Поэтому $R_1(\vec{z}_4) = 2$, $R_2(\vec{z}_4) = 3$, $R_3(\vec{z}_4) = 1$, $R_4(\vec{z}_4) = 4$. #

Определение 5.2. ***Рангом элемента*** Z_i ***случайной выборки*** $\vec{Z}_N = (Z_1, \ldots, Z_N)$ называют случайную величину $R_i(\vec{Z}_N)$, реализация которой $R_i(\vec{z}_N)$ есть ранг реализации z_i случайной величины Z_i в вариационном ряду $z_{(1)}, \ldots, z_{(N)}$.

Обозначим через $R_i = R_i(\vec{X}_n)$ — ранг элемента X_i случайной выборки $X_1, \ldots, X_n$, а через $S_i = S_i(\vec{Y}_n)$ — ранг элемента Y_i случайной выборки $Y_1, \ldots, Y_n$.

Ранговым коэффициентом корреляции Спирмена назовем случайную величину

$$\rho(\vec{X}_n, \vec{Y}_n) = \frac{\sum\limits_{i=1}^{n} (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum\limits_{i=1}^{n} (R_i - \overline{R})^2 \sum\limits_{i=1}^{n} (S_i - \overline{S})^2}}, \qquad (5.19)$$

где

$$\overline{R} = \frac{1}{n} \sum_{i=1}^{n} R_i, \qquad \overline{S} = \frac{1}{n} \sum_{i=1}^{n} S_i.$$

Статистика (5.19) является выборочным коэффициентом корреляции последовательностей рангов $R_1, \ldots, R_n$ и $S_1, \ldots, S_n$.

Согласно определению рангов R_i, S_i, $i = \overline{1,n}$,

$$\overline{R} = \overline{S} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2},$$

и можно показать, что

$$\rho(\vec{X}_n, \vec{Y}_n) = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - S_i)^2. \qquad (5.20)$$

Без ограничения общности можно считать, что значения пар наблюдений (x_i, y_i), $i = \overline{1,n}$, занумерованы в порядке возрастания их первых элементов, т.е. так, что выполняются неравенства

$$x_1 < x_2 < \ldots < x_n.$$

В этом случае реализация r_i ранга R_i равна i, $i = \overline{1,n}$, и значение $\rho(\vec{x}_n, \vec{y}_n)$ статистики $\rho(\vec{X}_n, \vec{Y}_n)$ можно вычислить по формуле

$$\rho(\vec{x}_n, \vec{y}_n) = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(i - s_i)^2, \qquad (5.21)$$

где s_i — реализация ранга S_i, $i = \overline{1,n}$.

Известно (см. **6.3**), что выборочный коэффициент корреляции приспособлен для обнаружения линейной зависимости между случайными величинами X и Y. Если же между X и Y имеется функциональная, но не линейная зависимость, то выборочный коэффициент корреляции может быть равен нулю. Примерно так же обстоит дело и с ранговым коэффициентом (5.19), с тем только (впрочем, немаловажным) различием, что он улавливает любую монотонную зависимость, а не только линейную.

Доказательство этого начнем с исследования статистики $\rho(\vec{X}_n, \vec{Y}_n)$ при линейной зависимости $Y = aX + b$, $a \in \mathbb{R}$, $b \in \mathbb{R}$, между случайными величинами X и Y.

Если $a > 0$, то бóльшим значениям x_i соответствуют бóльшие значения y_i, и, наоборот, меньшим значениям x_i — меньшие значения y_i, $i = \overline{1, n}$. Если пары наблюдений (x_i, y_i), $i = \overline{1, n}$, занумерованы по возрастанию первых элементов, то будут иметь место неравенства $y_1 < \ldots < y_n$. Поэтому $r_i = s_i$ для всех $i = \overline{1, n}$, и из (5.21) следует, что $\rho(\vec{x}_n, \vec{y}_n) = 1$.

Если же $a < 0$, то бóльшим значениям x_i соответствуют меньшие значения y_i, а меньшим значениям x_i — бóльшие значения y_i, $i = \overline{1, n}$. В этом случае $r_i = s_{n-i+1}$, $s_i = r_{n-i+1}$, $i = \overline{1, n}$, и $\rho(\vec{x}_n, \vec{y}_n) = -1$.

Заметим, что если $\varphi(x)$ — возрастающая функция, то ранг элемента x_i в последовательности $x_1, \ldots, x_n$ равен рангу $\varphi(x_i)$ в последовательности $\varphi(x_1), \ldots, \varphi(x_n)$. Поэтому если случайные величины X и Y связаны функциональной зависимостью $Y = \varphi(X)$, то $\rho(\vec{x}_n, \vec{y}_n) = 1$.

Аналогично, если $Y = \varphi(X)$, где $\varphi(x)$ — убывающая функция, то $\rho(\vec{x}_n, \vec{y}_n) = -1$.

Условие $|\rho(\vec{X}_n, \vec{Y}_n)| \leqslant 1$ выполняется всегда, так как оно выполняется для выборочного коэффициента корреляции, а $\rho(\vec{X}_n, \vec{Y}_n)$ — это выборочный коэффициент корреляции, построенный по последовательностям рангов наблюдений.

Рассмотрим теперь другой крайний случай, когда случайные величины X и Y независимы, т.е. когда основная гипотеза H_0 является истинной. В этой ситуации случайный вектор $(S_1, \ldots, S_n)$ принимает с равной вероятностью любое свое возможное значение, являющееся одной из $n!$ перестановок, составленной из чисел $1, 2, \ldots, n$. Следовательно, вероятность того, что статистика $\rho(\vec{X}_n, \vec{Y}_n)$ примет любое из своих возможных значений $\rho(\vec{x}_n, \vec{y}_n)$ при истинности основной гипотезы (5.18), не зависит от распределений случайных величин X и Y.

Можно показать, что при истинности основной гипотезы (5.18)

$$\mathbf{M}\,\rho(\vec{X}_n,\vec{Y}_n)=0,\qquad \mathbf{D}\,\rho(\vec{X}_n,\vec{Y}_n)=\frac{1}{n-1},\tag{5.22}$$

и, следовательно, при этом *выборочные значения* статистики $\rho(\vec{X}_n,\vec{Y}_n)$ невелики и группируются около нуля. Поэтому (и это кажется достаточно естественным) ранговый критерий Спирмена отклоняет H_0 на *уровне значимости* α, если

$$|\rho(\vec{x}_n,\vec{y}_n)|>\rho_{1-\alpha/2},$$

где $\rho_{1-\alpha/2}$ — квантиль уровня $1-\alpha/2$ распределения случайной величины $\rho(\vec{X}_n,\vec{Y}_n)$ при истинности основной гипотезы (5.18). При небольших n это распределение табулировано*. Известно, что при $n\to\infty$ и при истинности основной гипотезы (5.18)

$$\lim_{n\to\infty}\mathbf{P}\left\{\frac{\rho(\vec{X}_n,\vec{Y}_n)-\mathbf{M}\,\rho(\vec{X}_n,\vec{Y}_n)}{\sqrt{\mathbf{D}\,\rho(\vec{X}_n,\vec{Y}_n)}}<t\right\}=\Phi_0(t)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{t}e^{-\frac{u^2}{2}}\,du,$$

т.е. квантили случайной величины $\rho(\vec{X}_n,\vec{Y}_n)$ можно приближенно вычислять при помощи таблиц квантилей стандартного нормального распределения.

Пример 5.6. В табл. 5.1 представлены $n=10$ значений $(x_i,\,y_i)$, $i=\overline{1,10}$, непрерывной двумерной случайной величины $(X,\,Y)$. Проверим на уровне значимости $\alpha=0{,}05$ гипотезу H_0 о независимости случайных величин X и Y.

Таблица 5.1

x_i	−1,63	1,11	1,15	−1,93	0,38	−1,08	−0,31	0,60	0,12	0,92
y_i	0,54	0,88	−1,21	0,89	−0,64	−0,21	0,08	−0,74	0,79	0,14

*См.: *Большев Л.Н., Смирнов Н.В.*

Строим последовательность рангов (табл. 5.2). По формуле (5.20) вычисляем реализацию статистики $\rho(\vec{X}_n, \vec{Y}_n)$

$$\rho(\vec{x}_n, \vec{y}_n) = 1 - \frac{6}{10(10^2-1)}\Big((2-7)^2 + (9-9^2) + (10-1)^2 + (1-10)^2 +$$

$$+ (6-3)^2 + (3-4)^2 + (4-5)^2 + (7-2)^2 + (5-8)^2 + (8-6)^2\Big) =$$

$$= 1 - \frac{6}{990}\Big(25 + 0 + 81 + 81 + 9 + 1 + 1 + 25 + 9 + 2\Big) \approx -0{,}4118.$$

Таблица 5.2

r_i	2	9	10	1	6	3	4	7	5	8
s_i	7	9	1	10	3	4	5	2	8	6

По таблицам распределения статистики $\rho(\vec{X}_n, \vec{Y}_n)$ рангового критерия Спирмена* находим квантили

$$\rho_{0{,}952} = 0{,}6726, \qquad \rho_{0{,}97} = 0{,}7374, \qquad \rho_{0{,}983} = 0{,}80223, \quad (5.23)$$

а квантили $\rho_{1-\alpha/2} = \rho_{0{,}975}$ нет, так как $\rho(\vec{X}_n, \vec{Y}_n)$ — дискретная случайная величина. Тем не менее, из значений квантилей (5.23) заключаем, что $|\rho(\vec{x}_n, \vec{y}_n)| < \rho_{0{,}952}$ и H_0 не отклоняется даже на большем уровне значимости.

Таблицы сопряженности признаков и критерий χ^2. Пусть имеется случайная выборка

$$(\vec{X}_n, \vec{Y}_n) = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$$

из генеральной совокупности двумерной дискретной случайной величины (X, Y), где случайная величина X может принимать значения $u_1, \ldots, u_r$, а случайная величина Y — значения $v_1, \ldots, v_s$. Определим случайную величину $n_{ij}(\vec{X}_n, \vec{Y}_n)$, реализация n_{ij} которой равна количеству элементов выборки $(\vec{x}_n, \vec{y}_n) = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$, совпадающих с элементом (u_i, v_j), $i = \overline{1, r}$, $j = \overline{1, s}$.

*См.: *Большев Л.Н., Смирнов Н.В.*

Введем случайные величины $n_{i\cdot}(\vec{X}_n, \vec{Y}_n)$ и $n_{\cdot j}(\vec{X}_n, \vec{Y}_n)$, значения $n_{i\cdot}$ и $n_{\cdot j}$ которых определим по формулам

$$n_{i\cdot} = \sum_{j=1}^{n} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{n} n_{ij}.$$

При этом $n_{i\cdot}$ — количество элементов выборки $(\vec{x}_n, \vec{y}_n)$, в которых встретилось значение u_i, а $n_{\cdot j}$ — количество элементов выборки $(\vec{x}_n, \vec{y}_n)$, в которых встретилось значение v_j. Кроме того, имеют место очевидные равенства

$$\sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{s} n_{\cdot j} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = n.$$

В рассматриваемом случае результаты наблюдений удобно оформлять в виде таблицы, называемой ***таблицей сопряженности признаков*** (табл. 5.3).

Таблица 5.3

X	Y				
	v_1	v_2	...	v_s	
u_1	n_{11}	n_{12}	...	n_{1s}	$n_{1\cdot}$
u_2	n_{21}	n_{22}	...	n_{2s}	$n_{2\cdot}$
...	...	...	...	...	...
u_r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot s}$	n

Пусть далее

$$p_{ij} = \mathbf{P}\{X = u_i, Y = v_j\}, \quad p_{i\cdot} = \mathbf{P}\{X = u_i\}, \quad p_{\cdot j} = \mathbf{P}\{Y = v_j\},$$
$$i = \overline{1, r}, \quad j = \overline{1, s}.$$

Дискретные случайные величины X и Y независимы тогда и только тогда, когда

$$\mathbf{P}\{X = u_i, Y = v_j\} = \mathbf{P}\{X = u_i\}\,\mathbf{P}\{Y = v_j\}, \quad i = \overline{1, r}, \quad j = \overline{1, s}.$$

Поэтому основную гипотезу о независимости дискретных случайных величин X и Y можно представить в следующем виде:

$$H_0\colon\ p_{ij} = p_{i\cdot}p_{\cdot j},\ \ i = \overline{1,r},\ \ j = \overline{1,s}. \tag{5.24}$$

При этом, как правило, в качестве альтернативной используют гипотезу

$$H_1\colon\ p_{ij} \neq p_{i\cdot}p_{\cdot j} \text{ для некоторых } i = \overline{1,r},\ j = \overline{1,s}. \tag{5.25}$$

Для проверки основной гипотезы (5.24) при альтернативной гипотезе (5.25) К. Пирсон предложил использовать статистику $\widehat{\chi}^2(\vec{X}_n, \vec{Y}_n)$, называемую ***статистикой Фишера — Пирсона***, реализация $\widehat{\chi}^2(\vec{x}_n, \vec{y}_n)$ которой определяется формулой

$$\widehat{\chi}^2(\vec{x}_n, \vec{y}_n) = n \sum_{i=1}^{r} \sum_{i=1}^{s} \frac{\left(n_{ij} - \dfrac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{n_{i\cdot}n_{\cdot j}}. \tag{5.26}$$

Из закона больших чисел следует, что при $n \to \infty$

$$\frac{n_{ij}(\vec{X}_n, \vec{Y}_n)}{n} \to p_{ij}, \quad \frac{n_{i\cdot}(\vec{X}_n, \vec{Y}_n)}{n} \to p_{i\cdot}, \quad \frac{n_{\cdot j}(\vec{X}_n, \vec{Y}_n)}{n} \to p_{\cdot j},$$

$$i = \overline{1,r}, \quad j = \overline{1,s}.$$

Поэтому при истинности гипотезы H_0 и больших объемах выборки $(\vec{x}_n, \vec{y}_n)$ должно выполняться приближенное равенство

$$n_{ij} \approx n_{i\cdot}n_{\cdot j}, \quad i = \overline{1,r},\ \ j = \overline{1,s},$$

и, следовательно, значения (5.26) статистики $\widehat{\chi}^2(\vec{X}_n, \vec{Y}_n)$ должны быть „не слишком велики". „Слишком большие" значения должны свидетельствовать о том, что H_0 неверна.

Ответ на вопрос о том, какие значения нужно считать слишком большими, а какие — нет, дает следующая теорема.

Теорема 5.3. Если истинна гипотеза H_0, то распределение статистики $\widehat{\chi}^2(\vec{X}_n, \vec{Y}_n)$ при $n \to \infty$ слабо сходится к случайной

величине, имеющей χ^2-распределение с числом степеней свободы $k=(r-1)(s-1)$:

$$\lim_{n\to\infty}\mathbf{P}\{\widehat{\chi}^2(\vec{X}_n,\vec{Y}_n)<z\}=\int\limits_0^z\frac{t^{\frac{k}{2}-1}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}e^{-\frac{t}{2}}\,dt,\quad z>0.\quad\#$$

В соответствии с теоремой 5.3 ***критерий независимости*** χ^2 отклоняет гипотезу H_0 на уровне значимости $1-\alpha$, если

$$\widehat{\chi}^2(\vec{x}_n,\vec{y}_n)>\chi^2_{1-\alpha}((r-1)(s-1)),$$

где $\chi^2_{1-\alpha}((r-1)(s-1))$ — квантиль уровня значимости $1-\alpha$ χ^2-распределения с числом степеней свободы $(r-1)(s-1)$. При этом считается*, что критерий χ^2 можно использовать, если $n_{i\cdot}n_{\cdot j}/n\geqslant 5$.

Правую часть равенства (5.26) можно преобразовать к форме, более удобной для практического использования:

$$\widehat{\chi}^2(\vec{x}_n,\vec{y}_n)=n\Big(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\Big).\tag{5.27}$$

В частном, но очень распространенном случае таблиц сопряженности при $r=s=2$ формула (5.26) для вычисления $\widehat{\chi}^2(\vec{x}_n,\vec{y}_n)$ имеет еще более простой вид:

$$\widehat{\chi}^2(\vec{x}_n,\vec{y}_n)=\frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}.\tag{5.28}$$

Заметим, что для таблиц сопряженности при $r=s=2$, как правило, используют статистику $\widetilde{\chi}^2(\vec{X}_n,\vec{Y}_n)$ с реализациями

$$\widetilde{\chi}^2(\vec{x}_n,\vec{y}_n)=\frac{\big(n|n_{11}n_{22}-n_{12}n_{21}|-n/2\big)^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}},\tag{5.29}$$

*См.: *Тюрин Ю.Н., Макаров А.А.*

называемую ***статистикой Фишера — Пирсона с поправкой Йейтса на непрерывность***, распределение которой лучше согласуется с χ^2-распределением.

Пример 5.7. В табл. 5.4 приведены результаты 145 наблюдений двумерного дискретного случайного вектора (X, Y). Проверим на уровне $\alpha = 0{,}05$ гипотезу H_0 о независимости случайных величин X и Y.

Таблица 5.4

X	Y			
	3	4	5	
0	45	25	15	85
1	11	11	13	35
2	9	9	7	25
	65	45	35	145

В рассматриваемом случае $r = 3$, $s = 3$, т.е. случайные величины X и Y принимают по три различных значения. Вычислим по формуле (5.27) значение $\hat{\chi}^2(\vec{x}_n, \vec{y}_n)$ величины $\hat{\chi}^2(\vec{X}_n, \vec{Y}_n)$:

$$\hat{\chi}^2(\vec{x}_n, \vec{y}_n) = 145\left(\frac{45^2}{65\cdot 85} + \frac{25^2}{45\cdot 85} + \frac{15^2}{35\cdot 85} + \frac{11^2}{65\cdot 35} + \right.$$

$$\left. + \frac{11^2}{45\cdot 35} + \frac{13^2}{35\cdot 35} + \frac{9^2}{65\cdot 25} + \frac{9^2}{45\cdot 25} + \frac{7^2}{35\cdot 25} - 1\right) =$$

$$= 145\big(0{,}3665 + 0{,}1634 + 0{,}0756 + 0{,}0532 +$$

$$+ 0{,}0768 + 0{,}1380 + 0{,}0498 + 0{,}072 + 0{,}056 - 1\big) =$$

$$= 145 \cdot 0{,}0513 = 7{,}4385.$$

По таблице квантилей χ^2-распределения (см. табл. П.3) с числом степеней свободы $(r-1)(s-1) = 4$ находим

$$\chi^2_{1-\alpha}((r-1)(s-1)) = \chi^2_{0{,}95}(4) = 9{,}49.$$

Таким образом, оснований для отклонения гипотезы H_0 о независимости случайных величин X и Y недостаточно.

5.4. Решение типовых примеров

Пример 5.8. Даны *выборка объема* $m = 25$

0,00;	−0,53;	1,47;	0,96;	3,98;	3,22;	0,25;
0,31;	−0,64;	−1,26;	−0,92;	−1,36;	0,96;	1,39;
−0,81;	1,12;	−0,62;	−0,66;	1,07;	−0,52;	0,48;
−1,00;	−0,96;	−1,43;	−1,09			

из *распределения Коши* с плотностью

$$p_X(x) = \frac{1}{\pi(1+x^2)}$$

и выборка объема $n = 28$

−0,88;	0,41;	−0,64;	−0,81;	−0,09;	−0,71;	−0,00;
0,49;	−0,65;	0,59;	0,17;	−0,46;	0,99;	−0,24;
−0,98;	−0,85;	−0,09;	−0,63;	0,68;	0,02;	−0,59;
−0,02;	−0,45;	−0,50;	0,40;	0,29;	−0,17;	−0,43

из равномерного распределения на отрезке $[-1, 1]$ с плотностью $p_Y(x)$. Проверим при помощи *критерия Смирнова статистическую гипотезу* о равенстве функций p_X и p_Y.

Объединив заданные выборки и построив вариационный ряд, по формуле (5.15) найдем соответствующие этому ряду значения δ_i, $i = \overline{1, N}$, $N = 45$:

0; 0; 0; 0; 0; 1; 0; 0; 1; 1; 1; 0; 1; 0; 1; 1; 0; 1; 0; 1;
0; 0; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 0; 1; 1; 0; 1; 0; 1; 1;
0; 1; 1; 1; 0; 0; 1; 0; 0; 0; 0; 0; 0.

Вычислив по формуле (5.16) значения s_j, $j = \overline{1, N}$, по формуле (5.17) получим $D(\vec{x}_n, \vec{y}_n) = 0{,}473$ и $\sqrt{\frac{mn}{m+n}} = 1{,}718$. Так как m и

n велики, то для проверки гипотезы H_0 об однородности воспользуемся асимптотической формулой (5.13), в соответствии с которой

$$\mathbf{P}\left\{\sqrt{\frac{25\cdot 28}{25+28}}D(\vec{x}_n,\vec{y}_n)>1{,}718\right\}\approx 0{,}004.$$

Поэтому гипотезу об однородности следует отклонить на *уровне значимости* $\alpha \geqslant 0{,}004$.

Пример 5.9. При 4040 бросаниях монеты Ж.Л.Л. Бюффон* получил 2048 выпадений „герба“ и 1992 выпадений „решки“. Совместимо ли это с гипотезой о том, что вероятность выпадения „герба“ при одном бросании равна 1/2?

Здесь $n=4040$, $r=2$, $n_1(\vec{x}_n)=2048$, $n_2(\vec{x}_n)=1992$, $p_{10}=p_{20}=0{,}5$, число степеней свободы $r-1=1$, и при $\alpha=0{,}05$ находим $\chi^2_{0{,}05}(1)=3{,}841$.

Проверим гипотезу H_0 о том, что вероятности p_1 и p_2 выпадения „герба“ и „решки“ равны 1/2. На основании (5.9) получаем

$$\chi^2(\vec{x}_n)=\frac{(2048-4040\cdot 0{,}5)^2}{4040\cdot 0{,}5}+\frac{(1992-4040\cdot 0{,}5)^2}{4040\cdot 0{,}5}=0{,}776.$$

Так как $0{,}776<3{,}841$, то статистические данные не противоречат гипотезе H_0.

Пример 5.10. В табл. 5.5 приведены данные о распределении цвета волос на голове и бровей у 46542 человек. Проверим на уровне значимости $\alpha=0{,}05$ гипотезу о независимости этих признаков.

Здесь $n=46592$, $r=s=2$, $n_{11}=30472$, $n_{12}=3238$, $n_{21}=3364$, $n_{22}=9468$, $n_{1\cdot}=33710$, $n_{2\cdot}=12832$, $n_{\cdot 1}=33836$, $n_{\cdot 2}=12706$, число степеней свободы $(r-1)(s-1)=1$. Из (5.28) получаем $\hat{\chi}^2(\vec{x}_n,\vec{y}_n)=19{,}288$. По таблице квантилей χ^2-распределения

*Ж.Л.Л. Бюффон (1707–1788) — французский естествоиспытатель.

Таблица 5.5

Цвет бровей	Цвет волос на голове		Сумма
	светлые	темные	
Светлые	30472	3238	33710
Темные	3364	9468	12832
Сумма	33836	12706	46542

(см. табл. П.3) находим $\chi^2_{0,95}(1) = 3{,}84$. Так как $19{,}288 > 3{,}84$, то гипотезу о независимости признаков следует отклонить.

Пример 5.11. Бегуны, ранги которых при построении по росту были $1, 2, \ldots, 10$, заняли на состязаниях следующие места:

$$6, \quad 5, \quad 1, \quad 4, \quad 2, \quad 7, \quad 8, \quad 10, \quad 3, \quad 9.$$

Существует ли зависимость между ростом спортсмена и быстротой бега?

Проверим основную гипотезу H_0 о независимости между ростом и скоростью бега. Полагая в формуле (5.21) $n = 10$, $s_1 = 6$, $s_2 = 5$, $s_3 = 1$, $s_4 = 4$, $s_5 = 2$, $s_6 = 7$, $s_7 = 8$, $s_8 = 10$, $s_9 = 3$, $s_{10} = 9$, находим $\widehat{\rho}(\vec{x}_n, \vec{y}_n) = 0{,}24$. По таблице распределения рангового коэффициента корреляции* для уровня значимости $\alpha = 0{,}05$ находим $\rho_{0,75} = 0{,}56$. Так как $0{,}24 < 0{,}56$, то оснований отклонить H_0 нет.

Вопросы и задачи

5.1. Какие критерии называются критериями согласия?

5.2. В чем состоит критерий Колмогорова проверки статистических гипотез?

5.3. Какую статистику используют для проверки гипотез при помощи критерия Колмогорова?

*См.: *Большев Л.Н., Смирнов Н.В.*

5.4. В чем состоит критерий ω^2 проверки гипотез?

5.5. Какую статистику используют для проверки гипотез при помощи критерия ω^2?

5.6. Какие гипотезы лучше проверять при помощи критерия Колмогорова, а какие — при помощи критерия ω^2?

5.7. Можно ли при помощи критериев Колмогорова и ω^2 проверять простые гипотезы о математическом ожидании нормального распределения в примерах 4.10, 4.11 и 4.13?

5.8. Как при помощи критерия χ^2 проверять гипотезу о виде распределения непрерывной случайной величины?

5.9. Можно ли при помощи критериев Колмогорова и ω^2 проверять сложные гипотезы о виде распределения?

5.10. Что называют рангом элемента последовательности, рангом элемента случайной последовательности?

5.11. Какими свойствами обладает ранговый коэффициент корреляции Спирмена?

5.12. В чем преимущества и недостатки рангового коэффициента корреляции Спирмена перед выборочным коэффициентом корреляции?

5.13. Какую статистику используют для проверки гипотезы о независимости дискретных случайных величин? По какому закону она распределена?

5.14. Что называют таблицей сопряженности признаков?

5.15. Можно ли при помощи рангового критерия и таблиц сопряженности признаков исследовать случайные объекты нечисловой природы?

5.16. Проверьте на уровне значимости $\alpha = 0{,}05$ при помощи критерия Колмогорова гипотезу о том, что выборка 2,1; $-0{,}6$; 0,2; 3,0; $-1{,}0$; 1,3 извлечена из распределения $N(1, 1)$?

Ответ: данные не противоречат гипотезе.

5.17. Решите предыдущую задачу при помощи критерия ω^2.

Ответ: данные не противоречат гипотезе.

5.18. В экспериментах с селекцией гороха Г.И.Мендель* наблюдал частоты появления различных видов семян при скрещивании растений с круглыми желтыми семенами и растений с морщинистыми зелеными семенами. Эти данные и значения теоретических вероятностей по теории наследственности приведены в табл. 5.6. Проверьте на уровне значимости $\alpha = 0{,}1$ гипотезу H_0 о согласовании частотных данных с теоретическими вероятностями.

Таблица 5.6

Виды семян	Частота	Вероятность
Круглые и желтые	315	9/16
Морщинистые и желтые	101	3/16
Круглые и зеленые	108	3/16
Морщинистые и зеленые	32	1/16

Ответ: Гипотеза принимается.

5.19. Решите задачу 4.27, не предполагая нормальность распределения контролируемого признака.

5.20. В таблице 5.7 для каждой из девяти партий сыра приведены его жирность (в процентах) и усредненные (по 80 опрошенным респондентам) результаты опроса вкусовых качеств сыра по шестибалльной системе („превосходно“ — 6 баллов, „очень хорошо“ — 5, „хорошо“ — 4, „так себе“ — 3, „плохо“ — 2, „неприемлемо“ — 1). Проверьте по результатам опроса гипотезу о связи жирности сыра и его вкусовых качеств на уровне значимости $\alpha = 0{,}05$.

Ответ: вкусовые качества сыра улучшаются с увеличением его жирности.

*Г.И.Мендель (1822–1884) — монах и австрийский естествоиспытатель.

Таблица 5.7

Партия	Жирность, %	Результат опроса
1	44,4	2,6
2	45,9	3,1
3	41,9	2,5
4	53,3	5,0
5	44,7	3,6
6	44,1	4,0
7	50,7	5,2
8	45,2	2,8
9	60,1	3,8

5.21. Из 300 абитуриентов, поступивших в институт, 97 человек имели оценку 5 в школе и получили оценку 5 на вступительных экзаменах по тому же предмету, причем только 18 человек имели оценку 5 и в школе, и на экзамене. С уровнем значимости 0,1 проверьте гипотезу о независимости оценок 5 в школе и на экзамене.

Ответ: гипотеза отклоняется.

6. ОСНОВЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА

6.1. Исходные понятия

При решении прикладных задач в различных областях человеческой деятельности, в том числе и в инженерной практике, исследователь нередко сталкивается с необходимостью установления факта существования функциональных или иных зависимостей между переменными величинами, которые могут быть и случайными. Для подтверждения сказанного рассмотрим несколько простейших примеров.

Пример 6.1. Пусть Y — величина износа (в мм) протектора шины на автомобилях определенного типа после 10000 км пробега, X_1 — величина нагрузки (в кг) на колесо автомобиля, X_2 — тип протектора (используются три типа протекторов). Если установить степень влияния X_1 и X_2 на Y, то можно дать рекомендации по продлению долговечности шины.

Пример 6.2. Пусть Y_1 — производительность химической установки (в т/ч), Y_2 — процент брака готовой продукции. Технолог предполагает, что на переменные Y_1 и Y_2 влияют в наибольшей степени такие технологические параметры, как: X_1 — влажность сырья (в %), X_2 — температура в реакторе установки, X_3 — содержание примеси (в %).

Как установить степень влияния контролируемых переменных X_1, X_2, X_3 на переменные Y_1 и Y_2? Если найти вид зависимости Y_1 и Y_2 от X_1, X_2, X_3, то можно выбрать оптимальный (т.е. наилучший в определенном смысле) технологический режим (при котором, например, процент брака будет минимальным при заданном уровне производительности).

Пример 6.3. Пусть Y — успеваемость студентов по некоторой дисциплине (измеряемая, например, средним баллом на экзамене). Деканат проводит обследование студентов данного вуза с целью установления наиболее значимых факторов, влияющих на Y. В результате предварительного анализа сделано предположение о том, что этими факторами могут быть: X_1 — время, затрачиваемое студентом на самостоятельную работу, X_2 — количество пропущенных занятий, X_3 — величина стипендии. Существует ли взаимосвязь между факторами X_1, X_2, X_3? В какой степени они оказывают влияние на успеваемость? #

Приведенные примеры далеко не полностью отражают возможные постановки задач рассматриваемого типа. Но даже их поверхностный анализ позволяет отметить следующее.

1. Зависимое переменное Y может быть случайной величиной, даже если переменные $X_1, \ldots, X_p$ таковыми не являются, так как значение Y определяется не только значениями переменных $X_1, \ldots, X_p$, которые исследователь выделил (по его мнению, они являются определяющими), но и многими другими неучтенными факторами, а также ошибками измерений. Это означает, что ***связь*** между $X_1, \ldots, X_p$ и Y является не функциональной, а ***стохастической*** — изменение переменных $X_1, \ldots, X_p$ влияет на значения переменного Y через изменение закона распределения случайной величины Y.

2. Некоторые переменные могут иметь количественный характер, а некоторые — качественный (см. пример 6.1).

3. Нас может интересовать либо зависимость переменного Y от переменных $X_1, \ldots, X_p$, либо взаимозависимость между несколькими переменными (не обязательно между всеми). Так, в примере 6.3 может существовать взаимозависимость между переменными X_1, X_2 и X_3.

Перечисленные особенности приводят к различным постановкам задач статистического исследования зависимостей, ко-

торые упрощенно можно классифицировать следующим образом:

1) ***задачи корреляционного анализа*** — задачи исследования наличия взаимосвязей между отдельными группами переменных;

2) ***задачи регрессионного анализа*** — задачи, связанные с установлением аналитических зависимостей между переменным Y и одним или несколькими переменными $X_1, \ldots, X_p$, которые носят количественный характер;

3) ***задачи дисперсионного анализа*** — задачи, в которых переменные $X_1, \ldots, X_p$ имеют качественный характер, а исследуется и устанавливается степень их влияния на переменное Y.

Анализу наличия взаимосвязей между отдельными группами переменных и посвящена эта глава. Задачи регрессионного и дисперсионного анализа рассмотрены в последующих главах (см. 7 и 8).

Кроме перечисленных типов задач выделяют и многие другие. Так, ковариационный анализ рассматривает одновременно и количественные и качественные переменные $X_1, \ldots, X_p$, ***конфлюентный анализ****** обобщает регрессионный на тот случай, когда переменные $X_1, \ldots, X_p$ и Y измеряют с ошибками, ***факторный анализ******* служит для выделения из множества исследуемых переменных $X_1, \ldots, X_p$ наиболее значимых********.

Для удобства дальнейших рассуждений обратимся к так называемой модели „черного ящика“ (рис. 6.1) как наиболее общей модели любой реальной системы, ассоциированной с понятием отображения $f\colon \vec{X} \to \vec{Y}$. На вход „черного ящика“ поступает входной сигнал — вектор $\vec{X}$, который посредством отображения f преобразуется в выходной сигнал — вектор $\vec{Y}$. При этом, в соответствии со сложившийся терминологией, $\vec{X} = (X_1, \ldots, X_p)$ — вектор ***входных переменных***, или вектор

*См.: *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.*, 1985.

**См.: Прикладная статистика. Классификация и снижение размерности / *С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин.*

***См., например: *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.*, 1985.

факторов; $\vec{Y} = (Y_1, \ldots, Y_m)$ — вектор ***выходных переменных***, или вектор ***откликов***; $\varepsilon = \vec{Y} - f(\vec{X})$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ — вектор ***случайных ошибок***, т.е. случайных переменных, отражающих влияние на переменные Y_i, $i = \overline{1, m}$, неучтенных факторов, а также случайных ошибок измерений анализируемых показателей.

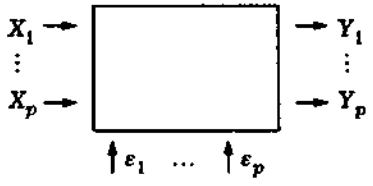


Рис. 6.1

При проведении корреляционного анализа исследователь должен уметь:

а) выбрать показатель стохастической связи анализируемых переменных;

б) оценить его значение по имеющимся *экспериментальным данным*, т.е. найти его *точечную* и *интервальную оценки*;

в) проверить *статистическую гипотезу* о том, что значение показателя стохастической связи значимо отличается от нуля.

Ниже дано описание методов и моделей, используемых для решения перечисленных задач.

6.2. Анализ парных связей

Выбор показателя связи. Для начала рассмотрим задачу выбора показателя *стохастической связи* между двумя случайными величинами* ξ и η, реализации которых будем обозначать соответственно через x и y.

*Использование новых обозначений (ξ и η вместо X и Y) связано с тем, что ξ и η могут выступать как в роли факторов, так и в роли откликов (или ξ может быть фактором, а η откликом).

Пример 6.4. Пусть случайный вектор (ξ,η) имеет нормальный закон распределения с математическим ожиданием $\mu=(\mu_1,\mu_2)$ и ковариационной матрицей

$$\Sigma=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix},$$

где σ_1^2 и σ_2^2 — дисперсии случайных величин ξ и η соответственно, а ρ — коэффициент корреляции между ξ и η.

В этом случае условная плотность распределения случайной величины η при условии, что $\xi=x$,

$$p(y|x)=\frac{1}{\sigma_{\eta/x}\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_{\eta/x}}{\sigma_{\eta/x}}\right)^2\right)$$

является плотностью нормального распределения [XVI] с параметрами $\mu_{\eta/x}$ (условное математическое ожидание) и $\sigma^2_{\eta/x}$ (условная дисперсия η) при значении $\xi=x$, которые связаны с параметрами исходного двумерного распределения следующим образом:

$$\mathbf{M}(\eta|\xi=x)=\mu_{\eta/x}=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \tag{6.1}$$

$$\mathbf{D}(\eta|\xi=x)=\sigma^2_{\eta/x}=\sigma_2^2(1-\rho^2). \tag{6.2}$$

В рассматриваемом случае линия регрессии является прямой, а условная дисперсия не зависит от x. #

Если закон распределения случайного вектора (ξ,η) не является нормальным, то характер изменения условного математического ожидания $\mathbf{M}(\eta|\xi=x)=f(x)$ может быть и нелинейным, причем, чем меньше условная дисперсия $\mathbf{D}(\eta|\xi=x)$, тем меньше при различных значениях x рассеяны возможные значения случайной величины η относительно линии регрессии $\mathbf{M}(\eta|\xi=x)=f(x)$ (рис. 6.2). Функцию $f(x)=\mathbf{M}(\eta|\xi=x)$ называют функцией регрессии, или регрессией.

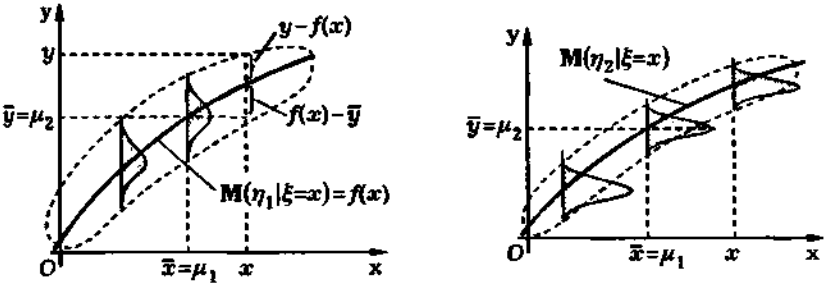


Рис. 6.2

Обозначим $\mathbf{M}\eta = \mu$, $\mathbf{D}\eta = \sigma_\eta^2$. Отклонение $y - \mu$ возможных значений η от μ складывается из двух слагаемых (см. рис. 6.2):

$$y - \mu = (f(x) - \mu) + (y - f(x)), \tag{6.3}$$

где $f(x) - \mu$ — отклонение функции регрессии $f(x)$ в точке x от математического ожидания μ; $y - f(x)$ — отклонение возможного значения η от значения функции регрессии в точке x.

Покажем, что рассеяние σ_η^2 случайной величины η относительно ее математического ожидания есть сумма двух слагаемых, а именно: математического ожидания квадрата отклонения η от ее условного математического ожидания $f(\xi)$ и математического ожидания квадрата отклонения $f(\xi)$ от μ.

Действительно [XVI],

$$\mathbf{M}\big(f(\xi)\big) = \mathbf{M}\big(\mathbf{M}(\eta|\xi)\big) = \mathbf{M}\eta = \mu,$$

$$\mathbf{D}\eta = \sigma_\eta^2 = \mathbf{M}(\eta - \mu)^2 = \mathbf{M}\Big((\eta - f(\xi)) + \big(f(\xi) - \mu\big)^2 =$$

$$= \mathbf{M}\big(\eta - f(\xi)\big)^2 + 2\mathbf{M}\big((\eta - f(\xi))(f(\xi) - \mu)\big) + \mathbf{M}\big(f(\xi) - \mu\big)^2 =$$

$$= \mathbf{M}\big(\eta - f(\xi)\big)^2 + \mathbf{M}\big(f(\xi) - \mu\big)^2,$$

так как $\mathbf{M}\big((\eta - f(\xi))(f(\xi) - \mu)\big) = 0$.

Докажем последнее равенство для непрерывных случайных величин ξ и η, предполагая, что их совместная плотность распределения $p(x,y)$ в $\mathbb{R}^2$ не обращается в нуль:

$$\mathrm{M}\big((\eta-f(\xi))(f(\xi)-\mu)\big)=\int\limits_{-\infty}^{\infty}\int(y-f(x))(f(x)-\mu)p(x,y)\,dx\,dy=$$

$$=\int\limits_{-\infty}^{\infty}(f(x)-\mu)p_\xi(x)\,dx\int\limits_{-\infty}^{\infty}(y-f(x))\frac{p(x,y)}{p_\xi(x)}\,dy=$$

$$=\int\limits_{-\infty}^{\infty}(f(x)-m)p_\xi(x)\,dx\left(\int\limits_{-\infty}^{\infty}y\frac{p(x,y)}{p_\xi(x)}\,dy-f(x)\int\limits_{-\infty}^{\infty}\frac{p(x,y)}{p_\xi(x)}\,dy\right)=0,$$

так как

$$\int\limits_{-\infty}^{\infty}y\frac{p(x,y)}{p_\xi(x)}\,dy=f(x),\qquad\int\limits_{-\infty}^{\infty}\frac{p(x,y)}{p_\xi(y)}\,dy=1.$$

Таким образом, если воспользоваться обозначениями

$$\sigma_f^2=\mathrm{D}\,f(\xi)=\mathrm{M}\big(f(\xi)-\mu\big)^2,\qquad\overline{\sigma}_\eta^2=\mathrm{M}\big(\eta-f(\xi)\big)^2,$$

то полученный результат может быть представлен в виде

$$\sigma_\eta^2=\sigma_f^2+\overline{\sigma}_\eta^2. \tag{6.4}$$

Из равенства (6.4) следует, что связь между ξ и η тем теснее, чем меньше слагаемое $\overline{\sigma}_\eta^2$ или чем больший вклад в дисперсию σ_η^2 вносит слагаемое σ_f^2, порожденное функцией регрессии $f(x)=\mathrm{M}(\eta)\xi=x$. Тем самым мы приходим к понятию общей характеристики степени тесноты связи — ***корреляционному отношению*** переменного η по переменному ξ:

$$r_{\eta\xi}=\sqrt{\frac{\sigma_f^2}{\sigma_\eta^2}}=\sqrt{1-\frac{\overline{\sigma}_\eta^2}{\sigma_\eta^2}}. \tag{6.5}$$

Непосредственно из (6.5) следует, что всегда выполняется неравенство

$$0 \leqslant r_{\eta\xi}^2 \leqslant 1, \tag{6.6}$$

причем равенство $r_{\eta\xi}^2 = 0$ означает, что с изменением ξ вариация функции регрессии $f(x)$ полностью отсутствует. Другими словами, случайные величины ξ и η являются независимыми. В этом случае линия регрессии есть горизонтальная прямая. Равенство $r_{\xi\eta}^2 = 1$ будет иметь место, если $\overline{\sigma}_\eta^2 = \mathrm{M}\big(\eta - f(\xi)\big)^2 = 0$, т.е. если η и ξ связаны функциональной зависимостью $\eta = f(\xi)$.

Аналогично определяется корреляционное отношение $r_{\xi\eta}$ переменной ξ по η.

Замечание 6.1. Между $r_{\eta\xi}$ и $r_{\xi\eta}$ нет какой-либо простой зависимости. Возможны ситуации, в которых один из этих показателей принимает нулевое значение, в то время как другой равен единице. Пусть, например, $\eta = \xi^2$, а ξ принимает следующие значения: -1, 0, 1 с вероятностями 1/3 каждое. В этом случае $r_{\eta\xi} = 1$, $r_{\xi\eta} = 0$ (в силу симметричности параболы относительно оси значений η и симметричности распределения ξ). #

Итак, решение задачи выбора показателя стохастической связи между двумя случайными величинами ξ и η для самой общей ситуации, когда закон распределения вектора (ξ, η) является произвольным, найдено — таким показателем являются корреляционные отношения $r_{\eta\xi}$ и $r_{\xi\eta}$.

Выясним, какую роль играет такой показатель связи между случайными величинами ξ и η, как коэффициент корреляции ρ:

$$\rho = \frac{\mathrm{M}\big((\xi - \mathrm{M}\xi)(\eta - \mathrm{M}\eta)\big)}{\sigma_1 \sigma_2}, \tag{6.7}$$

где $\sigma_1 = \sqrt{\mathrm{D}\xi}$, $\sigma_2 = \sqrt{\mathrm{D}\eta}$, $\mathrm{M}\big((\xi - \mathrm{M}\xi)(\eta - \mathrm{M}\eta)\big)$ — второй смешанный момент случайного вектора (ξ, η).

Напомним, что случайные величины ξ и η называют некоррелированными, если $\rho = 0$, и коррелированными при $\rho \neq 0$.

Известно [XVI], что из независимости случайных величин ξ и η следует их некоррелированность, однако обратное утверждение в общем случае неверно.

Если случайный вектор (ξ, η) имеет нормальный закон распределения, то линия регрессии η по ξ (и ξ по η) является прямой (см. пример 6.4), т.е. коэффициент корреляции ρ может служить мерой связи между ξ и η. Для нормального закона распределения на основании (6.2) и (6.5) имеем

$$r^2_{\eta\xi} = r^2_{\xi\eta} = \rho^2.$$

Действительно, из (6.2) получаем, что условная дисперсия η не зависит от значений случайной величины ξ, и, следовательно,

$$\sigma^2_\eta(1-\rho^2) = \mathbf{D}(\eta|\xi) = \mathbf{M}\Big((\eta - \mathbf{M}(\eta|\xi))^2\big|\xi\Big) =$$

$$= \mathbf{M}\Big((\eta - f(\xi))^2\big|\xi\Big) = \mathbf{M}(\eta - f(\xi))^2 = \overline{\sigma}^2_\eta.$$

Наконец, учитывая (6.5) и полученный результат, приходим к равенству $r^2_{\eta\xi} = \rho^2$. Аналогично можно доказать равенство $r^2_{\xi\eta} = \rho^2$. Таким образом, корреляционные отношения совпадают между собой и с абсолютной величиной коэффициента корреляции ρ. При этом равенство $|\rho| = 1$ означает линейную функциональную зависимость между ξ и η, а равенство $\rho = 0$ свидетельствует об их линейной независимости.

Понятно, что рассмотренными свойствами двумерного нормального закона не могут обладать все двумерные законы распределения или хотя бы их большая часть. Поэтому в общем случае не имеет смысла использование коэффициента корреляции ρ как меры взаимосвязи случайных величин ξ и η.

В общем случае показатели $r^2_{\eta\xi}$ и ρ^2 связаны неравенствами [XVI]

$$0 \leqslant \rho^2 \leqslant r^2_{\eta\xi} \leqslant 1. \qquad (6.8)$$

При этом возможны следующие варианты:

а) $\rho^2 = 0$, если ξ и η независимы, но обратное (в общем случае) неверно;

б) $\rho^2 = r^2_{\eta\xi} = 1$ тогда и только тогда, когда имеется строгая линейная функциональная зависимость η от ξ;

в) $\rho^2 \leqslant r^2_{\eta\xi} = 1$ тогда и только тогда, когда имеется строгая нелинейная функциональная зависимость η от ξ;

г) $\rho^2 = r^2_{\eta\xi} < 1$ тогда и только тогда, когда регрессия η по ξ строго линейна, но нет функциональной зависимости;

д) $\rho^2 < r^2_{\eta\xi} < 1$ указывает на то, что не существует функциональной зависимости, а некоторая нелинейная кривая регрессии „подходит" лучше, чем „наилучшая" прямая линия.

Итак, в качестве показателя стохастической связи между двумя случайными количественными переменными ξ и η следует выбрать корреляционное отношение $r_{\eta\xi}$ (или $r_{\xi\eta}$), если закон распределения вектора (ξ, η) вызывает сомнение. Если же можно с большой степенью уверенности считать закон распределения вектора (ξ, η) нормальным, то вместо корреляционного отношения следует использовать коэффициент корреляции ρ.

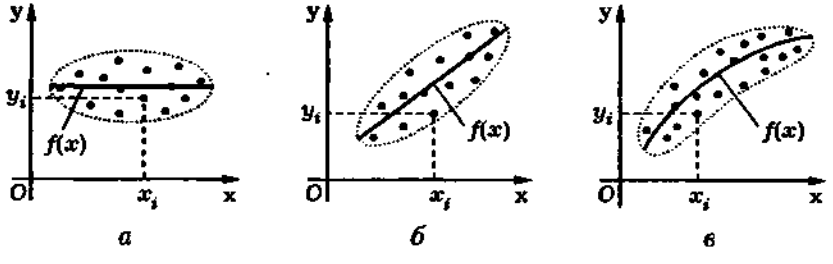
Оценка показателя связи по выборочным данным. После выбора показателя стохастической связи задача корреляционного анализа, как уже отмечалось в **6.1**, состоит в нахождении его *оценки* (*точечной* и *интервальной*), а также в проверке *статистической гипотезы* о значимом отличии его от нуля на основе *экспериментальных данных*.

Пусть в результате эксперимента получены n *выборочных значений* случайного вектора (ξ, η), которые будем записывать в виде

$$(x_i, y_i), \quad i = \overline{1, n}. \tag{6.9}$$

При изучении корреляционной зависимости двух случайных величин (ξ, η) по *выборке* (x_i, y_i), $i = \overline{1, n}$, общую картину их взаимной изменчивости можно получить, изобразив на координатной плоскости все точки. Это изображение называют ***корреляционным полем***.

Уже по виду корреляционного поля можно иногда сделать вывод о наличии и характере связи между случайными величинами ξ и η. Так, на рис. 6.3, *а* выборочные точки $(x_i,\ y_i)$ лежат внутри некоторого эллипса (эллипса рассеяния) с осями, параллельными координатным. Следовательно, с изменением, например, ξ величина η не будет менять своего условного распределения, т.е. ξ и η, по-видимому, некоррелированы. Напротив, на рис. 6.3, *б* видно, что условное математическое ожидание $\mathrm{M}(\eta|\xi = x) = f(x)$ имеет линейный характер изменения, и, значит, следует ожидать, что коэффициент корреляции ρ близок к единице. На рис. 6.3, *в* расположение точек $(x_i,\ y_i)$ говорит о наличии нелинейного характера изменения $f(x)$, и, следовательно, коэффициент корреляции может оказаться близким к нулю, а корреляционное отношение $r_{\eta\xi}$ — близким к единице.



а *б* *в*

Рис. 6.3

Следует отметить, что в том случае, когда среди x_i есть повторяющиеся с частотой n_i значения, выборочные значения представляют в виде

$$(x_i,\ y_{ij}),\quad j = \overline{1, n_i},\ \ i = \overline{1, m},\quad \sum_{i=1}^{m} n_i = n. \tag{6.10}$$

Если выборочные значения сгруппированы по каждой из переменных, т.е. значения x_i разделены на m групп, а значения y_i — на l групп, то выборочные значения представляют в виде

$$(x_i,\ y_j,\ n_{ij}),\quad i = \overline{1, m},\ \ j = \overline{1, l},\quad \sum_{i,j} n_{ij} = n, \tag{6.11}$$

или в виде ***корреляционной таблицы***, в каждой клетке которой указывают число n_{ij} попавших в нее выборочных значений, причем сумма всех этих значений равна n (табл. 6.1).

Таблица 6.1

Значения ξ	Значения η				
	y_1	...	y_j	...	y_l
x_1	n_{11}	...	n_{1j}	...	n_{1l}
...	...	...	...	...	...
x_i	n_{i1}	...	n_{ij}	...	n_{il}
...	...	...	...	...	...
x_m	n_{m1}	...	n_{mj}	...	n_{ml}

6.3. Анализ коэффициента корреляции

Точечная оценка показателя ρ. Пусть *экспериментальные данные* представлены в форме (6.9). Тогда $\widehat{\rho}$ — значение *точечной оценки* коэффициента корреляции ρ — вычисляют по формуле

$$\widehat{\rho} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}}. \qquad (6.12)$$

Пример 6.5. Вычислим значение $\widehat{\rho}$ для пары случайных величин (ξ, η), где ξ — рост (в см), а η — масса тела (в кг) наугад выбранного студента-первокурсника. Выборка объема $n = 15$ представлена в табл. 6.2.

Чтобы оценить показатель ρ связи двух случайных величин, сначала найдем выборочные средние этих величин:

$$\overline{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = \frac{2620}{15} = 173{,}3; \qquad \overline{y} = \frac{1}{15}\sum_{i=1}^{15} y_i = \frac{945}{15} = 63{,}1.$$

Таблица 6.2

Номер наблюдения	Рост, см		Масса тела, кг	
	x_i	$x_i - \overline{x}$	y_i	$y_i - \overline{y}$
1	165	−8,3	72,9	9,8
2	171	−2,3	48,4	−14,7
3	182	8,7	66,3	3,2
4	165	−8,3	64,1	1,0
5	183	9,7	62,7	−0,4
6	180	6,7	76,0	12,9
7	183	9,7	73,8	10,7
8	166	−7,3	50,6	−12,5
9	173	−0,3	52,3	−10,8
10	172	−1,3	56,5	−6,6
11	174	0,7	66,8	3,7
12	170	−3,3	61,6	−1,5
13	164	−9,3	72,8	9,7
14	168	−5,3	52,6	−10,5
15	184	10,7	68,6	5,5
Σ	2600		945	

Затем определяем суммы

$$\sum_{i=1}^{15}(x_i - \overline{x})^2 = 747{,}33; \qquad \sum_{i=1}^{15}(y_i - \overline{y})^2 = 1171{,}4;$$

$$\sum_{i=1}^{15}(x_i - \overline{x})(y_i - \overline{y}) = 293{,}3.$$

Таким образом, $\widehat{\rho} = \dfrac{293{,}3}{\sqrt{747{,}33 \cdot 1171{,}4}} = 0{,}313.$

Замечание 6.2. Если экспериментальные данные представлены в виде (6.10) или (6.11), т.е. сгруппированы по одному или по обоим переменным, то расчетная формула (6.12) для $\widehat{\rho}$ изменяется соответствующим образом. Например, если *выборка* представлена в виде (6.10), то значения оценок $\widehat{\mu}_{11}$, $\widehat{\sigma}_1$ и $\widehat{\sigma}_2$

вычисляют по формулам

$$\widehat{\mu}_{1,1}=\frac{1}{m}\sum_{i=1}^{m}(x_i-\overline{x})(\overline{y}_i-\overline{\overline{y}}),\qquad \overline{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij},\qquad \overline{\overline{y}}=\frac{1}{m}\sum_{i=1}^{m}\overline{y}_i,$$

$$\widehat{\sigma}_1=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i-\overline{x})^2},\qquad \widehat{\sigma}_2=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\overline{y}_i-\overline{\overline{y}})^2}.$$

Интервальная оценка и проверка значимости. При построении *доверительного интервала* для коэффициента корреляции и проверки его значимости будем предполагать, что *генеральная совокупность* имеет двумерный нормальный закон распределения. В этом случае оценка коэффициента корреляции $\widehat{\rho}(\vec{X}_n,\vec{Y}_n)$ имеет асимптотически нормальный закон распределения с математическим ожиданием $\mathrm{M}\,\widehat{\rho}(\vec{X}_n,\vec{Y}_n)\approx\rho-\frac{1}{2n}\bigl(\rho(1-\rho^2)\bigr)$ и дисперсией $\mathrm{D}\,\widehat{\rho}(\vec{X}_n,\vec{Y}_n)\approx\frac{1}{n}(1-\rho^2)^2$.

Заметим, что если распределение генеральной совокупности не является нормальным, то приближенное выражение для $\mathrm{D}\,\widehat{\rho}(\vec{X}_n,\vec{Y}_n)$ содержит вторые и четвертые моменты генеральной совокупности.

Используя общий метод построения доверительного интервала при $\rho^2\ll 1$, основанный на нормальном законе распределения соответствующей оценки при *доверительной вероятности* $\gamma=1-\alpha$ (см. **3.3**), можно получить следующее представление для значений *нижней* и *верхней границ интервальной оценки*:

$$\underline{\rho}\approx\widehat{\rho}+\frac{\widehat{\rho}(1-\widehat{\rho}^2)}{2n}-u_{1-\alpha/2}\frac{1-\widehat{\rho}^2}{\sqrt{n}},\tag{6.13}$$

$$\overline{\rho}\approx\widehat{\rho}+\frac{\widehat{\rho}(1-\widehat{\rho}^2)}{2n}+u_{1-\alpha/2}\frac{1-\widehat{\rho}^2}{\sqrt{n}}.\tag{6.14}$$

Однако пользоваться оценками (6.13) и (6.14) можно только при больших *объемах выборки* (не менее* 500).

*См.: *Кендалл М., Стюарт А.*

При малых объемах выборки можно рекомендовать построение доверительного интервала для ρ, которое основано на преобразовании Р. Фишера*

$$z = \frac{1}{2}\ln\frac{1+\widehat{\rho}}{1-\widehat{\rho}}, \quad \text{или} \quad z = \operatorname{arcth}\widehat{\rho}. \tag{6.15}$$

Оказывается, что случайная величина

$$Z = \frac{1}{2}\ln\frac{1+\widehat{\rho}(\vec{X}_n, \vec{Y}_n)}{1-\widehat{\rho}(\vec{X}_n, \vec{Y}_n)}$$

уже для небольших значений n приблизительно распределена по нормальному закону с параметрами

$$\mathbf{M}Z \approx \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}, \qquad \mathbf{D}Z = \frac{1}{n-3}.$$

Это приводит к представлению

$$\underline{\rho} = \operatorname{th}\underline{z}, \quad \overline{\rho} = \operatorname{th}\overline{z}, \tag{6.16}$$

где

$$\underline{z} = \frac{1}{2}\ln\frac{1+\widehat{\rho}}{1-\widehat{\rho}} - \frac{\widehat{\rho}}{2(n-1)} - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}; \tag{6.17}$$

$$\overline{z} = \frac{1}{2}\ln\frac{1+\widehat{\rho}}{1-\widehat{\rho}} - \frac{\widehat{\rho}}{2(n-1)} + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}. \tag{6.18}$$

Заметим, что равенствами (6.17), (6.18) можно пользоваться и в тех случаях, когда закон распределения генеральной совокупности отличен от нормального. Но в этих случаях ухудшается качество оценивания, т.е. увеличивается длина интервала $(\underline{z}, \overline{z})$, а значит, ухудшается точность оценивания.

*См.: *Крамер Г.*

При проверке статистической гипотезы H_0: $\rho = 0$ (т.е. гипотезы о том, что нормально распределенные случайные величины независимы) используют *статистику*

$$t = \frac{\widehat{\rho}(\vec{X}_n, \vec{Y}_n)\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2(\vec{X}_n, \vec{Y}_n)}}, \tag{6.19}$$

которая имеет *распределение Стьюдента* с $n-2$ степенями свободы*. Если окажется, что

$$\frac{|\widehat{\rho}|\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}} < t_{1-\alpha/2}(n-2),$$

то гипотезу H_0 принимают при *уровне значимости* α.

Пример 6.6. В примере 6.5 найдено значение точечной оценки $\widehat{\rho} = 0{,}519$. Определим значения $\underline{\rho}$ и $\overline{\rho}$ при $\gamma = 0{,}9$ и проверим гипотезу H_0: $\rho = 0$ на уровне значимости $\alpha = 0{,}1$.

Определив по таблице квантилей нормального распределения (см. табл. П.2) значение $u_{1-\alpha/2} = u_{0{,}95} = 1{,}65$ и воспользовавшись формулой (6.13), получим

$$\underline{\rho} \approx 0{,}313 + 0{,}009 - 1{,}65 \cdot \frac{0{,}902}{\sqrt{15}} = 0{,}322 - 0{,}384 \approx -0{,}062,$$

$$\overline{\rho} \approx 0{,}313 + 0{,}009 + 1{,}65 \cdot \frac{0{,}902}{\sqrt{15}} = 0{,}322 + 0{,}384 \approx 0{,}706.$$

Равенства (6.16) дают следующий результат:

$$\underline{\rho} = \operatorname{th}\underline{z} \approx -0{,}162, \qquad \overline{\rho} = \operatorname{th}\overline{z} \approx 0{,}658,$$

который является более надежным.

Для того чтобы проверить гипотезу H_0: $\rho = 0$, по таблице квантилей распределения Стьюдента (см. табл. П.4) находим

*См.: *Кендалл М., Стюарт А.*

квантиль $t_{0,95}(13) = 1{,}77$ и сравниваем со значением

$$\widehat{\rho}\frac{\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}} = 0{,}313\frac{\sqrt{13}}{\sqrt{0{,}902}} = 1{,}19.$$

Поскольку $1{,}19 < 1{,}77$, то гипотезу $\rho = 0$ принимаем.

6.4. Анализ корреляционного отношения

Точечная оценка показателя $r_{\eta\xi}$. Пусть *экспериментальные данные* представлены в форме (6.10), т.е. сгруппированы по значениям x_i случайной величины ξ.

Тогда за значение *точечной оценки* величины σ_f^2 принимают

$$\widehat{\sigma}_f^2 = \frac{1}{n}\sum_{i=1}^{m} n_i(\overline{y}_i - \overline{y})^2.$$

Значение точечной оценки дисперсии σ_η^2 находим по известной формуле (см. **2.1**)

$$\widehat{\sigma}_\eta^2 = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2.$$

Отсюда на основании (6.5) получаем значение точечной оценки показателя $r_{\eta\xi}$:

$$\widehat{r}_{\eta\xi} = \sqrt{\frac{\widehat{\sigma}_f^2}{\widehat{\sigma}_\eta^2}}. \tag{6.20}$$

Напомним, что точечная оценка $\widehat{r}(\vec{X}_n, \vec{Y}_n)$ определяет степень зависимости случайной величины η от случайной величины ξ. Аналогично можно ввести точечную оценку $\widehat{r}_{\xi\eta}$ для *корреляционного отношения* $r_{\xi\eta}$.

Пусть экспериментальные данные получены в форме (6.9) и не допускают удовлетворительной группировки по оси значений ξ (так как недостаточно велико n или точки (x_i, y_i) слишком „разрежены“ на плоскости).

В этом случае нужно выдвинуть некоторое предположение (*статистическую гипотезу*) о виде функции регрессии $\mathrm{M}(\eta|\xi=x)=f(x)$. Проверка таких гипотез будет рассмотрена ниже (см. 7).

Допустим, что параметрический вид этой функции задан, т.е. принято предположение о том, что

$$f(x)=f(x;\theta_1,\ldots,\theta_k)$$

и найдены значения $\widehat{\theta}_i$ оценок параметров θ_i, $i=\overline{1,k}$ (см. 7). Тогда значение точечной оценки $\widehat{\sigma}_\eta^2$ для дисперсии σ_η^2 находим по формуле

$$\widehat{\sigma}_\eta^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2,$$

а значение $\widehat{\overline{\sigma}}_\eta^2$ оценки $\overline{\sigma}_\eta^2$ можно записать в виде

$$\widehat{\overline{\sigma}}_\eta^2=\frac{1}{n-k}\sum_{i=1}^{n}\Big(y_i-f\big(x_i;\widehat{\theta}_1,\ldots,\widehat{\theta}_k\big)\Big)^2. \tag{6.21}$$

Следовательно, согласно (6.5), точечную оценку показателя $r_{\eta\xi}$ можно определить равенством

$$\widehat{r}_{\eta\xi}=\sqrt{\frac{1-\widehat{\overline{\sigma}}_\eta^2}{\widehat{\sigma}_\eta^2}}. \tag{6.22}$$

Интервальная оценка и проверка значимости $r_{\eta\xi}$. Построение *доверительного интервала* для показателя $r_{\eta\xi}$ основано на том, что *статистика**

$$W=\frac{(n-m)\widehat{r}^2_{\eta\xi}(\vec{X}_m,\vec{Y}_n)}{(m-1)(1-\widehat{r}^2_{\eta\xi}(\vec{X}_m,\vec{Y}_n))}\,\frac{m-1}{m-1+nr^2_{\eta\xi}(\vec{X}_m,\vec{Y}_n)}$$

*См.: *Кендалл М., Стюарт А.*

приближенно имеет *распределение Фишера* с числом степеней свободы r_1^* и $r_2 = n - m$, где

$$r_1^* = \frac{(m - 1 + n\widehat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n))^2}{m - 1 + 2n\widehat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n)}, \tag{6.23}$$

в предположении, что при условии $\xi = x$ случайная величина η имеет нормальный закон распределения с постоянной дисперсией для любого x.

Используя квантили $F_{\alpha/2}(r_1^*, r_2)$ и $F_{1-\alpha/2}(r_1^*, r_2)$ распределения Фишера для $\alpha = 1 - \gamma$, где γ — заданная *доверительная вероятность*, можно записать границы доверительного интервала в следующем виде:

$$\underline{r}_{\eta\xi} = \sqrt{\frac{(n-m)\widehat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n)}{n(1 - \widehat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n))F_{\alpha/2}(r_1^*, r_2)} - \frac{m-1}{n}}, \tag{6.24}$$

$$\overline{r}_{\eta\xi} = \sqrt{\frac{(n-m)\widehat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n)}{n(1 - \widehat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n))F_{1-\alpha/2}(r_1^*, r_2)} - \frac{m-1}{n}}. \tag{6.25}$$

Проверка значимости показателя $r_{\eta\xi}$ (т.е. проверка статистической гипотезы $H_0 : r_{\eta\xi} = 0$) основана на том*, что *статистика*

$$W_0 = \frac{(n-m)\widehat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n)}{(m-1)(1 - \widehat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n))} \tag{6.26}$$

имеет распределение Фишера с числом степеней $r_1 = m - 1$ и $r_2 = n - m$, если гипотеза H_0: $r_{\eta\xi} = 0$ верна.

Границу *критического множества* для гипотезы H_0: $r_{\eta\xi} = 0$ на *уровне значимости* α определяет квантиль $f_{1-\alpha}(r_1, r_2)$. Величину показателя $r_{\eta\xi}$ следует считать значимо отличающейся

*См.: *Кендалл М., Стюарт А.*

от нуля, если значение статистики W_0 принадлежит критическому множеству, т.е. ее значение больше $f_{1-\alpha}(r_1, r_2)$. В противном случае делаем вывод об отсутствии *стохастической связи* между η и ξ.

Пример 6.7. Пусть в результате обработки $n = 132$ экспериментальных точек (x_i, y_i), $i = \overline{1, n}$, получено *выборочное значение* корреляционного отношения $\widehat{r}_{\eta\xi} = 0{,}60$, причем промежуток, содержащий все *выборочные значения* случайной величины ξ, был разбит на $m = 12$ равных интервалов (см. 1.3). Найдем значения границ доверительного интервала $(\underline{r}, \overline{r})$ для показателя $r_{\eta\xi}$ с уровнем доверия $\gamma = 0{,}9$ и проверим значимость этого показателя на уровне значимости $\alpha = 0{,}1$.

Сначала определим по формуле (6.23) число степеней свободы r_1^* (округляя до целого числа):

$$r_1^* = \frac{(12 - 1 + 132 \cdot 0{,}36)^2}{12 - 1 + 2 \cdot 132 \cdot 0{,}36} \approx 27.$$

По таблице квантилей распределения Фишера с числом степеней свободы $r_1^* = 27$ и $r_2 = n - m = 132 - 12 = 120$ (см. табл. П.4) находим квантили уровней $\alpha/2 = (1 - \gamma)/2 = 0{,}05$ и $1 - \alpha/2 = 0{,}95$:

$$f_{0{,}05}(27, 120) = 1{,}58;$$

$$f_{0{,}95}(27, 120) = \frac{1}{f_{0{,}05}(120, 27)} = \frac{1}{1{,}73} \approx 0{,}58.$$

По формулам (6.24), (6.25) находим значения границ доверительного интервала:

$$\underline{r} = \sqrt{\frac{120 \cdot 0{,}36}{132 \cdot 0{,}64 \cdot 1{,}58} - \frac{11}{132}} = 0{,}49,$$

$$\overline{r} = \sqrt{\frac{120 \cdot 0{,}36}{132 \cdot 0{,}64 \cdot 0{,}58} - \frac{11}{132}} = 0{,}93.$$

Таким образом, с вероятностью $\gamma = 0{,}9$ истинное значение показателя $r_{\eta\xi}$ (при точечной оценке $\widehat{r}_{\eta\xi} = 0{,}60$) заключено в пределах $0{,}49 < r_{\eta\xi} < 0{,}93$.

Для проверки значимости $r_{\eta\xi}$ (хотя она и так очевидна) найдем квантиль распределения Фишера $f_{1-\alpha}(r_1, r_2)$ при $\alpha = 0{,}1$, $r_1 = 11$, $r_2 = 120$. Поскольку $f_{0{,}9}(120, 11) = 1{,}58$, то $f_{0{,}1}(11, 120) = 1/f_{0{,}9}(120, 11) = 0{,}63$. Значение статистики W_0 равно $6{,}1 > f_{0{,}1} = 0{,}63$, следовательно, гипотеза H_0: $r_{\eta\xi} = 0$ уверенно отклоняется, т.е. между переменными ξ и η имеет место стохастическая связь.

6.5. Анализ множественных связей

Перейдем к рассмотрению *стохастических связей* между совокупностью $p+1$ случайных величин $X_0, X_1, \ldots, X_p$, где *переменные* $X_1, \ldots, X_p$ являются *входными*, а *переменное* $X_0 = Y$ — *выходным*. Такое выделение переменного X_0 не является обязательным, т.е. все переменные могут быть входными, или выходных переменных может быть несколько, но выделенный случай является, по-видимому, наиболее типичным.

Предположим, что случайный вектор $(X_0, X_1, \ldots, X_p)$ имеет нормальный закон распределения, определяемый вектором математических ожиданий $\vec{\mu} = (\mu_0, \mu_1, \ldots, \mu_p)$ и ковариационной матрицей $\Sigma = (\sigma_{ij})$. Таким образом, известна корреляционная матрица

$$P = \begin{pmatrix} 1 & \rho_{01} & \rho_{02} & \ldots & \rho_{0p} \\ \rho_{10} & 1 & \rho_{12} & \ldots & \rho_{1p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho_{p0} & \rho_{p1} & \rho_{p2} & \ldots & 1 \end{pmatrix}, \qquad (6.27)$$

где ρ_{ij} является коэффициентом корреляции между случайными величинами X_i и X_j, $i, j = \overline{0, p}$.

Частные коэффициенты корреляции. При рассмотрении трех и более случайных величин $X_0, X_1, \ldots, X_p$ коэффици-

енты корреляции любой пары из этих случайных величин могут не дать правильного представления о степени связи между всеми случайными величинами. Это объясняется тем, что на закон распределения вероятностей исследуемой пары случайных величин могут оказывать влияние и другие рассматриваемые случайные величины (см. примеры 6.1–6.3).

Это обстоятельство делает необходимым введение показателей стохастической связи между парой случайных величин X_i и X_j $(i=\overline{0,p},\ j=\overline{0,p},\ i\neq j)$ при условии, что значения других случайных величин зафиксированы. В этом случае говорят о статистическом анализе ***частных связей.***

Частный коэффициент корреляции — мера линейной стохастической зависимости между двумя случайными величинами из некоторой совокупности случайных величин X_0, $X_1, \ldots, X_p$, когда исключено влияние остальных, т.е. (для пары X_i и X_j)

$$\rho_{ij(J(i,j))}=\frac{\mathrm{M}\big((Z_i-\mathrm{M}Z_i)(Z_j-\mathrm{M}Z_j)\big)}{\sqrt{\mathrm{D}Z_i\,\mathrm{D}Z_j}}, \qquad (6.28)$$

где

$$Z_i=\alpha_0^i+\sum_{k\in J(i,j)}\alpha_k^i X_k, \qquad Z_j=\beta_0^j+\sum_{k\in J(i,j)}\beta_k^j X_k$$

и

$$J(i,j)=\{0,1,\ldots,p\}\setminus\{i,j\}.$$

При этом i, j называют ***первичными индексами***, а остальные — ***вторичными***. Коэффициенты α_0^i, β_0^j, α_k^i, β_k^j, $k\in J(i,j)$, находят из условия минимизации следующих функций:

$$\mathrm{M}\Big(X_i-\alpha_0^i-\sum_{k\in J(i,j)}\alpha_k^i X_k\Big)^2, \qquad \mathrm{M}\Big(X_j-\beta_0^j-\sum_{k\in J(i,j)}\beta_k^j X_k\Big)^2.$$

Если случайный вектор $(X_0, \ldots, X_p)$ распределен по нормальному закону, то частный коэффициент корреляции между

случайными величинами X_i и X_j вычисляют по формуле*

$$\rho_{ij(J(i,j))} = -\frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}, \tag{6.29}$$

где Σ_{ij} — алгебраическое дополнение для элемента ρ_{ij} корреляционной матрицы (6.27).

Например, при $i=0$, $j=1$ по формуле (6.29) имеем

$$\rho_{01(2)} = \frac{\rho_{01} - \rho_{02}\rho_{12}}{\sqrt{(1-\rho_{02}^2)(1-\rho_{12}^2)}}.$$

Из формулы (6.29) следует, что для вычисления частных коэффициентов корреляции нужны лишь все коэффициенты корреляции случайных величин X_i, X_j, $i \neq j$.

Численные расчеты могут быть упрощены, если использовать рекуррентные соотношения**

$$\rho_{01(2,3,\ldots,k+1)} = \frac{\rho_{01(2,\ldots,k)} - \rho_{0(k+1)(2,\ldots,k)}\rho_{1(k+1)(2,\ldots,k)}}{\sqrt{\left(1-\rho^2_{0(k+1)(2,\ldots,k)}\right)\left(1-\rho^2_{1(k+1)(2,\ldots,k)}\right)}}. \tag{6.30}$$

Согласно (6.30), любой частный коэффициент корреляции может быть выражен через частные коэффициенты с меньшим на единицу числом вторичных индексов.

Замечание 6.3. Практика многомерного статистического анализа показала, что частные коэффициенты корреляции, определенные соотношениями (6.28)–(6.30), — вполне приемлемые характеристики линейной связи и в том случае, когда распределение анализируемых переменных X_0, X_1, ..., X_p отличается от нормального.

Статистический анализ частных коэффициентов корреляции. Вычисление значений $\widehat{\rho}_{ij(J(i,j))}$ точечной оценки частного коэффициента корреляции $\rho_{ij(J(i,j))}$ проводят по тем же

*См.: *Крамер Г.*

**См. там же.

формулам (6.29), (6.30) путем подстановки вместо коэффициентов корреляции ρ_{ij} их выборочных значений $\widehat{\rho}_{ij}$.

Исследуя статистические свойства выборочного частного коэффициента корреляции $\widehat{\rho}_{ij(J(i,j))}(\vec{X}_{0n},\ldots,\vec{X}_{pn})$, можно воспользоваться тем, что он распределен* точно так же, как и выборочный коэффициент корреляции тех же случайных величин X_i, X_j, но с единственной поправкой: *объем выборки* n надо заменить на $n-k$, где k — ***порядок частного коэффициента корреляции*** (см. (6.30)). Поэтому все формулы для *доверительных интервалов* и *критерии* значимости, приведенные в предыдущем пункте, сохраняются и для частных коэффициентов корреляции с учетом замены n на $n-k$.

Пример 6.8. По итогам работы 37 однотипных прядильных фабрик в течение года были измерены следующие показатели: $X_0 = Y$ — среднемесячная характеристика качества пряжи (в баллах), X_1 — среднемесячное количество профилактических наладок автоматической линии, X_2 — среднемесячное число обрывов нити.

По матрице исходных данных X_{0i}, X_{1i}, X_{2i}, $i=\overline{1,37}$, были подсчитаны выборочные коэффициенты корреляции $\widehat{\rho}_{ij}$ по формуле (6.12):

$$\widehat{\rho}_{01}=0{,}105; \qquad \widehat{\rho}_{02}=0{,}024; \qquad \widehat{\rho}_{12}=0{,}966.$$

Значения $\widehat{\rho}_{01}$ и $\widehat{\rho}_{02}$ дали основание предполагать, что случайные величины X_0, X_i, $i=1,2$, некоррелированные.

Гипотезы о равенстве нулю ρ_{01} и ρ_{02} были приняты на уровне значимости $\alpha=0{,}1$. Это свидетельствует об отсутствии стохастической связи между X_0 (качество ткани) и X_1, X_2, но не согласуется с профессиональными представлениями технологов.

Однако расчет значений частных коэффициентов корреляции по формуле (6.29) дает $\widehat{\rho}_{01(2)}=0{,}907$ и $\widehat{\rho}_{02(1)}=-0{,}906$, что

*См.: *Андерсон Т.*

вполне соответствует представлениям специалистов о характере связей между рассмотренными показателями.

Построение доверительных интервалов для истинных значений $\rho_{01(2)}$ и $\rho_{02(1)}$, согласно формулам (6.16), с учетом того, что объем выборки $n = 37$ должен быть уменьшен на 1 (ибо число „мешающих" переменных в данном случае равно $k = 1$), дает следующие результаты (на уровне доверия $\gamma = 0{,}9$):

$$0{,}821 < r_{01(2)} < 0{,}950; \qquad -0{,}950 < r_{02(1)} < -0{,}819.$$

Пример 6.9. С целью исследования влияния погодных условий (X_1 — весеннее количество осадков, см; X_2 — накопленная за весну сумма температур, ° С) на урожайность (в ц/га) кормовых трав X_0 в районе с одинаковыми метеорологическими условиями были получены выборочные значения вектора (X_0, X_1, X_2) на $n = 20$ участках. По этим *экспериментальным данным* (X_{0i}, X_{1i}, X_{2i}), $i = \overline{1, 20}$, были вычислены значения коэффициентов корреляции $\widehat{\rho}_{01} = 0{,}80$; $\widehat{\rho}_{02} = -0{,}40$; $\widehat{\rho}_{12} = -0{,}56$.

Значение $\widehat{\rho}_{02} = -0{,}40$ вызывает вопрос: действительно ли высокая температура X_2 отрицательно влияет на урожайность, или здесь сказывается влияние „мешающего" фактора — количества осадков?

Вычисление значений частных коэффициентов корреляции по формуле (6.29) дает следующие значения:

$$\widehat{\rho}_{01(2))} = 0{,}759; \qquad \widehat{\rho}_{02(1)} = 0{,}097; \qquad \widehat{\rho}_{12(0)} = -0{,}436.$$

Как видим, если исключить одновременное влияние количества осадков X_1 на X_0 (с ростом X_1 урожайность повышается) и на X_2 (с ростом X_1 температура X_2 понижается), то мы уже не обнаружим отрицательной корреляции между температурой X_2 и урожайностью X_0, ибо $\widehat{\rho}_{02(1)} = 0{,}097$, что не является значимой степенью стохастической связи.

Множественный коэффициент корреляции. Для того чтобы результаты, изложенные в **6.4**, были частным случаем рассматриваемой общей ситуации, сохраним обозначение η для

„выходной" переменной X_0 и обозначение ξ для „входной" переменной, но теперь ξ будет вектором размерности p, т.е. $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_p)$. Возможные значения переменной η будем обозначать y, а возможные значения $\vec{\xi}$ — $\vec{x} = (x_1, \ldots, x_p)$.

При решении практических задач, связанных с анализом стохастических связей между многими случайными переменными, чаще других рассматривают ситуацию, в которой поведение какой-то одной (выходной) переменной η стараются объяснить поведением совокупности других (входных) переменных $\vec{\xi} = (\xi_1, \ldots, \xi_p)$.

Прежде всего убедимся, что наилучшим прогнозом (аппроксимацией) для неизвестного значения η (в смысле средней квадратичной ошибки) является условное математическое ожидание η при условии $\vec{\xi} = \vec{x}$, т.е. величина $\mathrm{M}(\eta | \vec{\xi} = \vec{x}) = f(\vec{x})$, где $\vec{x} = (x_1, \ldots, x_n)$.

Действительно, пусть $\tilde{f}(\vec{x})$ — любая функция. Тогда

$$\mathrm{M}(\eta - \tilde{f}(\vec{\xi}))^2 = \mathrm{M}((\eta - f(\vec{\xi})) + (f(\vec{\xi}) - \tilde{f}(\vec{\xi})))^2 =$$

$$= \mathrm{M}(\eta - f(\vec{\xi}))^2 + \mathrm{M}(f(\vec{\xi}) - \tilde{f}(\vec{\xi}))^2 + 2\mathrm{M}((f(\vec{\xi}) - \tilde{f}(\vec{\xi}))(\eta - f(\vec{\xi}))).$$

Поскольку последнее слагаемое равно нулю (доказательство этого аналогично тому, которое приведено в **6.2**), то

$$\min \mathrm{M}(\eta - \tilde{f}(\vec{\xi}))^2 = \mathrm{M}(\eta - f(\vec{\xi}))^2,$$

если $\tilde{f}(\vec{\xi}) = f(\vec{\xi})$. Следовательно, при каждом данном значении $\vec{\xi} = \vec{x}$ и любой функции $\tilde{f}(\vec{x}) \neq f(\vec{x})$ имеет место неравенство

$$\mathrm{M}(\eta - \tilde{f}(\vec{x}))^2 > \mathrm{M}(\eta - f(\vec{x}))^2.$$

Таким образом, мы снова (как и в **6.1**) пришли к *функции регрессии* $f(\vec{x}) = \mathrm{M}(\eta | \vec{\xi} = \vec{x})$, но уже функции от p переменных $x_1, \ldots, x_p$, которая наиболее точно (в смысле сред-

ней квадратичной ошибки) воспроизводит значения исследуемого результирующего переменного η по заданным величинам $\vec{x} = (x_1, \ldots, x_p)$ входных переменных $\vec{\xi} = (\xi_1, \ldots, \xi_p)$.

Теперь вернемся к соотношению (6.4), которое связывает дисперсию σ_η^2 случайной величины η с величинами $\sigma_f^2 = \mathbf{D} f(\xi)$ и $\overline{\sigma}_\eta^2 = \mathbf{MD}(\eta|\xi)$. Соотношение (6.4) остается справедливым и в случае вектора входных переменных $\vec{\xi} = (\xi_1, \ldots, \xi_p)$.

Следовательно, так же как и в случае парной зависимости, случайный разброс (вариация) выходного переменного η складывается из контролируемой нами (посредством $\vec{x} = (x_1, \ldots, x_p)$) вариации функции регрессии $f(\vec{x})$ и из неподдающегося нашему контролю случайного разброса значений η (при фиксированном $\vec{x}$) относительно функции регрессии. Именно этот неконтролируемый разброс определяет меру зависимости переменной η от переменной $\vec{\xi}$, которая характеризуется величиной $\overline{\sigma}_\eta^2$. Чем меньше значение $\overline{\sigma}_\eta^2$, тем точнее прогноз. При $\overline{\sigma}_\eta^2 = 0$ случайные величины η и $\vec{\xi}$ связаны функциональной зависимостью.

Эти соображения подводят нас к определению ***множественного коэффициента корреляции*** R_η, под которым понимают величину

$$R_\eta = \sqrt{1 - \frac{\overline{\sigma}_\eta^2}{\sigma_\eta^2}}. \tag{6.31}$$

Заметим, что квадрат R_η^2 показателя R_η принято называть ***коэффициентом детерминации***.

Покажем, что R_η есть коэффициент корреляции между η и $f(\vec{\xi})$ (тем самым оправдаем его название). Имеем

$$\begin{aligned}
\mathbf{cov}(\eta, f(\vec{\xi})) &= \mathbf{M}\big((\eta - \mathbf{M}\eta)(f(\vec{\xi}) - \mathbf{M}\eta)\big) = \\
&= \mathbf{M}\big((f(\vec{\xi}) - \mathbf{M}\eta)^2 + (\eta - f(\vec{\xi}))(f(\vec{\xi}) - \mathbf{M}\eta)\big) = \\
&= \mathbf{M}\big(f(\vec{\xi}) - \mathbf{M}\eta\big)^2 + \mathbf{M}\big((\eta - f(\vec{\xi}))(f(\vec{\xi}) - \mathbf{M}\eta)\big) = \sigma_f^2,
\end{aligned}$$

поскольку

$$\mathrm{M}\big((\eta - f(\vec{\xi}))(f(\vec{\xi}) - \mathrm{M}\eta)\big) = 0.$$

Далее,

$$R_\eta = \sqrt{1 - \frac{\overline{\sigma}_\eta^2}{\sigma_\eta^2}} = \sqrt{\frac{\sigma_\eta^2 - \overline{\sigma}_\eta^2}{\sigma_\eta^2}} = \sqrt{\frac{\sigma_f^2}{\sigma_\eta^2}} = \frac{\sigma_f^2}{\sqrt{\sigma_\eta^2 \sigma_f^2}} = \frac{\mathbf{cov}(\eta, f(\vec{\xi}))}{\sigma_\eta \sigma_f}.$$

Отметим свойства показателя R_η, которые непосредственно вытекают из соотношения (6.31), справедливого и в многомерном случае.

1°. $0 \leqslant R_\eta \leqslant 1$.

2°. $R_\eta = 0$ соответствует $\sigma_f^2 = \mathrm{D} f(\vec{\xi}) = 0$. В частности, функция регрессии f не зависит от значений ее аргументов $\vec{x}$: $f(\vec{x}) = \mathrm{const}$.

3°. $R_\eta = 1$ соответствует $\overline{\sigma}_\eta^2 = 0$ и означает наличие чисто функциональной связи между η и $\vec{\xi} = (\xi_1, \ldots, \xi_p)$: $\eta = f(\xi_1, \ldots, \xi_p)$.

Определение показателя R_η в виде (6.31) и отмеченные свойства 1°–3° справедливы при любом законе распределения вектора $(\eta, \xi_1, \ldots, \xi_p)$.

Если же предположить, что исходные *статистические данные* $(x_{1i}, x_{2i}, \ldots, x_{pi})$, y_i, $i = \overline{1, n}$, могут интерпретироваться как выборка объема n из $(p+1)$-мерной *генеральной совокупности*, распределенной по нормальному закону с вектором средних значений $\vec{\mu} = (\mu_0, \mu_1, \ldots, \mu_p)$, где $\mu_0 = \mathrm{M}\eta$, $\mu_i = \mathrm{M}\xi_i$, $i = \overline{1, p}$, и ковариационной матрицей Σ, то можно отметить дополнительные свойства показателя R_η и правила его вычисления.

Прежде всего укажем на то, что в рассматриваемой ситуации (ср. с примером 6.4) условное математическое ожидание η при фиксированных значениях $\xi_1 = x_1, \ldots, \xi_p = x_p$ (т.е. функция регрессии $f(x)$) является линейной функцией переменных

$x_1, \ldots, x_p$, а условная дисперсия $\mathbf{D}(\eta|\vec{\xi}=\vec{x})$ не зависит от $\vec{x}=(x_1,\ldots,x_p)$ и имеет вид

$$\mathbf{D}(\eta)\vec{\xi}=\vec{x}=\sigma_\eta^2(1-R_\eta^2).$$

Последнее выражение — полная аналогия формулы (6.2), только роль коэффициента корреляции ρ играет множественный коэффициент корреляции R_η.

Приведем без доказательства* следующие дополнительные свойства показателя R_η в случае совместного нормального закона распределения переменных η и $\vec{\xi}=(\xi_1, \ldots, \xi_p)$.

4°. С помощью корреляционной матрицы P (6.27) показатель R_η можно вычислить по формуле

$$R_\eta=\sqrt{1-\frac{\det P}{P_{00}}}, \tag{6.32}$$

где $\det P$ — определитель матрицы P, а P_{00} — алгебраическое дополнение элемента $\rho_{00}=1$.

5°. Показатель R_η можно вычислить, используя частные коэффициенты корреляции следующим образом:

$$R_\eta^2=1-(1-\rho_{01}^2)\prod_{j=2}^{p}(1-\rho_{0j(12\ldots j-1)}^2). \tag{6.33}$$

6°. Множественный коэффициент корреляции мажорирует любой парный коэффициент корреляции, характеризующий стохастическую связь результирующего показателя η с остальными, т.е.

$$|\rho_{0j}|\leqslant R_\eta, \quad |\rho_{0j(\cdot)}\leqslant R_\eta, \quad j=\overline{0,p},$$

где $\rho_{0j(\cdot)}$ — произвольный частный коэффициент корреляции, содержащий нуль среди первичных индексов.

*Доказательство см.: *Кендалл М., Стюарт А.*

7°. Присоединение каждого нового предсказывающего (входного) переменного не может уменьшить величины R_η (независимо от порядка присоединения).

Статистический анализ множественного коэффициента корреляции. Вычисление значений точечной оценки $\widehat{R}_\eta$ показателя R_η проводится по тем же формулам (6.31)-(6.33) путем подстановки в них вместо значений *теоретических характеристик* соответствующих значений *выборочных характеристик*.

Например, при использовании формулы (6.32) матрицу P нужно заменить матрицей $\widehat{P}$, в которой все элементы ρ_{ij} заменены на $\widehat{\rho}_{ij}$, $i,j=\overline{0,p}$, а при использовании формулы (6.33) коэффициент корреляции ρ_{01} и все частные коэффициенты корреляции $\rho_{ij(\cdot)}$ нужно заменить значениями $\widehat{\rho}_{ij(\cdot)}$.

Для проверки гипотезы H_0: $R_\eta=0$ будем предполагать, что случайный вектор (ξ,η) имеет $(p+1)$-мерный нормальный закон распределения, и воспользуемся тем*, что *статистика*

$$W_1=\frac{\widehat{R}_\eta^2}{1-\widehat{R}_\eta^2}\,\frac{n-p-1}{p}$$

имеет *распределение Фишера* с p и $n-p-1$ степенями свободы, если истинное значение $R_\eta=0$.

Гипотеза об отсутствии множественной корреляционной связи между η и $\vec{\xi}=(xi_1,\ \ldots,\ \xi_p)$ отвергается на уровне значимости α, если

$$\frac{\widehat{R}^2}{1-\widehat{R}^2}\,\frac{n-p-1}{p}>F_{1-\alpha}(p,n-p-1). \tag{6.34}$$

В предположении, что η при условии $\vec{\xi}=\vec{x}$ имеет нормальный закон с постоянной дисперсией для любого $\vec{x}$, можно показать**, что значения приближенных доверительных границ $\underline{R}_\eta$

*См.: *Кендалл М., Стюарт А.*

**См.: *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.*, 1985.

и $\overline{R}_\eta$ для показателя R_η, отвечающие доверительной вероятности $\gamma = 1 - \alpha$ и выборке объема n, имеют вид (справедливый при условии $p \geqslant 8$):

$$\underline{R}_\eta = \sqrt{\frac{\hat{R}_\eta^2[1-(p+1)/n]}{(1-\hat{R}_\eta^2)F_{1-\alpha/2}(r_1,r_2)} - \frac{p}{n}}, \tag{6.35}$$

$$\overline{R}_\eta = \sqrt{\frac{\hat{R}_\eta^2[1-(p+1)/n]}{(1-\hat{R}_\eta^2)F_{\alpha/2}(r_1,r_2)} - \frac{p}{n}}, \tag{6.36}$$

где

$$r_1 = \frac{(p+n\hat{R}_\eta)^2}{p+2n\hat{R}_\eta^2}, \quad r_2 = n-p-1.$$

Пример 6.10. Вернемся к примерам 6.8 и 6.9.

В примере 6.8 найдем значения оценок множественного коэффициента корреляции R_η между показателем качества η пряжи и совокупностью двух факторов: количеством ξ_1 профилактических наладок и числом ξ_2 обрывов нити.

Используя формулу (6.33), в которой вместо истинных значений показателей корреляции использованы значения их выборочных оценок (см. пример 6.3), получаем

$$\hat{R}^2 = 1-(1-\hat{\rho}_{01}^2)(1-\hat{\rho}_{02(1)}^2) =$$

$$= 1-(1-0{,}105^2)(1-0{,}906^2) = 0{,}823,$$

откуда $\hat{R} = \sqrt{0{,}823} = 0{,}907$.

В примере 6.9 найдем значения оценок показателя R_η множественной корреляции между урожайностью η кормовых трав и природными факторами: весенним количеством ξ_1 осадков и накопленной суммой ξ_2 температур.

Используя найденные в примере 6.4 оценки $\hat{\rho}_{01} = 0{,}8$ и $\hat{\rho}_{02(1)} = 0{,}097$, по той же формуле (6.33) находим (с заменой

истинных значений показателей корреляции значениями их оценок)

$$\widehat{R}^2 = 1 - (1 - \widehat{\rho}_{01}^2)(1 - \widehat{\rho}_{02(1)}^2) = \\ = 1 - (1 - 0{,}80^2)(1 - 0{,}097^2) = 0{,}644,$$

откуда $\widehat{R} = \sqrt{0{,}644} = 0{,}802$.

Заметим, что формулами (6.35), (6.36) для вычисления границ доверительного интервала воспользоваться нельзя, так как не выполнено условие $p \geqslant 8$ (у нас $p = 3$).

6.6. Решение типовых примеров

Пример 6.11. Двумерная случайная величина имеет нормальный закон распределения. Определим *доверительный интервал* для коэффициента корреляции ρ с *коэффициентом доверия* $\gamma = 0{,}99$, если значение $\widehat{\rho}$, найденное по *выборке объема* $n = 300$, равно 0,14.

Воспользуемся тем, что при больших объемах выборки *оценка* $\widehat{\rho}(\vec{X}_n, \vec{Y}_n)$ распределена почти по нормальному закону с параметрами $\rho - \rho(1-\rho^2)/2n$ и $(1-\rho^2)^2/n$ (см. **6.3**). По таблице квантилей нормального распределения (см. табл. П.2) найдем квантиль $u_{(1+\gamma)/2} = u_{0{,}995} = 2{,}575$. Имеем

$$\mathbf{P}\left\{\left|\widehat{\rho}(\vec{X}_n, \vec{Y}_n) - \frac{\rho - \rho(1-\rho^2)}{2n}\right| \frac{\sqrt{n}}{1-\rho^2} < u_{0{,}995}\right\} = 0{,}99.$$

Отсюда

$$\widehat{\rho}(\vec{X}_n, \vec{Y}_n) + \frac{\rho(1-\rho^2)}{2n} - u_{0{,}995}\frac{1-\rho^2}{\sqrt{n}} < \rho < \\ < \widehat{\rho}(\vec{X}_n, \vec{Y}_n) + \frac{\rho(1-\rho^2)}{2n} + u_{0{,}995}\frac{1-\rho^2}{\sqrt{n}}.$$

Заменяя в левой и правой частях неравенств ρ на $\widehat{\rho}$ и подставляя значение $u_{0{,}995} = 2{,}575$, для данной выборки получаем границы

доверительного интервала в виде

$$\underline{\rho} = 0{,}14 + \frac{0{,}14(1-0{,}14^2)}{600} - 2{,}575\frac{1-0{,}14^2}{\sqrt{300}},$$

$$\overline{\rho} = 0{,}14 + \frac{0{,}14(1-0{,}14^2)}{600} + 2{,}575\frac{1-0{,}14^2}{\sqrt{300}}.$$

После вычислений окончательно получаем доверительный интервал $(0{,}13,\ 0{,}16)$.

Пример 6.12. Двумерная случайная величина имеет нормальный закон распределения. Построим доверительный интервал для коэффициента корреляции ρ с коэффициентом доверия $\gamma = 0{,}95$, если значение $\widehat{\rho}(\vec{X}_n, \vec{Y}_n)$, найденное по выборке объема $n = 12$, равно $-0{,}65$.

Поскольку объем выборки мал, используем случайную величину

$$Z = \frac{1}{2}\ln\frac{1+\widehat{\rho}(\vec{X}_n, \vec{Y}_n)}{1-\widehat{\rho}(\vec{X}_n, \vec{Y}_n)},$$

которая имеет приближенно нормальный закон распределения с параметрами

$$\mu = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}, \qquad \sigma = \frac{1}{n-3}.$$

Используя таблицу квантилей нормального распределения (см. табл. П.2), находим квантиль $u_{(1+\gamma)/2} = u_{0{,}975} = 1{,}96$. Имеем

$$\mathbf{P}\left\{\left|\frac{1}{2}\ln\frac{1+\widehat{\rho}(\vec{X}_n, \vec{Y}_n)}{1-\widehat{\rho}(\vec{X}_n, \vec{Y}_n)} - \frac{1}{2}\ln\frac{1+\rho}{1-\rho} - \frac{\rho}{2(n-1)}\right| < \frac{1{,}96}{\sqrt{n-3}}\right\} = 0{,}95,$$

откуда

$$\frac{1}{2}\ln\frac{1+\widehat{\rho}(\vec{X}_n, \vec{Y}_n)}{1-\widehat{\rho}(\vec{X}_n, \vec{Y}_n)} - \frac{1{,}96}{\sqrt{n-3}} < \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} <$$

$$< \frac{1}{2}\ln\frac{1+\widehat{\rho}(\vec{X}_n, \vec{Y}_n)}{1-\widehat{\rho}(\vec{X}_n, \vec{Y}_n)} + \frac{1{,}96}{\sqrt{n-3}}.$$

Учитывая условия задачи, получаем

$$\frac{1}{2}\ln\frac{0{,}35}{1{,}65}-\frac{1{,}96}{3}<\frac{1}{2}\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{22}<\frac{1}{2}\ln\frac{0{,}35}{1{,}65}+\frac{1{,}96}{3},$$

или

$$-\ln\frac{33}{7}-1{,}31<\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{11}<-\ln\frac{33}{7}+1{,}31.$$

Решая уравнения

$$\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{11}=1{,}31-\ln\frac{33}{7},$$
$$\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{11}=-1{,}31-\ln\frac{33}{7},$$

находим нижнюю $\underline{\rho}$ и верхнюю $\overline{\rho}$ границы доверительного интервала: $\overline{\rho}\approx -0{,}12$, $\underline{\rho}\approx -0{,}88$. Таким образом, доверительный интервал для ρ имеет вид $(-0{,}988, -0{,}12)$.

Заметим, что границы доверительного интервала можно определить с помощью (6.16).

Пример 6.13. В условиях примера 6.11 проверим *гипотезу* H_0: $\rho=0$ на *уровне значимости* $\alpha=0{,}01$.

Если гипотеза H_0 верна, *статистика*

$$t=\frac{\widehat{\rho}(\vec{X}_n,\vec{Y}_n)\sqrt{n-2}}{1-\widehat{\rho}^2(\vec{X}_n,\vec{Y}_n)}$$

имеет *распределение Стьюдента* с $n-2$ степенями свободы. Поскольку объем выборки большой, соответствующую квантиль можно найти по таблице квантилей нормального распределения (см. табл. П.2): $u_{1-\alpha/2}=u_{0{,}995}=2{,}575$. По данным задачи вычисляем *выборочное значение* статистики t:

$$t_в=\frac{0{,}14\sqrt{300-2}}{\sqrt{1-0{,}0196}}=\frac{0{,}14\sqrt{298}}{0{,}9804}\approx 2{,}44.$$

Поскольку $2{,}44 < 2{,}575$, то гипотезу H_0: $\rho = 0$ принимаем на уровне значимости $\alpha = 0{,}01$.

Пример 6.14. По выборке объема $n = 28$ из двумерной *генеральной совокупности*, распределенной по нормальному закону, найдено значение оценки $\widehat{\rho} = 0{,}88$ коэффициента корреляции. Проверим гипотезу $H_0 : \rho \geqslant 0{,}90$ при *альтернативной гипотезе* H_1: $\rho < 0{,}90$ на уровне значимости $\alpha = 0{,}01$.

Для проверки гипотезы H_0 воспользуемся статистикой

$$Z(\vec{X}_n, \vec{Y}_n) = \frac{1}{2} \ln \frac{1 + \widehat{\rho}(\vec{X}_n, \vec{Y}_n)}{1 - \widehat{\rho}(\vec{X}_n, \vec{Y}_n)}$$

(см. (6.15)), для которой имеем

$$\mathbf{P}\left\{\left(Z(\vec{X}_n, \vec{Y}_n) - \frac{1}{2} \ln \frac{1+\rho}{1-\rho} - \frac{\rho}{2(n-1)}\right)\sqrt{n-3} < u_{0{,}01}\right\} = 0{,}01.$$

С помощью таблицы квантилей нормального распределения (см. табл. П.3) находим квантиль $u_{0{,}01} = 0{,}504$, а затем — границу *критической области* для $Z(\vec{X}_n, \vec{Y}_n)$:

$$\frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} + \frac{u_{0{,}01}}{\sqrt{n-3}} = \frac{1}{2} \ln \frac{1+0{,}9}{0{,}1} + \frac{0{,}9}{2 \cdot 27} + \frac{0{,}504}{5} =$$

$$= \frac{1}{2} \ln 19 + \frac{0{,}1}{6} + 0{,}1008 \approx \frac{1}{2} \ln 19 + 0{,}118.$$

Вычислим выборочное значение

$$Z_{\text{в}} = \frac{1}{2} \ln \frac{1 + 0{,}88}{1 - 0{,}88} \approx \frac{1}{2} \ln 15{,}7.$$

Поскольку выборочное значение попало в критическую область ($Z_{\text{в}} < \frac{1}{2} \ln 19 + 0{,}118$), гипотезу H_0 отклоняем.

Пример 6.15. По выборке объема $n = 19$, заданной в виде таблицы (табл. 6.3), найдем значение оценки *корреляционного*

Таблица 6.3

x	0	0	0	1	2	2	3	3	4	4
y	22,8	21,9	22,1	24,5	26,0	26,1	26,8	27,3	28,2	28,5
x	5	6	6	6	7	8	8	9	1	
y	28,9	30,0	30,3	29,8	30,4	31,4	31,5	31,8	33,7	

отношения и границы соответствующего доверительного интервала с коэффициентом доверия $\gamma = 0{,}8$.

Вычисляем выборочное среднее

$$\overline{y} = \frac{1}{19}\Big(22{,}8 + 21{,}9 + 22{,}1 + 24{,}5 + \\ + 26{,}0 + 26{,}1 + 26{,}8 + 27{,}3 + 28{,}2 + 28{,}5 + 28{,}9 + 30{,}0 + \\ + 30{,}3 + 29{,}8 + 30{,}4 + 31{,}4 + 31{,}5 + 31{,}8 + 33{,}1\Big) \approx 28{,}0$$

и выборочную дисперсию

$$\widehat{\sigma}^2 = \frac{1}{19}\Big(22{,}8^2 + 21{,}9^2 + 22{,}1^2 + 24{,}5^2 + 26{,}0^2 + 26{,}1^2 + 26{,}8^2 + \\ + 27{,}3^2 + 28{,}2^2 + 28{,}5^2 + 28{,}9^2 + 30{,}0^2 + 30{,}3^2 + 29{,}8^2 + 30{,}4^2 + \\ + 31{,}4^2 + 31{,}5^2 + 31{,}8^2 + 33{,}1^2\Big) - 28^2 \approx 292{,}99 - 782{,}32 = 10{,}67.$$

Чтобы вычислить значение $\widehat{\sigma}_f^2$, составим *статистический ряд* (табл. 6.4). С помощью этого ряда находим

$$\widehat{\sigma}_f^2 = \frac{1}{19}\Big(3(22{,}3 - 28{,}0)^2 + (24{,}5 - 28{,}0)^2 + 2(26{,}0 - 28{,}0)^2 + \\ + 2(27{,}0 - 28{,}0)^2 + 2(28{,}4 - 28{,}0)^2 + (28{,}9 - 28{,}0)^2 + \\ + 3(30{,}0 - 28{,}0)^2 + (30{,}4 - 28{,}0)^2 + 2(31{,}5 - 28{,}0)^2 + \\ + (31{,}8 - 28{,}0)^2 + (33{,}1 - 28{,}0)^2\Big) \approx 8{,}18.$$

Согласно формуле (6.20),

$$\widehat{r}_{\xi\eta} = \sqrt{\frac{8{,}18}{10{,}67}} \approx \sqrt{0{,}77} \approx 0{,}9.$$

Таблица 6.4

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	3	1	2	2	2	1	3	1	2	1	1
y_i	22,3	24,5	26,0	27,0	28,4	28,9	30,0	30,4	31,5	31,8	33,1

Чтобы определить границы доверительного интервала, предварительно найдем степени свободы

$$r_1^* = \frac{(11-1+19\cdot 0{,}77)^2}{11-1+2\cdot 19\cdot 0{,}77} + \frac{(10+19\cdot 0{,}77)^2}{10+38\cdot 0{,}77} \approx 15$$

и $r_2 = n - m = 19 - 11 = 8$, а также квантили распределения Фишера

$$F_{1-0,1} = F_{0,9} = 2{,}46, \qquad F_{0,1} = \frac{1}{F_{0,9}} \approx 0{,}41.$$

Далее с помощью формул (см. (6.24), (6.25)) получаем

$$\underline{r}_{\eta\xi} = \sqrt{\frac{(19-11)\cdot 0{,}77}{19(1-0{,}77)\cdot 2{,}46} - \frac{10}{19}} = \sqrt{\frac{10{,}90-10}{19}} \approx \sqrt{0{,}05} \approx 0{,}2,$$

$$\overline{r}_{\eta\xi} = \sqrt{\frac{(19-11)\cdot 0{,}77}{19(1-0{,}77)\cdot 0{,}41} - \frac{10}{19}} = \sqrt{\frac{10{,}90-10}{19}} \approx \sqrt{2{,}91} \approx 1{,}7.$$

Итак, $0{,}2 < r_{\eta\xi} < 1$.

Пример 6.16. По выборке объема $n = 24$ (табл. 6.5) найдем значение оценки корреляционного отношения и проверим гипотезу H_0: $\sigma_{\eta\xi} = 0$ на уровне значимости $\alpha = 0{,}05$.

Вычислим выборочное среднее $\overline{y}$ и выборочную дисперсию $\widehat{\sigma}_\eta^2$:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx 14{,}08, \qquad \widehat{\sigma}_\eta^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2 \approx 28{,}92.$$

Таблица 6.5

x	10	10	10	10	10	20	20	20	20	35	35	35
y	5	6	5	6	7	12	13	14	13	17	19	16
x	35	35	40	40	40	40	60	60	60	60	60	
y	15	18	20	21	18	20	17	19	16	14	16	

Таблица 6.6

x_i	10	20	35	40	60
n_i	5	4	5	5	5
y_i	5,8	13	16,2	19,4	16,4

Чтобы найти значение $\widehat{\sigma}_f^2$, по результатам выборки составим статистический ряд (табл. 6.6). Далее по формуле

$$\widehat{\sigma}_f^2 = \frac{1}{n}\sum_{i=1}^{n} n_i (y_i - \overline{y})^2$$

получим $\widehat{\sigma}_f^2 \approx 4{,}66$. Отсюда

$$\widehat{r}_{\eta\xi} = \sqrt{\frac{\widehat{\sigma}_f^2}{\widehat{\sigma}_\eta^2}} = \sqrt{\frac{4{,}66}{28{,}92}} \approx \sqrt{0{,}16} = 0{,}4.$$

Для проверки гипотезы H_0: $r_{\eta\xi} = 0$ используем статистику (6.26)

$$F = \frac{(n-m)\widehat{r}^2_{\eta\xi}(\vec{X}_n, \vec{Y}_n)}{(m-1)(1-\widehat{r}^2_{\eta\xi}(\vec{X}_n, \vec{Y}_n))},$$

которая приближенно имеет *распределение Фишера* со степенями свободы $r_1 = m - 1 = 5 - 1 = 4$ и $r_2 = n - m = 24 - 5 = 19$. По таблице квантилей распределения Фишера (см. табл. П.4) находим $F_{1-\alpha} = F_{0{,}95} = 2{,}92$. Поскольку значение статистики

$$F = \frac{(24-5)\cdot 0{,}16}{4 \cdot 0{,}94} = \frac{19 \cdot 16}{4 \cdot 94} \approx 0{,}8$$

меньше квантили $F_{0{,}95} = 2{,}92$, гипотезу H_0 принимаем.

Пример 6.17. По результатам 12 наблюдений найдены значения оценок коэффициентов корреляции $\widehat{\rho}_{01} = 0{,}64$, $\widehat{\rho}_{02} = 0{,}46$, $\widehat{\rho}_{12} = -0{,}07$. Найдем значения оценок *частных коэффициентов корреляции* $\rho_{01(2)}$, $\rho_{02(1)}$ и границы доверительного интервала для них с коэффициентом доверия $\gamma = 0{,}9$.

Значения оценок для частных коэффициентов корреляции находим по формулам (6.29):

$$\widehat{\rho}_{01(2)} = \frac{\widehat{\rho}_{01} - \widehat{\rho}_{02}\widehat{\rho}_{12}}{\sqrt{(1-\widehat{\rho}_{02}^2)(1-\widehat{\rho}_{12}^2)}} = \frac{0{,}64 - 0{,}46(-0{,}07)}{\sqrt{(1-0{,}46^2)(1-0{,}07^2)}} \approx 0{,}76,$$

$$\widehat{\rho}_{02(1)} = \frac{\widehat{\rho}_{02} - \widehat{\rho}_{01}\widehat{\rho}_{12}}{\sqrt{(1-\widehat{\rho}_{01}^2)(1-\widehat{\rho}_{12}^2)}} = \frac{0{,}46 - 0{,}64(-0{,}07)}{\sqrt{(1-0{,}64^2)(1-0{,}07^2)}} \approx 0{,}66.$$

Для вычисления границ доверительного интервала используем формулы (6.17), (6.18), в которых объем выборки следует понизить на величину k порядка частного коэффициента корреляции (см. **6.5**), т.е. в данном случае заменить n на $n-1$. По таблице квантилей нормального распределения (см. табл. П.2) находим квантиль $u_{(1+\gamma)/2} = u_{0{,}95} = 1{,}65$ и решаем уравнения

$$\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{10} = \ln\frac{1{,}76}{0{,}24} \pm \frac{1{,}65}{\sqrt{2}},$$

дающие границы доверительного интервала для показателя $\rho_{01(2)}$, и уравнения

$$\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{10} = \ln\frac{1{,}66}{0{,}34} \pm \frac{1{,}65}{\sqrt{2}},$$

дающие границы доверительного интервала для показателя $\rho_{02(1)}$. В результате получаем:

(0,38, 0,91) — доверительный интервал для $\rho_{01(2)}$;

(0,20, 0,87) — доверительный интервал для $\rho_{02(1)}$.

Пример 6.18. В условиях примера 6.17 найдем значение оценки *коэффициента детерминации* $\widehat{R}^2$.

Искомое значение $\widehat{R}^2$ вычисляем по формуле (6.33), используя полученное в примере 6.17 значение $\widehat{\rho}_{02(1)} \approx 0{,}66$:

$$\widehat{R}^2 = 1 - (1 - \widehat{\rho}_{01}^2)(1 - \widehat{\rho}_{02(1)}^2) =$$
$$= 1 - (1 - 0{,}64^2)(1 - 0{,}66^2) \approx 1 - 0{,}33 = 0{,}67.$$

Вопросы и задачи

6.1. Что такое вектор входных переменных (факторов), вектор выходных переменных (откликов)?

6.2. Перечислите основные задачи статистического исследования зависимостей.

6.3. Что называют корреляционным полем, корреляционной таблицей?

6.4. Запишите преобразование, используемое при построении доверительного интервала для ρ.

6.5. Какую статистику используют для проверки гипотезы H_0: $\rho = 0$?

6.6. Какую статистику используют при построении доверительного интервала для корреляционного отношения? По какому закону она распределена?

6.7. Какую статистику используют для проверки гипотезы о равенстве нулю корреляционного отношения?

6.8. Что называют частным коэффициентом корреляции? Запишите формулу для частных коэффициентов корреляции.

6.9. Что называют множественным коэффициентом корреляции, коэффициентом детерминации?

6.10. Какими свойствами обладает множественный коэффициент корреляции?

6.11. Запишите формулу, по которой может быть вычислен множественный коэффициент корреляции в случае нормального закона распределения?

6.12. Покажите, что из (6.32) следует (6.33).

6.13. Двумерная случайная величина имеет нормальный закон распределения. Определите значения границ доверительного интервала для коэффициента корреляции ρ с коэффициентом доверия $\gamma = 0{,}95$, если значение $\widehat{\rho}$, найденное по выборке объема $n = 300$, равно $-0{,}2$.

Ответ: $(-0{,}26, -0{,}14)$.

6.14. Двумерная случайная величина имеет нормальный закон распределения. Определите значения границ доверительного интервала для коэффициента корреляции ρ с коэффициентом доверия $\gamma = 0{,}9$, если значение $\widehat{\rho}$, найденное по выборке объема $n = 28$, равно $\widehat{\rho} = -0{,}36$.

Ответ: $(-0{,}70, -0{,}04)$.

6.15. В условиях предыдущей задачи проверьте гипотезу H_0: $\rho = 0$ при уровне значимости $\alpha = 0{,}05$ и альтернативной гипотезе H_1: $\rho < 0$.

Ответ: гипотеза отклоняется.

6.16. По выборке объема $n = 20$ (табл. 6.7) найдите значение оценки корреляционного отношения.

Таблица 6.7

x	1,0	1,0	1,5	2,0	2,0	2,5	2,5	3,0	3,0	3,5
y	0,2	0,3	0,3	0,3	0,4	0,4	0,5	0,5	0,5	0,8
x	3,5	4,0	4,0	4,5	5,0	5,5	5,5	5,5	6,0	6,5
y	0,7	1,1	1,0	1,2	1,7	2,3	2,2	2,4	2,7	3,3

Ответ: $\widehat{r}_{\eta\xi} = 0{,}95$.

6.17. По результатам 10 наблюдений, заданным таблицей (табл. 6.8), найдите:

а) значения оценок коэффициентов корреляции $\widehat{\rho}_{01}$, $\widehat{\rho}_{02}$, $\widehat{\rho}_{12}$;

б) значения оценок частных коэффициентов корреляции $\widehat{\rho}_{01(2)}$ и $\widehat{\rho}_{02(1)}$;

в) значения границ доверительного интервала для $\rho_{01(2)}$ и $\rho_{02(1)}$ с коэффициентом доверия 0,95;

г) значения оценки коэффициента детерминации.

Таблица 6.8

x_1	1	4	0	5	−3	3	−5	−1	2	−2
x_2	4	−6	2	−4	12	−2	14	6	0	8
y	−4	−5	4	−1	4	0	5	1	2	7

Ответ: а) $\widehat{\rho}_{12} = -0{,}98$; $\widehat{\rho}_{01} = -0{,}73$; $\widehat{\rho}_{02} = 0{,}69$; б) $\widehat{\rho}_{01(2)} = -0{,}36$; $\widehat{\rho}_{02(1)} = -0{,}15$; в) $-0{,}3 \pm 0{,}57$; $-0{,}15 \pm 0{,}64$; г) 0,54.

7. ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА

После обнаружения *стохастических связей* между изучаемыми переменными величинами (см. 6) исследователь приступает к математическому описанию интересующих его зависимостей. Для достижения этой цели ему необходимо решить следующие задачи:

1) подобрать класс функций, в котором целесообразно искать наилучшую (в определенном смысле) аппроксимацию искомой зависимости;

2) найти *оценки* для неизвестных значений параметров, входящих в уравнение искомой зависимости;

3) установить адекватность полученного уравнения искомой зависимости;

4) выявить наиболее информативные *входные переменные (факторы)*.

Совокупность перечисленных задач и составляет предмет исследований ***регрессионного анализа***.

7.1. Исходные предположения

Во многих прикладных задачах требуется построить *математическую модель*, связывающую *входные переменные* (*факторы*) $X_1, \ldots, X_p$ и *выходное переменное* (*отклик*) Y. В дальнейших рассуждениях будем предполагать, что Y является случайной величиной при каждом фиксированном наборе $\vec{x} = (x_1, \ldots, x_p)$ значений переменных $\vec{X} = (X_1, \ldots, X_p)$. В этом случае искомая *математическая модель* может быть представлена в следующем виде:

$$Y = f(\vec{x}) + \varepsilon(\vec{x}), \tag{7.1}$$

где $f(\vec{x})$ — скалярная функция, $\varepsilon(\vec{x})$ — *случайная ошибка*, т.е. случайная составляющая, порожденная либо действием случайных факторов, не включенных в набор $X_1, \ldots, X_p$, либо случайными ошибками измерений величины $f(\vec{x})$, либо и тем и другим одновременно.

Будем считать, что для каждого $\vec{x}$ математическое ожидание $\varepsilon(\vec{x})$ равно нулю, т.е. отсутствует ***систематическая погрешность*** модели. Следовательно для условного математического ожидания $\overline{y}(\vec{x}) = \mathrm{M}\left(Y \mid \vec{X} = \vec{x}\right)$ выходного переменного Y при условии, что вектор входных переменных $\vec{X}$ принял значение $\vec{x}$, согласно (7.1), имеем $\overline{y}(\vec{x}) = f(\vec{x})$.

Функцию $f(\vec{x})$, описывающую зависимость условного среднего значения $\overline{y}(\vec{x})$ выходного переменного Y от заданных фиксированных значений входных переменных $X_1, \ldots, X_p$, называют *функцией регрессии* (или *регрессией*).

Функция регрессии полностью определена, если известен условный закон распределения выходного переменного Y при условии, что $\vec{X} = \vec{x}$. Поскольку в реальных ситуациях никогда не располагают такой информацией, то обычно ограничиваются поиском подходящей аппроксимации $f_a(\vec{x})$ для $f(\vec{x})$, основываясь на *статистических данных* вида $\left(\vec{x}^i, y_i\right)$, $i = \overline{1, n}$, где $\vec{x}^i = \left(x_1^i, \ldots, x_p^i\right)$. Эти данные есть результат n независимых наблюдений $y_1, \ldots, y_n$ случайной величины Y при значениях входных переменных $\vec{x}^1 = \left(x_1^1, \ldots, x_p^1\right)$, $\vec{x}^2 = \left(x_1^2, \ldots, x_p^2\right)$, $\ldots$, $\vec{x}^n = \left(x_1^n, \ldots, x_p^n\right)$, т.е. результат специально организованного эксперимента.

Говоря о подходящей аппроксимации функции $f(\vec{x})$ — ***модели регрессии***, нужно, во-первых, задать класс ***допустимых моделей регрессии*** $\mathcal{F}$, т.е. класс функций, среди которых будем искать наилучшую аппроксимирующую функцию $f_a(\vec{x})$, и, во-вторых, выбрать *критерий*, по которому будем получать наилучшую аппроксимирующую функцию $f_a(\vec{x})$ из заданного класса $\mathcal{F}$.

Чтобы задать критерий, используют функцию $\rho(\varepsilon_f(\vec{X}))$, где $\varepsilon_f(\vec{X}) = f(\vec{X}) - f_a(\vec{X})$ — случайная величина, а $\rho(u)$ — некоторая неотрицательная функция аргумента u, как правило, *неубывающая* и *выпуклая*, например $\rho = u^2$ или $\rho = |u|$.

Функцию $f_a(x)$ считают наилучшей аппроксимирующей функцией из заданного класса $\mathcal{F}$, если она обеспечивает минимальное значение функционала

$$\Delta(f_a) = \mathrm{M}\,\rho(\varepsilon_f(\vec{X})) \quad \text{или} \quad \Delta_n(f_a) = \frac{1}{n}\sum_{i=1}^{n}\rho(\varepsilon_f(\vec{x}^i)),$$

где усреднение проводится по всем возможным значениям случайного вектора $\vec{X}$ в первом равенстве и по всем имеющимся наблюдениям — во втором.

В случае функции $\rho(u) = u^2$ получаемую ***регрессию*** называют ***средней квадратичной***, а метод, реализующий минимизацию функционала $\Delta_n(f_a)$, принято называть ***методом наименьших квадратов*** (***МНК***). Далее будем рассматривать только этот тип регрессии. Поэтому, говоря о регрессии, будем опускать слова „средняя квадратичная“.

В дальнейшем будем предполагать, что класс $\mathcal{F}$ допустимых моделей регрессии можно задать некоторым параметрическим семейством функций, т.е. представить в виде $\mathcal{F}_\beta = \{f_a(\vec{x};\vec{\beta})\}$, $\vec{\beta} \in \mathbf{R}^m$. Тогда задача отыскания наилучшей аппроксимации для $f(\vec{x})$ сводится к определению таких значений параметров $\hat{\vec{\beta}}$, при которых $\Delta_n(f_a)$ достигает минимума.

Следует отметить, что проблема выбора параметрического семейства функций $\mathcal{F}_\beta$, являясь ключевой в *регрессионном анализе*, не имеет, к сожалению, формализованных процедур для своего решения. Иногда выбор определяют на основе *экспериментальных данных* $(\vec{x}^i,\ y_i)$, $i = \overline{1,n}$ (см. пример 7.1), чаще — из теоретических соображений.

Например, если известно, что скорость протекания химической реакции между некоторыми компонентами пропорцио-

нальна объему исходного вещества, то объем вещества $V(t)$ в момент t изменяется по экспоненциальному закону

$$V(t) = \theta_0 e^{-\theta_1 (t - t_0)}, \quad t > t_0,$$

где θ_0, θ_1 — неизвестные параметры модели, которые нужно оценить наилучшим образом по результатам наблюдений, а t_0 — начальный момент времени.

К сожалению, такие случаи редки. Более реальной является ситуация, когда о механизме явления ничего не известно и можно лишь предполагать, что искомая функция $f(\vec{x})$ является достаточно гладкой. Тогда аппроксимирующая ее функция $f_a(\vec{x})$ может быть представлена в виде линейной комбинации некоторого набора линейно независимых ***базисных функций*** $\{\psi_k(\vec{x})\}$, $k = \overline{0, m-1}$, т.е. в виде

$$f_a(\vec{x}; \vec{\beta}) = \vec{\beta}^{\mathrm{T}} \vec{\psi}(\vec{x}) = \sum_{k=0}^{m-1} \beta^k \psi_k(\vec{x}), \tag{7.2}$$

где $\vec{\beta} = (\beta_0 \ \beta_1 \ \dots \ \beta_{m-1})^{\mathrm{T}}$ — вектор неизвестных параметров; $\vec{\psi} = (\psi_0 \ \psi_1 \ \dots \ \psi_{m-1})^{\mathrm{T}}$ — вектор базисных функций (известных заранее); m — число неизвестных параметров β_k, в общем случае неизвестная величина, уточняемая в ходе построения модели.

Следует заметить, что, согласно (7.2), *функция* $f_a(\vec{x}) = f_a(\vec{x}; \vec{\beta})$ является *линейной* по параметрам, представленным вектором $\vec{\beta}$. Поэтому в рассматриваемом случае говорят о ***модели, линейной по параметрам.***

Другими словами, исходный класс функций $\mathcal{F}$, содержащий истинную функцию регрессии $f(\vec{x})$, заменяют некоторым классом $\mathcal{F}_\beta = \{f_a(\vec{x}; \vec{\beta})\}$, $\vec{\beta} \in \mathbb{R}^m$, более простых по структуре функций, представимых в виде (7.2), и задача сводится к наилучшей оценке вектора неизвестных параметров $\vec{\beta} = (\beta_0, \beta_1, \dots, \beta_{m-1})^{\mathrm{T}}$.

При такой постановке задачи общая погрешность

$$\Delta = \sqrt{\sum_{i=1}^{n} \bigl(y_i - f_a(\vec{x}^{\,i})\bigr)^2}$$

от аппроксимации результатов наблюдений

$$y_i = f(\vec{x}^{\,i}) + \varepsilon_i, \quad i = \overline{1, n},$$

полученных в эксперименте, значениями функции $f_a(\vec{x}) \in \mathcal{F}_\beta$ обусловлена рассеянием отклика Y относительно истинной регрессии $f(\vec{x})$, т.е. величиной

$$\Delta_\varepsilon = \sqrt{\sum_{i=1}^{n} \bigl(y_i - f(\vec{x}^{\,i})\bigr)^2}$$

и систематической погрешностью аппроксимации, связанной с заменой исходного класса функций $\mathcal{F}$ более узким $\mathcal{F}_\beta \in \mathcal{F}$:

$$\Delta_a = \sqrt{\sum_{i=1}^{n} \bigl(f(\vec{x}^{\,i}) - f_a(\vec{x}^{\,i})\bigr)^2}.$$

Следовательно, приближение $f(\vec{x}) \approx f_a\bigl(\vec{x};\vec{\beta}\bigr)$ (см. 7.2) нужно понимать в том смысле, что систематической погрешностью Δ_a при замене класса $\mathcal{F}$ на $\mathcal{F}_\beta$ можно пренебречь по сравнению со случайной погрешностью Δ_ε. Именно на сопоставлении этих двух типов погрешностей и основаны правила проверки адекватности модели $f_a\bigl(\vec{x};\hat{\vec{\beta}}\bigr) = \hat{f}_a(\vec{x})$, где вектор параметров $\vec{\beta}$ заменен значением вектора оценок $\hat{\vec{\beta}}$.

Одним из наиболее распространенных аппроксимирующих классов функций $\mathcal{F}_\beta$ является класс полиномов, в котором в качестве базисных функций выбраны степени переменных $x_1, \ldots, x_p$.

Простейшей полиномиальной моделью является модель 1-го порядка, линейная по всем переменным:

$$f_a(x) = \sum_{k=0}^{m-1} \beta_k x_k, \quad m \leqslant p+1,$$

где $x_0 \equiv 1$ — фиктивное переменное, т.е. здесь $\psi_0(x) \equiv 1$, $\psi_1(x) = x_1, \ldots, \psi_{m-1}(x) = x_{m-1}$.

Следует подчеркнуть, что представление (7.2) является самым общим видом линейной по параметрам модели и описывает не только полиномиальные модели. Например, в качестве базисных функций $\psi_k(x)$ могут выступать тригонометрические функции $\sin kx$, $\cos kx$, показательные e^{kx} и др.

Если неизвестная функция регрессии $f(\vec{x})$ представлена в виде (7.2), то задача ее поиска сведена тем самым к оценке вектора неизвестных параметров $\vec{\beta} = (\beta_0, \ldots, \beta_{m-1})^{\mathrm{T}}$ и последующей проверке качества аппроксимации $f(\vec{x}) \approx f_a(\vec{x})$, т.е. адекватности модели $f_a(x)$. Если модель (7.2) окажется неадекватной, то вид аппроксимирующей функции $f_a(x)$ нужно уточнять либо увеличением числа m базисных функций, либо заменой самих базисных функций другими, более подходящими.

Пример 7.1. Анализируется поведение двумерной случайной величины (X, Y), где X — возраст (в годах) наугад выбранного школьника из группы в $n = 40$ человек, а Y — масса его тела (в кг). На рис. 7.1 исходные статистические данные (x_i, y_i), $i = \overline{1, n}$, отмечены крестиками. Поскольку имелась возможность контролировать значения входной переменной X, то это позволило разбить обследованную группу школьников на четыре равные по объему подгруппы с примерно одинаковым возрастом.

На рис. 7.1 видно, что в пределах каждой подгруппы рост подвержен неконтролируемому разбросу, т.е. налицо отмеченный выше стохастический характер связи между X и Y. Однако расположение точек (x_i, y_i) на плоскости xOy обнаруживает

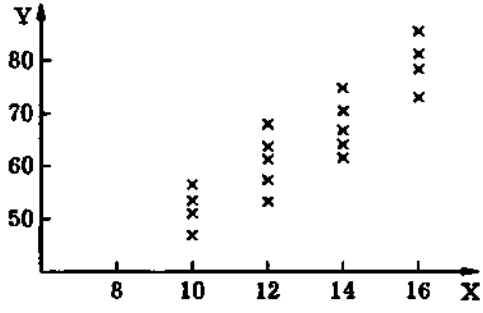


Рис. 7.1

вполне определенную тенденцию, характеризующую увеличение „в среднем" массы тела Y при увеличении возраста в рассматриваемый период интенсивного роста (от 12 до 16 лет).

Целью проведенного исследования является прогноз роста конкретного школьника по заданному значению его возраста и определение среднего роста $\overline{y}(x)$ школьников, достигших возраста x.

Для достижения этой цели необходимо математически описать закономерность изменения условных средних значений $\overline{y}(x) = \mathrm{M}(Y \mid X = x)$ в зависимости от значения x случайного переменного X, а также изучить характер случайного разброса массы тела Y отдельных школьников возраста x относительно своего среднего значения $\overline{y}(x)$.

Таким образом, возникла необходимость рассмотрения математической модели (7.1), где $\varepsilon(x)$ — случайное отклонение массы тела Y школьников возраста x от среднего значения $\overline{y}(x) = \mathrm{M}(Y \mid X = x)$. Если $\mathrm{M}(\varepsilon \mid X = x) = 0$ при любых x, то $\overline{y}(x) = f(x)$, и построение искомой зависимости сводится к отысканию функции $f(x)$, описывающей изменение условного среднего значения выходного переменного Y при различных значениях $X = x$ входного переменного X.

Остается определить, в каком классе $\mathcal{F}$ функций мы будем искать аппроксимацию для $f(x)$. Для нашего примера по

расположению точек $(x_i,\ y_i)$, $i=\overline{1,n}$, можно заключить, что

$$f(x)=\beta_0+\beta_1 x,$$

где β_0 и β_1 — неизвестные параметры модели, т.е. $\mathcal{F}_\beta = \{f(x;\beta_0,\beta_1)\}$ есть класс полиномов первого порядка, к которому принадлежит функция регрессии $f(x)$. Значения оценок $\widehat{\beta}_0$, $\widehat{\beta}_1$ параметров β_0, β_1 можно найти с помощью метода наименьших квадратов (см. 7.2)

Матричная форма записи линейной регрессионной модели. Результаты эксперимента для исследования связи между откликом Y и вектором факторов $\vec{X}=(X_1,\ \ldots,\ X_p)$ удобно представлять в виде матрицы D исходных данных:

$$D=\begin{pmatrix} \vec{x}^1 & \vec{x}^2 & \ldots & \vec{x}^i & \ldots & \vec{x}^n \\ r_1 & r_2 & \ldots & r_i & \ldots & r_n \\ \vec{y_1} & \vec{y_2} & \ldots & \vec{y_i} & \ldots & \vec{y_n} \end{pmatrix}, \qquad \sum_{i=1}^{n} r_i = N,$$

где $\vec{x}^i=\left(x_1^i,\ \ldots,\ x_p^i\right)$, $i=\overline{1,n}$, — различные значения вектора факторов $\vec{X}$, для которых проводился эксперимент; r_i — число независимых повторных (параллельных) опытов для $\vec{x}^i$; N — общее число наблюдений за откликом Y; $\vec{y_i}=(y_{i1},\ \ldots,\ y_{ir_i})$, $i=\overline{1,n}$, — значения отклика Y, полученные в эксперименте для значения $\vec{x}^i$ вектора факторов. Заметим, что матрицу

$$P=\begin{pmatrix} \vec{x}^1 & \vec{x}^2 & \ldots & \vec{x}^i & \ldots & \vec{x}^n \\ r_1 & r_2 & \ldots & r_i & \ldots & r_n \end{pmatrix},$$

образованную двумя первыми строками матрицы D, называют часто ***планом эксперимента***, совокупность возможных значений вектора факторов $\vec{X}$ называют ***факторным пространством*** и обозначают $\mathcal{X}^p$.

Если $r_i=1$, $i=\overline{1,n}$, то результаты эксперимента представляют собой n точек $\left(\vec{x}^i,\ y_i\right)$, $i=\overline{1,n}$, в пространстве $\mathbb{R}^{p+1}$.

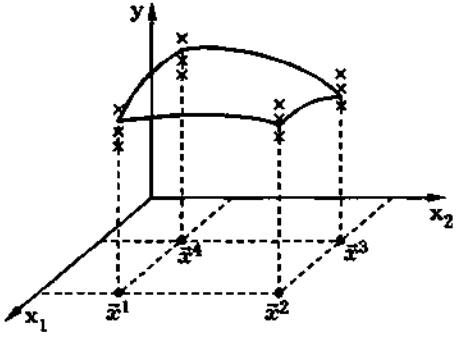


Рис. 7.2

Геометрическая интерпретация матрицы D представлена на рис. 7.2 для $p=2$, $n=4$, $r_i=3$, $i=\overline{1,4}$ (крестиками отмечены соответствующие значения отклика Y).

Для удобства дальнейших рассуждений в соответствии с равенством (7.1) будем считать, что значению $\vec{x}^i=(x_1^i,\ \ldots,\ x_p^i)$ вектора факторов $\vec{X}^i=(X_1^i,\ \ldots,\ X_p^i)$ соответствует отклик Y_i и случайная ошибка $\varepsilon_i=\varepsilon(\vec{x}^i)$, т.е.

$$Y_i=f(\vec{x}^i)+\varepsilon_i. \tag{7.3}$$

При этом в случае модели, линейной по параметрам, согласно (7.2), имеем

$$Y_i=\sum_{k=0}^{m-1}\beta_k\psi_k(\vec{x}^i)+\varepsilon_i,\quad i=\overline{1,n}. \tag{7.4}$$

Если на основе системы равенств (7.4), которая содержит в себе всю информацию, полученную в эксперименте, мы сумеем оценить неизвестные параметры β_k (некоторым наилучшим образом), т.е. сумеем найти значения $\widehat{\beta}_k\approx\beta_k$, то тем самым будет найдена наилучшая (для выбранных базисных функций) модель следующего вида:

$$\widehat{f}_a(\vec{x})=\sum_{k=0}^{m-1}\widehat{\beta}_k\psi_k(\vec{x}). \tag{7.5}$$

Эта модель будет наилучшей в классе $\mathcal{F}_\beta$ для выбранного набора базисных функций $\psi_i(x)$, $i=\overline{1,m-1}$. При этом общую погрешность Δ можно уменьшить лишь за счет уменьшения погрешности аппроксимации Δ_a, связанной с выбором класса аппроксимирующих функций $\mathcal{F}_\beta$ (если удачно подобрать как сами функции $\psi_k(x)$, так и их количество m).

Таким образом, модель (7.5) требует в общем случае проверки на адекватность (на соответствие результатам эксперимента) и при необходимости уточнения (это рассмотрено ниже, (см. 7.3).

Введем в рассмотрение следующие матрицы:

– ***матрицу отклика*** $Y=(Y_1,\ \ldots,\ Y_n)^T$ типа $n\times 1$, если повторных опытов не было (т.е. $r_i=1$, $i=\overline{1,n}$), или ***матрицу выборочных средних значений отклика*** $\overline{Y}$ типа $n\times 1$ в противном случае, i-й элемент которой есть

$$\overline{Y}_i=\frac{1}{r_i}\sum_{j=1}^{r_i}Y_{ij},\quad i=\overline{1,n};$$

– ***матрицу*** F ***базисных функций*** (***матрицу наблюдений***) типа $n\times m$

$$F=\begin{pmatrix}\psi_0(\vec{x}^1) & \psi_1(\vec{x}^1) & \ldots & \psi_{m-1}(\vec{x}^1)\\ \psi_0(\vec{x}^2) & \psi_1(\vec{x}^2) & \ldots & \psi_{m-1}(\vec{x}^2)\\ \ldots & \ldots & \ldots & \ldots\\ \psi_0(\vec{x}^n) & \psi_1(\vec{x}^n) & \ldots & \psi_{m-1}(\vec{x}^n)\end{pmatrix};$$

– ***матрицу*** (***вектор-столбец***) ***ошибок*** $\vec{\varepsilon}=(\varepsilon_1,\ \ldots,\ \varepsilon_n)^T$ типа $n\times 1$ и вектор-столбец $\vec{\beta}=(\beta_0,\ \ldots,\ \beta_{m-1})^T$ параметров модели.

Тогда систему равенств (7.3) можно представить в матричном виде:

$$Y=F\vec{\beta}+\vec{\varepsilon}. \tag{7.6}$$

Уравнение (7.6) называют ***линейной регрессионной моделью***. Подчеркнем, что линейность в этой модели понимается как линейность по параметрам $\beta_0, \beta_1, \ldots, \beta_{m-1}$, называемым также ***коэффициентами регрессии***. По переменным $X_1, \ldots, X_p$ модель (7.6) может быть (и, как правило, так и бывает) нелинейной. Возможные ситуации рассмотрены в примере 7.2.

Замечание 7.1. При наличии повторных опытов в равенстве (7.6) вместо матрицы Y будет стоять матрица $\overline{Y}$. #

Рассмотрим возможные конкретные случаи реализации соотношения (7.2), которые приводят к общей модели (7.6).

Пример 7.2. а. Пусть имеется лишь один фактор X (т.е. $p=1$), а множество точек (x_i, y_i), $i=\overline{1,n}$, расположено на плоскости xOy вдоль некоторой прямой (рис. 7.3, *а*). В этом случае в качестве функции $f_a(x)$, аппроксимирующей функцию регрессии $f(x)=\mathrm{M}(Y\,|\,x)$, естественно взять линейную функцию

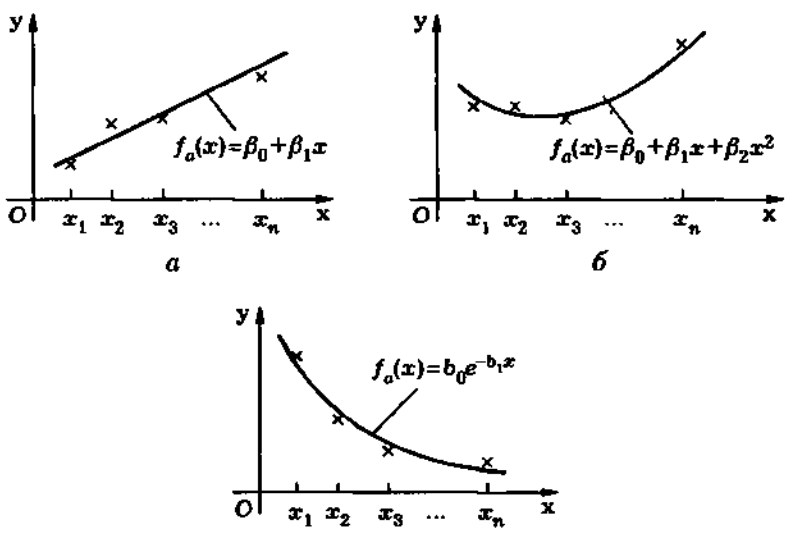


Рис. 7.3

аргумента x:

$$f_a(x) = \beta_0 + \beta_1 x,$$

т.е. в качестве базисных функций здесь выбраны $\psi_0(x) \equiv 1$ и $\psi_1(x) = x$. Такую регрессию называют ***простой линейной регрессией.***

Если множество точек (x_i, y_i), $i = \overline{1, n}$, расположено вдоль некоторой кривой (рис. 7.3, *б*), то в качестве $f_a(x)$ естественно попробовать выбрать семейство парабол:

$$f_a(x) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

т.е. в качестве базисных функций здесь выступают функции $\psi_0(x) \equiv 1$ и $\psi_1(x) = x$, $\psi_2(x) = x^2$.

Наконец, в случае расположения точек (x_i, y_i), $i = \overline{1, n}$, показанного на рис. 7.3, *в*, можно попробовать подобрать функцию $f_a(x)$ из семейства экспонент:

$$f_a(x) = \beta_0 e^{-\beta_1 x}.$$

В последнем случае функция $f_a(x)$ является нелинейной по параметрам β_0 и β_1 и не приводит к линейной регрессионной модели (7.5). Однако после некоторого функционального преобразования нелинейную по параметрам функцию $f_a(x)$ часто можно привести к функции $\widetilde{f}_a(x)$, линейной по параметрам. В данном случае после логарифмирования получаем

$$\ln f_a(x) = \ln \beta_0 - \beta_1 x,$$

т.е. функция $\widetilde{f}_a(x) = \ln f_a(x)$ уже линейна по параметрам $\theta_0 = \ln \beta_0$ и $\theta_1 = -\beta_1$.

б. Пусть имеется два фактора X_1 и X_2 (т.е. $p = 2$), а множество точек (x^i, y_i), $i = \overline{1, n}$, где $x^i = (x_1^i, x_2^i)$, расположены вдоль некоторой плоскости в пространстве трех переменных y, x_1 и x_2. Тогда набор наилучшей аппроксимации $f_a(x)$ можно начинать с линейной по переменным X_1 и X_2 функции

$$f_a(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

т.е. выбрать в качестве базисных функций $\psi_0(x) \equiv 1$, $\psi_1(x) = x_1$ и $\psi_2(x) = x_2$.

Если точки (x_i, y_i), $i = \overline{1,n}$, расположены в пространстве переменных y, x_1, x_2 так, что есть основание предполагать наличие у функции $f(x)$ точки экстремума, то естественно искать $f_a(x)$ среди полиномов второго порядка, т.е. принять

$$f_a(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2.$$

В этом случае базисными функциями будут $\psi_0(x) \equiv 1$, $\psi_1(x) = x_1$, $\psi_2(x) = x_2$, $\psi_3(x) = x_1 x_2$, $\psi_4(x) = x_1^2$, $\psi_5(x) = x_2^2$.

в. В качестве базисных функций могут быть выбраны не только степени переменных $x_1, \ldots, x_p$, но, вообще говоря, любые линейно независимые функции, не содержащие неизвестных параметров. Например, при $\psi_0(\vec{x}) = 1$, $\psi_1(\vec{x}) = e^{x_1 + x_2}$, $\psi_2(\vec{x}) = \sin x_1$ получаем линейную по параметрам модель регрессии

$$f_a(x) = \beta_0 + \beta_1 e^{x_1 + x_2} + \beta_2 \sin x_1.$$

7.2. Метод наименьших квадратов

Матрицы F и Y в *линейной регрессионной модели* (7.6) содержат всю информацию, получаемую в результате эксперимента. По этим данным нам нужно оценить вектор неизвестных параметров $\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_{m-1})^{\mathrm{T}}$. Для получения оценок, как отмечалось выше, будем использовать *метод наименьших квадратов*. Предварительно сформулируем предположения, лежащие в его основе.

1. $\mathrm{M}\varepsilon_i = 0$, $i = \overline{1,n}$, т.е. систематическая *погрешность модели* отсутствует.

2. $\mathrm{M}(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$, т.е. *случайные ошибки* некоррелированы (это ограничение можно снять, если матрица ковариаций $\mathrm{D}(\vec{\varepsilon})$ *вектор-столбца ошибок* известна*).

*См.: *Ивченко Г.И., Медведев Ю.И.*

3. $\mathbf{D}\varepsilon_i = \mathbf{M}\varepsilon_i^2 = \sigma^2$, $i = \overline{1,n}$, т.е. в любых точках *факторного пространства* $\mathcal{X}^p$ случайные ошибки имеют одинаковую дисперсию.

4. Значения x_i переменных X_i, $i = \overline{1,p}$, в процессе эксперимента измеряются без ошибок.

Отметим, что предположения 2 и 3 можно объединить и представить в следующем виде:

$$\mathbf{D}\vec{\varepsilon} = \sigma^2 I_n,$$

где I_n — единичная матрица порядка n.

Четвертое предположение означает, что, согласно соотношениям (7.3), верны равенства

$$\mathbf{M}Y_i = \sum_{k=0}^{m-1} \beta_k \psi_k(\vec{x}^i), \quad \mathbf{D}Y_i = \mathbf{D}\varepsilon_i = \sigma^2, \quad i = \overline{1,n},$$

которые в матричной записи имеют вид

$$\mathbf{M}Y = F\vec{\beta}, \qquad \mathbf{D}Y = \sigma^2 I_n.$$

Подчеркнем, что никаких предположений о законе распределения случайных величин Y_i, $i = \overline{1,n}$, мы пока не делаем.

Теорема 7.1. Пусть $M = F^{\mathrm{T}}F$ — невырожденная матрица. *Несмещенной эффективной оценкой* в классе всех *линейных оценок* для параметра $\vec{\beta} = (\beta_0 \ \beta_1 \ \dots \ \beta_{m-1})^{\mathrm{T}}$ в *линейной регрессионной модели* (7.6) является ***оценка метода наименьших квадратов (МНК-оценка)***, определяемая матричным равенством

$$\hat{\vec{\beta}}(\vec{Y}_n) = (F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}Y. \quad \# \tag{7.7}$$

Поясним идею метода наименьших квадратов и происхождение формулы (7.7). Докажем несмещенность и эффективность оценки $\hat{\vec{\beta}}(\vec{Y}_n)$ в классе линейных оценок.

Пусть отклик Y зависит лишь от одного фактора X $(p=1)$, а искомая функция регрессии $\mathrm{M}(Y\,|\,x)=f(x)$ имеет график, изображенный пунктирной линией на рис. 7.4. Функция $f(x)$ нам не известна, известны лишь значения отклика $y_1, \ldots, y_n$, полученные в эксперименте при значениях факторов $x_1, \ldots, x_n$ (на рис. 7.4 точки (x_i, y_i), $i=\overline{1,n}$, отмечены „крестиками“).

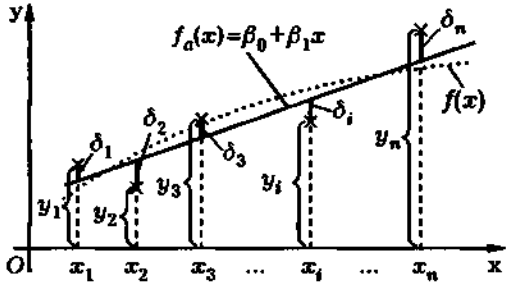


Рис. 7.4

Неизвестную функцию $f(x)$ на основании характера расположения экспериментальных точек (они визуально расположены вдоль прямой) естественно аппроксимировать линейной функцией $f_a\big(x;\vec{\beta}\big)=\beta_0+\beta_1 x$. Отклонения $\delta_i=y_i-(\beta_0+\beta_1 x_i)$, $i=\overline{1,n}$, ординат экспериментальных точек (x_i, y_i) от любой прямой $f_a\big(x;\vec{\beta}\big)=\beta_0+\beta_1 x$ называют ***невязками***.

В общем случае для линейной регрессионной модели (7.6)

$$\delta_i=y_i-\sum_{k=0}^{m-1}\beta_k\psi_k(x_i),\quad i=\overline{1,n},$$

и невязку δ_i можно рассматривать как реализацию случайной ошибки $\varepsilon_i=\varepsilon(x_i)$, $i=\overline{1,n}$.

Согласно методу наименьших квадратов, оценку $\hat{\vec{\beta}}(\vec{Y}_n)=$ $=\big(\hat{\beta}_0(\vec{Y}_n)\ \ldots\ \hat{\beta}_{m-1}(\vec{Y}_n)\big)^{\mathrm{T}}$ вектора параметров $\vec{\beta}=(\beta_0\ \ldots\ \beta_{m-1})^{\mathrm{T}}$ выбирают так, чтобы сумма квадратов невязок δ_i была мини-

мальной, т.е.

$$\Delta(\vec{\beta}) = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} \Big(y_i - \sum_{k=0}^{m-1} \beta_k \psi_k(x^i) \Big)^2 \to \min,$$

или, что то же самое,

$$\Delta(\vec{\beta}) = \delta^{\mathrm{T}} \delta = (Y - F\vec{\beta})^{\mathrm{T}} (Y - F\vec{\beta}) = \Delta(\vec{\beta}) \to \min,$$

где $\delta = (\delta_1, \ldots, \delta_n)$ — вектор невязок.

Необходимым условием экстремума функции $\Delta(\vec{\beta})$ переменных $\beta_0, \beta_1, \ldots, \beta_{m-1}$ (а, следовательно, и условием существования МНК-оценки параметра $\vec{\beta}$), как известно, является равенство нулю ее частных производных [V], т.е.

$$\frac{\partial \Delta}{\partial \beta_\nu} = -2 \sum_{i=1}^{n} \psi_\nu(\vec{x}^i) \Big(y_i - \sum_{k=0}^{m-1} \psi_k(\vec{x}^i) \beta_k \Big) = 0, \quad \nu = \overline{0, m-1}.$$

Эту систему можно представить, используя матричную запись $-2F^{\mathrm{T}}(Y - F\vec{\beta}) = 0$, или

$$F^{\mathrm{T}} F \vec{\beta} = F^{\mathrm{T}} Y. \tag{7.8}$$

Из геометрических соображений очевидно, что решением системы (7.8) является точка минимума функции $\Delta(\beta)$, в чем можно убедиться непосредственно, воспользовавшись достаточным условием экстремума [V]. Систему линейных алгебраических уравнений (7.8) называют ***системой нормальных уравнений*** Гаусса. Она всегда имеет решение (хотя не всегда единственное).

Пусть матрица $F^{\mathrm{T}}F$ имеет обратную матрицу $(F^{\mathrm{T}}F)^{-1}$ (для этого необходимо и достаточно, чтобы $\operatorname{rang} F$ был равен числу столбцов матрицы F). Тогда, умножая обе части равенства (7.8) слева на матрицу $(F^{\mathrm{T}}F)^{-1}$, приходим к формуле (7.7), которая дает единственное решение системы (7.8).

Если матрица $F^{\mathrm{T}}F$ не имеет обратной (случай, когда ранг матрицы F меньше числа m ее столбцов), то МНК-оценка параметра $\vec{\beta}$ существует, но не является единственной*.

*См.: *Рао С.Р.*

Несмещенность оценки $\widehat{\vec{\beta}}(\vec{Y}_n)$, заданной равенством (7.7), и эффективность в классе всех линейных несмещенных оценок непосредственно следуют из исходных предположений для метода наименьших квадратов. Действительно,

$$\mathrm{M}\widehat{\vec{\beta}}(\vec{Y}_n) = \mathrm{M}\big((F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}Y\big) = \\ = (F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\mathrm{M}Y = (F^{\mathrm{T}}F)^{-1}FF^{\mathrm{T}}\vec{\beta} = \vec{\beta},$$

т.е. $\widehat{\vec{\beta}}(\vec{Y}_n)$ — несмещенная оценка для β. Докажем ее эффективность.

Пусть LY — произвольная линейная несмещенная оценка для $\vec{\beta}$. Тогда из равенства

$$\mathrm{M}(LY) = L\mathrm{M}Y = LF\vec{\beta} = \vec{\beta}$$

получаем $LF = I_n$ и

$$\mathrm{D}(LY) = \mathrm{M}(LY - L\mathrm{M}Y)^2 = \mathrm{M}L(Y - \mathrm{M}Y)^2L^{\mathrm{T}} = \\ = L\mathrm{D}YL^{\mathrm{T}} = L\sigma^2I_nL^{\mathrm{T}} = \sigma^2LL^{\mathrm{T}}.$$

Наша задача минимизировать диагональные элементы матрицы LL^{T}, которые с точностью до множителя σ^2 являются дисперсиями оценок параметров β_k, $k = \overline{0, m-1}$. Для этого рассмотрим равенство

$$LL^{\mathrm{T}} = (M^{-1}F^{\mathrm{T}})(M^{-1}F^{\mathrm{T}})^{\mathrm{T}} + (L - M^{-1}F^{\mathrm{T}})(L - M^{-1}F^{\mathrm{T}})^{\mathrm{T}},$$

в справедливости которого можно убедиться непосредственно, перемножив матрицы в правой его части с учетом равенств $LF = I_n$ и $M = F^{\mathrm{T}}F$. Поскольку диагональные элементы матрицы вида AA^{T} являются неотрицательными, то можно утверждать, что диагональные элементы матрицы LL^{T} будут минимальными, если $L = M^{-1}F^{\mathrm{T}}$, т.е. оценка $\widehat{\vec{\beta}}(\vec{Y}_n)$ является эффективной в классе всех линейных оценок.

Итак, по теореме 7.1 МНК-оценки являются наилучшими в указанном выше смысле в классе линейных оценок. Тем самым равенство (7.5) определяет наилучшую ***модель регрессии*** для выбранных базисных функций и значений $\widehat{\beta}_k$, $k = \overline{0, m-1}$, найденных по методу наименьших квадратов, которую будем записывать (обозначив $\widehat{f}_a(\vec{x}) = \widehat{y}(\vec{x})$) в виде

$$\widehat{y}(\vec{x}) = \sum_{k=0}^{m-1} \widehat{\beta}_k \psi_k(\vec{x}). \tag{7.9}$$

Случайную величину

$$\widehat{Y}(\vec{x}) = \sum_{k=0}^{m-1} \widehat{\beta}_k(\vec{Y}_n)\psi_k(\vec{x})$$

будем называть ***оценкой среднего значения отклика*** Y.

Согласно (7.9), можно определить оценки $\widehat{Y}_i = \widehat{Y}(\vec{x^i})$ среднего значения отклика (условного математического ожидания отклика) в каждой точке $\vec{x}^i$ факторного пространства:

$$\widehat{Y}_i = \sum_{k=0}^{m-1} \widehat{\beta}_k(\vec{Y}_n)\,\psi_k(\vec{x}^i), \quad i = \overline{1, n}.$$

При этом, если ввести ***матрицу*** $\widehat{Y} = \left(\widehat{Y}_1, \ldots, \widehat{Y}_n\right)^{\mathrm{T}}$ ***оценок среднего значения отклика***, то

$$\widehat{Y} = F\widehat{\vec{\beta}}(\vec{Y}_n). \tag{7.10}$$

Замечание 7.2. В ряде случаев интерес представляют не сами параметры $\beta_0, \beta_1, \ldots, \beta_{m-1}$ в линейной регрессионной модели (7.6), а их некоторые линейные комбинации, т.е. новый вектор параметров $\vec{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{q-1})$, $q \leqslant m$, связанный с вектором $\vec{\beta} = (\beta_0, \ldots, \beta_{m-1})^{\mathrm{T}}$ соотношением $\vec{\alpha} = A\vec{\beta}$, где A — некоторая матрица типа $q \times m$.

Для вектора $\vec{\alpha}$ МНК-оценка $\widehat{\vec{\alpha}}(\vec{Y}_n)$ определяется равенством

$$\widehat{\vec{\alpha}}(\vec{Y}_n) = A\widehat{\vec{\beta}}(\vec{Y}_n), \tag{7.11}$$

где $\widehat{\vec{\beta}}(\vec{Y}_n)$ — МНК-оценка для $\vec{\beta}$. #

Укажем теперь правило определения ковариационной матрицы $\Sigma(\widehat{\vec{\beta}})$ МНК-оценки $\widehat{\vec{\beta}}(\vec{Y}_n)$ вектора параметров $\vec{\beta}$. Это правило будет вытекать как частный случай из следующей теоремы.

Теорема 7.2. Пусть $\vec{\beta}$ — m-мерный вектор-столбец линейной регрессионной модели (7.6), A — произвольная матрица типа $q \times m$, где $1 \leqslant q \leqslant m$, а матрица $F^{\mathrm{T}}F$ является обратимой (т.е. $\det(F^{\mathrm{T}}F) \neq 0$). Тогда для вектора $\vec{\alpha} = A\vec{\beta}$ МНК-оценка $\widehat{\vec{\alpha}}(\vec{Y}_n)$, определяемая равенством (7.11), является несмещенной оценкой с матрицей ковариаций

$$\Sigma(\widehat{\vec{\alpha}}) = \sigma^2 A(F^{\mathrm{T}}F)^{-1}A^{\mathrm{T}}, \tag{7.12}$$

где σ^2 — дисперсия отклика.

◀ Согласно (7.7) и (7.6), имеем

$$\widehat{\vec{\beta}}(\vec{Y}_n) = (F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}(F\vec{\beta}+\vec{\varepsilon}) =$$
$$= (F^{\mathrm{T}}F)^{-1}(F^{\mathrm{T}}F)\vec{\beta}+(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\vec{\varepsilon} = \vec{\beta}+(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\vec{\varepsilon}.$$

Отсюда для оценки $\widehat{\vec{\alpha}}(\vec{Y}_n) = A\widehat{\vec{\beta}}(\vec{Y}_n)$ параметра $\vec{\alpha} = A\vec{\beta}$ имеем представление

$$\widehat{\vec{\alpha}}(\vec{Y}_n) = A\widehat{\vec{\beta}}(\vec{Y}_n) = A\vec{\beta}+A(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\vec{\varepsilon} =$$
$$= \vec{\alpha}+A(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\vec{\varepsilon}, \tag{7.13}$$

из которого вытекает несмещенность оценки $\widehat{\vec{\alpha}}(\vec{Y}_n)$, так как, согласно исходным предположениям метода наименьших квад-

ратов, $\mathbf{M}\vec{\varepsilon} = \vec{0}$ и, следовательно,

$$\mathbf{M}\hat{\vec{\alpha}}(\vec{Y}_n) = \vec{\alpha} + A(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\mathbf{M}\vec{\varepsilon} = \vec{\alpha}.$$

Далее, матрица ковариаций вектора МНК-оценок $\hat{\vec{\alpha}}(\vec{Y}_n)$ в силу несмещенности оценки $\vec{\alpha}$ имеет вид

$$\Sigma(\hat{\vec{\alpha}}) = \mathbf{M}\Big(\big(\hat{\vec{\alpha}}(\vec{Y}_n) - \vec{\alpha}\big)\big(\hat{\vec{\alpha}}(\vec{Y}_n) - \vec{\alpha}\big)^{\mathrm{T}}\Big).$$

Используя представление (7.13), преобразуем выражение для $\Sigma(\hat{\vec{\alpha}})$ следующим образом:

$$\Sigma(\hat{\vec{\alpha}}) = \mathbf{M}\Big(\big(A(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\vec{\varepsilon}\big)\big(A(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\vec{\varepsilon}\big)^{\mathrm{T}}\Big) =$$

$$= \mathbf{M}\big(A(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\vec{\varepsilon}\vec{\varepsilon}^{\mathrm{T}}F(F^{\mathrm{T}}F)^{-1}A^{\mathrm{T}}\big) =$$

$$= A(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\mathbf{M}\big(\vec{\varepsilon}\vec{\varepsilon}^{\mathrm{T}}\big)F(F^{\mathrm{T}}F)^{-1}A^{\mathrm{T}}.$$

(При переходе к правой части мы воспользовались правилом транспонирования произведения матриц $(AB)^{\mathrm{T}} = B^{\mathrm{T}}A^{\mathrm{T}}$ [III] и симметричностью матрицы $(F^{\mathrm{T}}F)^{-1}$.)

Если теперь учесть, что, согласно исходным предположениям метода наименьших квадратов, $\mathbf{M}\big(\vec{\varepsilon}^{\mathrm{T}}\vec{\varepsilon}\big) = I_n\sigma^2$, то

$$\Sigma(\hat{\vec{\alpha}}) = A(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}I_n\sigma^2F(F^{\mathrm{T}}F)^{-1}A^{\mathrm{T}} =$$

$$= \sigma^2A(F^{\mathrm{T}}F)^{-1}(F^{\mathrm{T}}F)(F^{\mathrm{T}}F)^{-1}A^{\mathrm{T}} = \sigma^2A(F^{\mathrm{T}}F)^{-1}A^{\mathrm{T}},$$

что и доказывает представление (7.12). ▶

Следствие 7.1. Если $A = I_m$, т.е. $\vec{\alpha} = A\vec{\beta} = \vec{\beta}$, то

$$\Sigma(\hat{\vec{\beta}}) = \sigma^2C, \tag{7.14}$$

где $C = (F^{\mathrm{T}}F)^{-1}$ — ***дисперсионная матрица Фишера***.

Следствие 7.2. Дисперсия оценки $\widehat{Y}(\vec{x})$ среднего значения отклика в произвольной точке $\vec{x}$ факторного пространства $\mathcal{X}^p$

определяется по формуле

$$\mathbf{D}\hat{Y}(\vec{x}) = \sigma^2 \psi^{\mathrm{T}}(\vec{x}) C \psi(\vec{x}), \tag{7.15}$$

где $\psi^{\mathrm{T}}(\vec{x}) = (\psi_0(\vec{x}), \ldots, \psi_{m-1}(\vec{x}))$.

◀ Действительно, согласно (7.9) и (7.14), имеем

$$\begin{aligned}
\mathbf{D}\hat{Y}(\vec{x}) &= \mathbf{D}\big(\psi^{\mathrm{T}}(\vec{x})\hat{\vec{\beta}}(\vec{Y}_n)\big) = \\
&= \mathbf{M}\big(\psi^{\mathrm{T}}(\vec{x})(\hat{\vec{\beta}}(\vec{Y}_n) - \vec{\beta})(\hat{\vec{\beta}}(\vec{Y}_n) - \vec{\beta})^{\mathrm{T}}\psi(\vec{x})\big) = \\
&= \psi^{\mathrm{T}}(\vec{x})\,\mathbf{M}\big((\hat{\vec{\beta}}(\vec{Y}_n) - \vec{\beta})(\hat{\vec{\beta}}(\vec{Y}_n) - \vec{\beta})^{\mathrm{T}}\big)\psi(\vec{x}) = \\
&= \psi^{\mathrm{T}}(\vec{x})\Sigma(\hat{\vec{\beta}})\psi(\vec{x}) = \sigma^2\psi^{\mathrm{T}}(\vec{x})C\psi(\vec{x}). \qquad \blacktriangleright
\end{aligned}$$

Формулы (7.14) и (7.15) содержат неизвестный параметр σ^2 — дисперсию отклика Y. Поэтому требуется правило определения оценки параметра σ^2. Такое правило устанавливает следующая теорема. Однако прежде чем формулировать теорему, отметим, что случайную величину

$$\hat{\vec{\varepsilon}}^{\mathrm{T}}\hat{\vec{\varepsilon}} = \big(Y - F\hat{\vec{\beta}}(\vec{Y}_n)\big)^{\mathrm{T}}\big(Y - F\hat{\vec{\beta}}(\vec{Y}_n)\big), \tag{7.16}$$

где $\hat{\vec{\varepsilon}} = Y - F\hat{\vec{\beta}}(\vec{Y}_n)$ — случайный вектор, а $\hat{\vec{\beta}}(\vec{Y}_n)$ — МНК-оценка вектора параметров $\vec{\beta}$ линейной регрессионной модели (7.6), называют ***остаточной суммой квадратов***.

Теорема 7.3. Если выполнены исходные предположения метода наименьших квадратов и ранг матрицы базисных функций F типа $n \times m$ равен m, то несмещенная оценка для ***остаточной дисперсии*** σ^2 определяется по формуле

$$S^2(\vec{Y}_n) = \frac{1}{n-m}\big(Y - F\hat{\vec{\beta}}(\vec{Y}_n)\big)^{\mathrm{T}}\big(Y - F\hat{\vec{\beta}}(\vec{Y}_n)\big), \tag{7.17}$$

где $\hat{\vec{\beta}}(\vec{Y}_n)$ — МНК-оценка вектора параметров $\vec{\beta}$ линейной регрессионной модели (7.6).

◀ Из равенства (7.6) и предположений о случайной составляющей модели следует:

$$\mathrm{M}\big((Y-F\vec{\beta})^{\mathrm{T}}(Y-F\vec{\beta})\big)=\mathrm{M}(\vec{\varepsilon}^{\mathrm{T}}\vec{\varepsilon})=$$

$$=\mathrm{M}\Big(\sum_{i=1}^{n}\varepsilon_i^2\Big)=\sum_{i=1}^{n}\mathrm{D}\varepsilon_i=n\sigma^2. \quad (7.18)$$

Рассмотрим равенства

$$(Y-F\vec{\beta})^{\mathrm{T}}(Y-F\vec{\beta})=$$

$$=\Big(Y-F\hat{\vec{\beta}}(\vec{Y}_n)+F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})\Big)^{\mathrm{T}}\Big(Y-F\hat{\vec{\beta}}(\vec{Y}_n)+F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})\Big)=$$

$$=(Y-F\hat{\vec{\beta}}(\vec{Y}_n))^{\mathrm{T}}(Y-F\hat{\vec{\beta}}(\vec{Y}_n))+(F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta}))^{\mathrm{T}}F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})+$$

$$+(Y-F\hat{\vec{\beta}}(\vec{Y}_n))^{\mathrm{T}}F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})+(F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta}))^{\mathrm{T}}(Y-F\hat{\vec{\beta}}(\vec{Y}_n)).$$

Поскольку $\hat{\vec{\beta}}(\vec{Y}_n)$ — МНК-оценка вектора параметров $\vec{\beta}$, то, согласно (7.7), $\hat{\vec{\beta}}(\vec{Y}_n)=(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}Y$, и, как следствие, имеем

$$(F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta}))^{\mathrm{T}}(Y-F\hat{\vec{\beta}}(\vec{Y}_n))=$$

$$=(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})^{\mathrm{T}}F^{\mathrm{T}}(Y-F(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}Y)=$$

$$=(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})^{\mathrm{T}}(F^{\mathrm{T}}Y-F^{\mathrm{T}}F(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}Y)=0$$

и, кроме того,

$$(Y-F\hat{\vec{\beta}}(\vec{Y}_n))^{\mathrm{T}}F(\vec{\beta}-\hat{\vec{\beta}}(\vec{Y}_n))=\big((F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta}))^{\mathrm{T}}(Y-F\hat{\vec{\beta}}(\vec{Y}_n))\big)^{\mathrm{T}}=0.$$

Таким образом,

$$(Y-F\vec{\beta})^{\mathrm{T}}(Y-F\vec{\beta})-(F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta}))^{\mathrm{T}}F(\hat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})=$$

$$=(Y-F\hat{\vec{\beta}}(\vec{Y}_n))^{\mathrm{T}}(Y-F\hat{\vec{\beta}}(\vec{Y}_n)). \quad (7.19)$$

Воспользовавшись свойствами следа матриц и следствием 7.1, получим

$$\begin{aligned}
&\mathrm{M}\Big((F(\widehat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta}))^{\mathrm{T}}F(\widehat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})\Big)=\\
&\quad=\mathrm{M}\Big((\widehat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})^{\mathrm{T}}(F^{\mathrm{T}}F)(\widehat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})\Big)=\\
&\quad=\mathrm{M}\Big(\mathrm{tr}\big((F^{\mathrm{T}}F)(\widehat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})(\widehat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})^{\mathrm{T}}\big)\Big)=\\
&\quad=\mathrm{tr}\Big((F^{\mathrm{T}}F)\,\mathrm{M}\big((\widehat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})(\widehat{\vec{\beta}}(\vec{Y}_n)-\vec{\beta})^{\mathrm{T}}\big)\Big)=\\
&\quad=\mathrm{tr}\big((F^{\mathrm{T}}F)\Sigma(\widehat{\vec{\beta}}(\vec{Y}_n))\big)=\mathrm{tr}\big((F^{\mathrm{T}}F)(F^{\mathrm{T}}F)^{-1}\sigma^2\big)=\\
&\quad=\sigma^2\mathrm{tr}I_m=m\sigma^2.
\end{aligned}$$

Таким образом, согласно (7.17)–(7.19), имеем

$$n\sigma^2-m\sigma^2=\mathrm{M}\Big(Y-F\widehat{\vec{\beta}}(\vec{Y}_n))^{\mathrm{T}}(Y-F\widehat{\vec{\beta}}(\vec{Y}_n))\Big),$$

откуда

$$\sigma^2=\frac{1}{n-m}\mathrm{M}\Big((Y-F\widehat{\vec{\beta}}(\vec{Y}_n))^{\mathrm{T}}(Y-F\widehat{\vec{\beta}}(\vec{Y}_n)\Big).$$

Следовательно,

$$S^2=\frac{1}{n-m}(Y-F\widehat{\vec{\beta}}(\vec{Y}_n))^{\mathrm{T}}(Y-F\widehat{\vec{\beta}}(\vec{Y}_n))$$

является несмещенной оценкой для σ^2. ▶

Замечание 7.3. а. Оценка S^2 остаточной дисперсии σ^2 представляет собой отношение остаточной суммы квадратов $\widehat{\vec{\varepsilon}}^{\mathrm{T}}\widehat{\vec{\varepsilon}}$, отнесенное к числу степеней свободы $n-m$, где n — количество наблюдений $\vec{Y}_n=(Y_1,\ldots,Y_n)$, представленных матрицей отклика Y, а m — число оцениваемых параметров, представленных вектором $\vec{\beta}$. Таким образом, S^2 — доля остаточной суммы квадратов линейной регрессионной модели (7.6), приходящаяся

на одну „степень свободы". Фактически число степеней свободы равно объему случайной выборки за вычетом числа независимых линейных связей, наложенных на выборочные значения.

б. Формула (7.17) верна лишь в том случае, если есть основания считать, что выбранная линейная регрессионная модель (7.6) является верной, т.е. $\mathrm{M}Y = F\vec{\beta}$. В противном случае в остаточную сумму квадратов кроме случайных ошибок входят и систематические, а потому она может давать завышенную оценку для σ^2.

в. Полезно обратить внимание на сходство результата (7.17) с несмещенной оценкой $S^2(\vec{Y}_n)$ дисперсии случайной величины Y по наблюдениям $\vec{Y}_n$, которая имеет вид

$$S^2(\vec{Y}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Здесь также сумма квадратов отклонений Y_i от $\overline{Y}$ делится на число степеней свободы $n-1$, так как неизвестный параметр $\mathrm{M}Y = \mu$ заменен его оценкой $\overline{Y}$, т.е. на экспериментальные данные наложена одна линейная связь. #

При решении реальных задач, связанных с практическим использованием регрессионных моделей, необходимо проверять выполнение исходных предположений для метода наименьших квадратов, т.е. проводить статистический анализ регрессионной модели (см. **7.3**).

Проиллюстрируем процедуру построения регрессионной модели на частных примерах, имеющих и самостоятельный интерес.

Пример 7.3. Рассмотрим случай *простой линейной регрессии*, когда отклик Y зависит от одного фактора X (т.е. $p=1$) и в качестве приближения искомой функции регрессии выбрана функция

$$f_a(x) = \beta_0 + \beta_1 x.$$

Эту функцию получают из общей модели (7.2) при $\psi_0(x) \equiv 1$, $\psi_1(x) = x$, т.е. размерность вектора X равна $p = 1$, а число параметров $m = 2$.

Роль фактора X могут играть время (иногда часто вместо x пишут t), температура, доза лечебного препарата и т.д. Задача состоит в изучении связи между откликом Y и фактором X на основании выборки (x^i, y_i), $i = \overline{1, n}$, полученной в результате эксперимента (вместо x^i далее будем писать x_i, так как x — скаляр).

Для конкретности будем считать, что X — это скорость движения автомобиля (в км/ч), а Y — тормозной путь (в м) по скользкой дороге до полной его остановки, и по результатам $n = 17$ замеров X и Y получены данные, представленные в табл. 7.1.

Таблица 7.1

x_i	28	29	32	35	40	44	45	51	53
y_i	0,53	0,92	1,52	2,07	2,17	3,65	3,97	5,27	5,54
x_i	58	64	65	73	75	80	83	93	
y_i	6,43	7,60	7,91	9,48	10,1	8,95	11,48	13,74	

Найдем значения оценок параметров β_0 и β_1, дисперсии отклика и дисперсии среднего значения отклика, а также дадим прогноз для длины тормозного пути при скорости $x_0 = 120\,км/ч$. В рассматриваемом примере матрицы Y и F имеют вид

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad F = \begin{pmatrix} 1 & x_1 \\ \dots & \\ 1 & x_n \end{pmatrix}.$$

Следовательно,

$$F^{\mathrm{T}}F = \begin{pmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix}, \qquad F^{\mathrm{T}}Y = \begin{pmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{pmatrix}.$$

Далее случайные векторы и их реализации будем обозначать одинаково: из текста всегда ясно, о чем идет речь.

Поскольку x_i — различные числа, то матрица $M = F^{\mathrm{T}}F$ обратима, причем

$$\det M = n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2 = n\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

С помощью присоединенной матрицы находим

$$M^{-1} = \frac{1}{\det M}\begin{pmatrix} \sum\limits_{i=1}^{n} x_i^2 & -\sum\limits_{i=1}^{n} x_i \\ -\sum\limits_{i=1}^{n} x_i & n \end{pmatrix}.$$

Используя обратную матрицу $M^{-1} = (F^{\mathrm{T}}F)^{-1}$, по формуле (7.7) получаем

$$\widehat{\vec{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = M^{-1}F^{\mathrm{T}}Y = \frac{1}{\det M}\begin{pmatrix} \sum\limits_{i=1}^{n} x_i^2\Big(\sum\limits_{i=1}^{n} y_i\Big) - \sum\limits_{i=1}^{n} x_i\Big(\sum\limits_{i=1}^{n} x_i y_i\Big) \\ -\sum\limits_{i=1}^{n} x_i\Big(\sum\limits_{i=1}^{n} y_i\Big) + n\sum\limits_{i=1}^{n} x_i y_i \end{pmatrix},$$

откуда

$$\widehat{\beta}_0 = \frac{\sum\limits_{i=1}^{n} x_i^2 \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i\Big(\sum\limits_{i=1}^{n} x_i y_i\Big)}{n\sum\limits_{i=1}^{n} x_i^2 - \Big(\sum\limits_{i=1}^{n} x_i\Big)^2}, \qquad \widehat{\beta}_1 = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i\Big(\sum\limits_{i=1}^{n} y_i\Big)}{n\sum\limits_{i=1}^{n} x_i^2 - \Big(\sum\limits_{i=1}^{n} x_i\Big)^2}.$$

Из последних двух равенств с помощью простых преобразований получаем

$$\begin{cases} \widehat{\beta}_1 = \dfrac{Q_{xy}}{Q_x}, \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}, \\ Q_x = \sum\limits_{i=1}^{n}(x_i - \overline{x})^2, \\ Q_{xy} = \sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}). \end{cases} \tag{7.20}$$

Из равенств (7.20) видно, что оценки $\widehat{\beta}_0(\vec{Y}_n)$ и $\widehat{\beta}_1(\vec{Y}_n)$ связаны линейной зависимостью.

Поскольку $Q_{xy}/n = \widehat{K}_{xy}$ является значением оценки ковариации $Q_{xy}(\vec{X}_n, \vec{Y}_n)/n = \widehat{K}_{xy}(\vec{X}_n, \vec{Y}_n)$ фактора и отклика, а $Q_x/n = \widehat{\sigma}_x^2$ — значением оценки дисперсии фактора, то для значения $\widehat{\beta}_1$ оценки параметра β_1 справедливо и такое представление:

$$\widehat{\beta}_1 = \frac{\widehat{K}_{xy}}{\widehat{\sigma}_x^2} = \frac{\widehat{\rho}\widehat{\sigma}_x\widehat{\sigma}_y}{\widehat{\sigma}_x^2},$$

где $\widehat{\rho}$ — значение оценки коэффициента корреляции ρ между фактором и откликом.

Таким образом, найдена модель простой регрессии

$$\widehat{y}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x,$$

где $\widehat{\beta}_0$ и $\widehat{\beta}_1$ определены по формулам (7.20).

Для данных из табл. 7.1 имеем $n = 17$. Далее находим $Q_x \approx 6557$, $Q_y \approx 250{,}8$, $Q_{xy} \approx 1271{,}5$.

По формулам (7.20) вычисляем значения оценок $\widehat{\beta}_0$ и $\widehat{\beta}_1$:

$$\widehat{\beta}_1 = 1271{,}5/6557 \approx 0{,}194, \qquad \widehat{\beta}_0 = 5{,}95 - 0{,}194 \cdot 55{,}76 \approx -4{,}87.$$

Следовательно, прогнозируемое значение y при $x = x_0 = 120$ равно

$$\widehat{y}(120) = -4{,}87 + 0{,}194 \cdot 120 = 18{,}4.$$

Найдем точность оценок $\widehat{\beta}_0(\vec{Y}_n)$, $\widehat{\beta}_1(\vec{Y}_n)$ и $\widehat{Y}(x)$. Используя формулу (7.14), получаем

$$\Sigma\left(\widehat{\vec{\beta}}\right) = \sigma^2 M^{-1} = \frac{\sigma^2}{\det M}\begin{pmatrix} \sum\limits_{i=1}^{n} x_i^2 & -\sum\limits_{i=1}^{n} x_i \\ -\sum\limits_{i=1}^{n} x_i & n \end{pmatrix},$$

где

$$\det M = n\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Таким образом,

$$\mathbf{D}\widehat{\beta}_0(\vec{Y}_n)=\frac{\sigma^2\sum\limits_{i=1}^{n}x_i^2}{n\sum\limits_{i=1}^{n}(x_i-\overline{x})^2},\qquad \mathbf{D}\widehat{\beta}_1(\vec{Y}_n)=\frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2},$$

$$\widehat{K}(\vec{X}_n,\vec{Y}_n)=\mathbf{cov}(\widehat{\beta}_0(\vec{Y}_n),\widehat{\beta}_1(\vec{Y}_n))=-\frac{\sigma^2\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}.$$

По формуле (7.15) находим $\mathbf{D}\widehat{Y}(x)$:

$$\mathbf{D}\widehat{Y}(x)=\frac{\sigma^2}{n\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}(1\ \ x)\begin{pmatrix}\sum\limits_{i=1}^{n}x_i^2 & -\sum\limits_{i=1}^{n}x_i\\ -\sum\limits_{i=1}^{n}x_i & n\end{pmatrix}\begin{pmatrix}1\\ x\end{pmatrix}=$$

$$=\sigma^2\left(\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right).\quad(7.21)$$

В точке прогноза $x=x_0=120$ и

$$\mathbf{D}\widehat{Y}(x_0)=\sigma^2\left(\frac{1}{n}+\frac{(120-55{,}76)^2}{6557}\right)=0{,}69\sigma^2.$$

Наконец, по формуле (7.17) находим значение S_y^2 оценки дисперсии отклика:

$$S_y^2=\frac{1}{n-2}\sum_{i=1}^{n}\bigl(y_i-\widehat{y}(x_i)\bigr)^2\approx\frac{1}{15}\cdot 4{,}2=0{,}28$$

и заменяем σ^2 на S_y^2 во всех предыдущих равенствах, где присутствует σ^2.

Считая оценку $\widehat{Y}(x)$ нормально распределенной с математическим ожиданием $\mathbf{M}\widehat{Y}(x)=f(x)$ и дисперсией $\mathbf{D}\widehat{Y}(x)$, вычисленной по формуле (7.21), можно по *правилу* „3σ“ указать

интервал возможных значений для Y в точке $x = x_0 = 120$, учитывая, что $\sigma \approx S_y = 0{,}53$:

$$\left(\widehat{y}(x_0) - 3S_y,\ \widehat{y}(x_0) + 3S_y\right) = (18{,}4 - 1{,}6;\ 18{,}4 + 1{,}6).$$

Доверительный интервал для заданной доверительной вероятности γ будет построен в 7.3.

Пример 7.4 (квадратичная регрессия). Исследуется эффективность системы охлаждения двигателя, работающего непрерывно в течение времени $t_0 = 60$ мин. Измерение температуры T (в °C) работающего двигателя проведено с интервалом 5 мин в течение 25 мин. Результаты сведены в таблицу (табл. 7.2). На рис. 7.5 дано графическое представление этих данных. Считая, что зависимость между переменными t (фактор) и T (отклик) является квадратичной, т.е. $T = a + bt + ct^2$, найдем по методу наименьших квадратов значения оценок параметров a, b и c.

Таблица 7.2

t	5	10	15	20	25
T	59,3	59,8	60,1	64,9	70,2

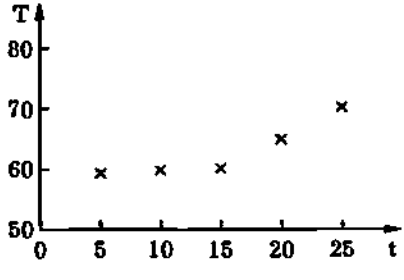


Рис. 7.5

Для удобства вычислений предварительно сделаем линейные преобразования переменных t и T по формулам

$$x = \frac{t - 15}{5}, \qquad y = 10(T - 60)$$

и вычислим вначале значения МНК-оценки параметров линейной (по параметрам) модели

$$y = \beta_0 + \beta_1 x + \beta_2 x^2.$$

В данном случае базисные функции такие: $\psi_0(x) \equiv 1$, $\psi_1(x) = x$, $\psi_2(x) = x^2$. Будем искать МНК-оценки не по формуле (7.7), а непосредственно решая систему уравнений (7.8), которая в данном случае имеет вид

$$\begin{cases} 5\beta_0 + 10\beta_2 = 143, \\ 10\beta_1 = 269, \\ 10\beta_0 + 34\beta_2 = 427. \end{cases}$$

Решение системы таково: $\beta_0 = 8{,}4$, $\beta_1 = 26{,}9$, $\beta_2 = 10{,}1$. Таким образом, $y = 8{,}4 + 26{,}9x + 10{,}1x^2$ и, переходя к исходным переменным t и T, окончательно получаем

$$\widehat{f}_a(t) = 61{,}86 - 0{,}67t + 0{,}04t^2.$$

7.3. Статистический анализ регрессионной модели

Статистический анализ *модели регрессии* (7.9), построенной на основе параметризации искомой функции регрессии $f(x)$ в виде (7.2) и на основе *МНК-оценок* параметров, состоит из следующих трех этапов:

– проверка адекватности модели регрессии;

– проверка значимости модели регрессии и ее параметров;

– анализ точности результатов, полученных с использованием регрессионной модели.

Для проведения статистического анализа требуется дополнить исходные предположения *метода наименьших квадратов* еще одним. Будем считать, что *случайные ошибки* ε_i, $i = \overline{1,n}$, в модели (7.3) не только независимы, но и распределены по нормальному закону: $\varepsilon_i \sim N(0, \sigma^2)$, $i = \overline{1,n}$, т.е. случайная составляющая $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ *линейной регрессионной модели* (7.6) имеет n-мерный нормальный закон распределения с нулевым средним значением и ковариационной матрицей $\sigma^2 I_n$.

Это предположение в силу (7.3) эквивалентно тому, что наблюдения Y_i, $i=\overline{1,n}$, являются независимыми нормально распределенными случайными величинами, т.е.

$$Y_i \sim N\left(f_a(\vec{x}^i), \sigma^2\right), \tag{7.22}$$

где

$$f_a(\vec{x}^i) = \sum_{k=0}^{m-1} \beta_k \psi_k(\vec{x}^i), \quad i=\overline{1,n}.$$

Проверку рассматриваемого предположения проводят на основе статистического анализа случайных величин

$$\varepsilon_i = Y_i - \widehat{Y}(\vec{x}^i), \quad i=\overline{1,n},$$

значения которых представляют собой отклонения наблюдаемых значений y_i отклика Y от его значений, предсказанных моделью регрессии

$$\widehat{y}(\vec{x}^i) = \sum_{k=0}^{m-1} \widehat{\beta}_i \psi_k(\vec{x}^i).$$

Таким образом, все сводится к проверке *статистической гипотезы* о выполнении исходных предположений: случайные величины ε_i, $i=\overline{1,n}$, являются независимыми и $\varepsilon_i \sim N(0,\sigma^2)$, $i=\overline{1,n}$. Критерии проверки указанных гипотез рассмотрены выше (см. 5).

Следует отметить, что, когда каждая случайная величина ε_i имеет единственную *реализацию* (нет повторных наблюдений), мы не можем проверить гипотезу о независимости случайных величин ε_i, $i=\overline{1,n}$. Однако, если у исследователя есть основания считать, что случайные величины ε_i, $i=\overline{1,n}$, независимы и одинаково распределены, можно ограничиться проверкой гипотезы о том, что $\widehat{\varepsilon}_i$, $i=\overline{1,n}$ — реализация случайной величины ε_i распределенной по нормальному закону.

Считая, что исходные предположения метода наименьших квадратов выполнены, перейдем к рассмотрению этапов статистического анализа регрессионной модели.

Проверка адекватности построенной модели регрессии. ***Линейную регрессионную модель*** называют ***адекватной***, если предсказанные по ней значения отклика Y согласуются с результатами наблюдений.

В основе процедуры проверки адекватности модели лежат предположения, что случайные ошибки наблюдений ε_i, $i = \overline{1, n}$, являются независимыми, нормально распределенными случайными величинами с нулевыми средними значениями и одинаковыми дисперсиями σ^2.

Пусть для каждого или некоторых значений переменного $x = (x_1, \ldots, x_p)$ имеется несколько (r_i, $i = \overline{1, n}$) повторных наблюдений отклика Y (т.е. исходные данные представлены матрицей D — см. (7.1)). Тогда для проверки адекватности модели можно использовать следующую процедуру.

Итак, повторные наблюдения получены при различных значениях $\vec{x}^1, \ldots, \vec{x}^n$ переменного x, причем в точке $\vec{x} = \vec{x}^i$ произведено r_i наблюдений $y_{i1}, \ldots, y_{ir_i}$ отклика Y, а $\sum\limits_{i=1}^{n} r_i = N$ — *объем выборки*. Введем обозначение

$$\overline{y}_i = \frac{1}{r_i} \sum_{j=1}^{r_i} y_{ij}.$$

Если линейная регрессионная модель адекватна, то значения $\overline{y}_i$ должны быть близки к значениям $\widehat{y}_i = \widehat{y}(\vec{x}^i)$, $i = \overline{1, n}$. Следовательно, сумму квадратов

$$Q_n = \sum_{i=1}^{n} r_i \left(\overline{y}_i - \widehat{y}(\vec{x}^i)\right)^2$$

можно рассматривать как меру неадекватности рассматриваемой модели.

Можно показать, что *статистики*

$$Q_n(\vec{Y}_N) = \sum_{i=1}^{n} r_i \big(\overline{Y}_i - \widehat{Y}(\vec{x}^{\,i})\big)^2,$$

$$Q_p(\vec{Y}_N) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \big(Y_{ij} - \overline{Y}(\vec{x}^{\,i})\big)^2$$

являются независимыми случайными величинами. Статистика $Q_p(\vec{Y}_N)/\sigma^2$ имеет *χ^2-распределение* с числом степеней свободы $\sum\limits_{i=1}^{n}(r_i - 1)$, а отношение

$$S_y^2(\vec{Y}_N) = \frac{Q_p(\vec{Y}_N)}{\sum\limits_{i=1}^{n}(r_i - 1)}$$

является несмещенной оценкой *остаточной дисперсии*. Эта статистика не связана с ошибкой в выборе модели. Статистика $Q_n(\vec{Y}_N)/\sigma^2$ имеет распределение χ^2 с числом степеней свободы $n - m$, если гипотеза H_0: $\mathbf{M}Y = F\vec{\beta}$ верна (здесь m — число неизвестных параметров в модели (7.2)). При этом $S^2_{\text{ад}} = Q_n(\vec{Y}_N)/(n-m)$ — несмещенная оценка σ^2.

Следовательно (см. **Д.3.1**), статистика имеет *распределение Фишера* со степенями свободы $n - m$ и $\sum\limits_{i=1}^{n}(r_i - 1)$:

$$F = \frac{S^2_{\text{ад}}(\vec{Y}_N)}{S_y^2(\vec{Y}_N)} = \frac{Q_n(\vec{Y}_N)}{n-m} \frac{\sum\limits_{i=1}^{n}(r_i - 1)}{Q_p(\vec{Y}_N)} \sim F\Big(n-m, \sum_{i=1}^{n}(r_i - 1)\Big).$$

Поэтому проверка гипотезы H_0 осуществляется стандартным образом по *критерию Фишера*.

Если *выборочное значение* $f_{\text{в}}$ статистики F не превышает критического $f_{\text{кр}}$, т.е.

$$f_{\text{в}} \leqslant f_{\text{кр}} = f_{1-\alpha}(r_n, r_p),$$

то гипотезу H_0 принимают (точнее, не отклоняют) на *уровне значимости* α, т.е. модель признается адекватной.

В противном случае модель признается неадекватной и нужно пытаться построить более сложную модель, увеличив, например, число *базисных функций* или выбрав другие базисные функции.

Пример 7.5. Найдем МНК-оценки параметров *простой линейной регрессии*

$$f_a(x) = \beta_0 + \beta_1 x$$

по данным табл. 7.3 и проверим адекватность модели регрессии на уровне значимости $\alpha = 0{,}05$.

Таблица 7.3

x_i	1	2	3	2,7	4,3	5,0
y_{ij}	0,5; 0,1	0,5; 1,2	1,2; 1,7	0,9; 2,2	1,1; 1,7; 2,5	2,0; 2,2
r_i	2	2	2	2	3	2

Имеем $\sum\limits_{i=1}^{n} r_i = N = 13$, $n = 6$, $m = 2$,

$$Q_p = \sum_{i=1}^{6}\sum_{j=1}^{r_i}(y_{ij} - \overline{y}_i) = 2{,}29.$$

По формулам (7.20) находим

$$\widehat{\beta}_1 = \frac{9{,}68}{23{,}12} = 0{,}419, \qquad \widehat{\beta}_0 = \frac{17{,}8 - 0{,}419 \cdot 40{,}3}{13} = 0{,}07.$$

Итак, $\widehat{y}(x) = 0{,}07 + 0{,}419x$. Далее вычисляем

$$Q_n = \sum_{i=1}^{6} r_i\left(\overline{y}_i - \widehat{y}(x_i)\right)^2 \approx 0{,}39$$

и рассчитываем выборочное значение

$$f_в = \frac{0{,}39/(6-2)}{2{,}29/(13-6)} \approx 0{,}3$$

статистики

$$F = \frac{Q_n(\vec{Y}_N)}{n-m} \frac{\sum_{i=1}^{n}(r_i - 1)}{Q_p(\vec{Y}_N)}.$$

Поскольку критическое значение $f_{\text{кр}} = f_{0,95}(4,7) = 4{,}14$ (см. табл. П.5) существенно больше $f_{\text{в}}$, то построенную модель регрессии можно считать адекватной результатам наблюдений.

Проверка значимости параметров модели регрессии. Напомним, что регрессионную модель мы выбрали в виде (7.3), т.е. неизвестную функцию регрессии $f(x)$ ищем в виде

$$f_a(x) = \sum_{k=0}^{m-1} \beta_k \psi_k(x), \tag{7.23}$$

где некоторые из базисных функций $\psi_k(x)$ могли быть включены в модель регрессии ошибочно, т.е. на самом деле *отклик* Y от этих $\psi_k(x)$ не зависит и потому соответствующие коэффициенты β_k должны быть равны нулю. Однако может оказаться, что полученные по формуле (7.7) значения МНК-оценок $\widehat{\beta}_k$ отличны от нуля, хотя обычно к нулю и близки.

Проверка значимости коэффициента β_k означает проверку гипотезы H_0: $\beta_k = 0$ против *альтернативной статистической гипотезы* H_1: $\beta_k \neq 0$. ***Коэффициент*** β_k считают ***значимым***, если верна гипотеза H_1.

В общем случае могут возникать более сложные гипотезы, например гипотеза H_0: $\beta_1 = -\beta_2 = \beta$, означающая, что $\beta_1 + \beta_2 = 0$. Такая гипотеза уместна, когда есть подозрение, что действует не каждый из *факторов* X_1 и X_2 по отдельности, а только их разность, т.е. вместо комбинации $\beta_1 X_1 + \beta_2 X_2$ в модель нужно включить выражение $\beta(X_1 - X_2)$.

Статистические гипотезы, которые включают утверждение о линейной комбинации параметров β_j, $j = \overline{0, m-1}$, называют ***линейными гипотезами***. Они обычно вытекают из знаний экспериментатора или его предположений относительно

возможных моделей. Под проверкой значимости параметров модели регрессии в этом случае понимают проверку всех возможных линейных гипотез.

Мы ограничимся здесь проверкой линейных гипотез двух типов:

1) гипотезы H_0: $\beta_0 = \beta_1 = \ldots = \beta_{m-1} = 0$ против альтернативной гипотезы H_1, согласно которой $\beta_k \neq 0$ хотя бы для одного номера k, $k = \overline{0, m-1}$;

2) гипотезы H_{0k}: $\beta_k = 0$ против альтернативной гипотезы H_{1k}: $\beta_k \neq 0$, рассматриваемых для некоторого фиксированного номера k, $k = \overline{0, m-1}$.

Если гипотеза H_0 верна, то ***модель регрессии*** называют ***незначимой***, т.е. условное математическое ожидание отклика $\mathrm{M}(Y)\,x = \overline{y}(x) = \beta_0$ постоянно и не меняется с изменением x. В противном случае ***модель регрессии*** называют ***значимой***.

Гипотезы второго типа связаны с анализом конкретного коэффициента β_k. Если гипотеза H_{0k} принимается, то коэффициент β_K незначим и может быть удален из модели.

Рассмотрим критерий проверки гипотез первого типа. Исходя из предположений о случайных величинах Y_i, $i = \overline{1, n}$, сделанных в начале параграфа, можно показать, что статистики $Q_l(\vec{Y}_n) = (Y - \widehat{Y})^{\mathrm{T}}(Y - \widehat{Y})$ (*остаточная сумма квадратов*) и $Q_f(\vec{Y}_n) = (\widehat{Y} - I\overline{Y})^{\mathrm{T}}(\widehat{Y} - I\overline{Y})$ являются независимыми случайными величинами. Здесь Y — *матрица отклика линейной регрессионной модели* (7.6), $\widehat{Y}$ — *матрица МНК-оценок средних значений отклика* и $\overline{Y}$ — *выборочное среднее* отклика.

Раскрывая матричное представление статистик $Q_l(\vec{Y}_n)$ и $Q_f(\vec{Y}_n)$, заключаем, что

$$Q_l(\vec{Y}_n) = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2, \qquad Q_f(\vec{Y}_n) = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2.$$

Статистика $Q_l(\vec{Y}_n)/\sigma^2$ имеет χ^2-распределение с числом степеней свободы $n - m$, а статистика $Q_f(\vec{Y}_n)/\sigma^2$ — χ^2-распределение

с числом степеней свободы $m-1$, если H_0 верна. Тогда статистика

$$F = \frac{Q_f(\vec{Y}_n)}{m-1} \frac{n-m}{Q_l(\vec{Y}_n)} \sim F(m-1, n-m),$$

т.е. имеет распределение Фишера со степенями свободы $m-1$ и $n-m$.

Статистика $Q_l(\vec{Y}_n)/(n-m)$ является несмещенной оценкой остаточной дисперсии (см. теорему 7.3), обусловленной как случайными ошибками измерений значений функции регрессии, так и неучтенными в регрессии факторами; статистика $Q_f(\vec{Y}_n)/(m-1)$ — несмещенная оценка дисперсии случайных ошибок при использовании функции регрессии (т.е. дисперсии случайных ошибок измерений значений функции регрессии). Поэтому статистика F может быть использована при проверке рассматриваемой гипотезы.

Таким образом, гипотеза H_0: $\beta_1 = \ldots = \beta_{m-1} = 0$ отклоняется на уровне значимости α (а следовательно, регрессия признается значимой), если вычисленное значение статистики F

$$f_в > f_{кр} = f_{1-\alpha}(m-1, n-m). \tag{7.24}$$

Замечание 7.4. Полезной характеристикой линейной регрессионной модели является *коэффициент детерминации* R^2 (или квадрат *множественного коэффициента корреляции*).

Оценка

$$\widehat{R}^2 = 1 - \frac{Q_l}{Q_y} = \frac{Q_f}{Q_y}$$

коэффициента детерминации показывает, какая доля в сумме квадратов отклонений отклика Y от его среднего значения, т.е. в $Q_Y(\vec{Y}_n) = (Y - I\overline{Y})^{\mathrm{T}}(Y - I\overline{Y})$, обусловлена регрессией (т.е. показывает, насколько значимы параметры модели регрессии). Величина $\widehat{R}(\vec{Y}_n)$ является оценкой коэффициента корреляции (мерой линейной связи) между случайными величинами Y и $\widehat{Y}(\vec{x})$. #

Перейдем к проверке линейных гипотез второго типа. Эти гипотезы проверяют после того, как обоснована значимость регрессии. Такая проверка позволяет более детально проанализировать структуру модели регрессии на уровне отдельных коэффициентов. Ясно, что возможна ситуация, когда вектор параметров $\vec{\beta}$ модели регрессии является значимым, в то время как отдельные коэффициенты модели незначимы (и, следовательно, их надо принять равными нулю).

Проверку любой из m гипотез H_{0k}, $0 \leqslant k \leqslant m-1$, против гипотезы H_{1k} проводят по *критерию Стьюдента*.

Напомним, что МНК-оценка $\widehat{\beta}_k(\vec{Y}_n)$ параметра β_k линейно зависит от матрицы отклика Y. Следовательно, в силу (7.22) эта оценка имеет нормальный закон распределения с математическим ожиданием β_k (ибо оценка $\widehat{\beta}_k(\vec{Y}_n)$ несмещенная) и дисперсией $\sigma_y^2 c_{kk}$ (см. следствие 7.1). Здесь c_{kk} — k-й диагональный элемент *дисперсионной матрицы Фишера* $C = (F^T F)^{-1}$. Поэтому

$$Z = \frac{\widehat{\beta}_k(\vec{Y}_n) - \beta_k}{\sigma\sqrt{c_{kk}}} \sim N(0,1).$$

В то же время

$$V = \frac{Q_l(\vec{Y}_n)}{\sigma} = \frac{(n-m)S_y^2(\vec{Y}_n)}{\sigma^2} \sim \chi^2(n-m).$$

Таким образом, если гипотеза H_{0k}: $\beta_k = 0$ верна, то

$$T_k = \frac{\widehat{\beta}_k(\vec{Y}_n)}{S_y\sqrt{c_{kk}}} \sim S(n-m), \quad k = \overline{0, m-1}. \tag{7.25}$$

Если модуль вычисленного значения t_k статистики T_k превысит критический уровень $t_k^{\text{кр}} = t_{1-\alpha/2}(n-m)$, то гипотезу H_{0k} следует отклонить на уровне значимости α и признать коэффициент β_k значимым.

Замечание 7.5. Проверку значимости коэффициента β_k модели регрессии (7.23) можно проводить также с помощью

доверительного интервала $J_\gamma(\beta_k) = (\underline{\beta}_k(\vec{Y}_n), \overline{\beta}_k(\vec{Y}_n))$, значения границ которого в силу (7.25) имеют вид (см. **3.3**)

$$\widehat{\beta}_k \pm t_{1-\alpha/2}(n-m)S_y\sqrt{c_{kk}}. \tag{7.26}$$

Гипотеза H_{0k}: $\beta_k = 0$ принимается, если интервал с границами (7.26) накрывает нуль, и отклоняется в противном случае.

Замечание 7.6. Для простой линейной регрессии

$$f_a(x) = \beta_0 + \beta_1 x$$

(см. пример 7.3) число параметров $m = 2$, а дисперсионная матрица Фишера имеет вид

$$\Sigma\widehat{\beta} = \begin{pmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{pmatrix},$$

где

$$c_{00} = \frac{\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum\limits_{i=1}^{n} x_i^2}{nQ_x}, \qquad c_{11} = \frac{1}{Q_x}.$$

Поэтому из (7.25) следует, что

$$T_0 = \frac{\widehat{\beta}_0(\vec{Y}_n)}{S_y}\sqrt{\frac{nQ_x}{\sum\limits_{i=1}^{n} x_i^2}} \sim S(n-2), \qquad T_1 = \frac{\widehat{\beta}_1(\vec{Y}_n)}{S_y}\sqrt{Q_x} \sim S(n-2),$$

а значения (7.26) границ доверительных интервалов для параметров β_0 и β_1 принимают соответственно вид

$$\widehat{\beta}_0 \pm t_{1-\alpha/2}(n-2)S_y\sqrt{\frac{\sum\limits_{i=1}^{n} x_i^2}{nQ_x}}, \qquad \widehat{\beta}_1 \pm t_{1-\alpha/2}(n-2)S_y\sqrt{\frac{1}{Q_x}}.$$

Пример 7.6. Результаты y_i, $i = \overline{1,n}$, наблюдений, проведенных над откликом Y при значениях x_i фактора X, представлены в табл. 7.4.

Таблица 7.4

x_i	0	1	2	3	4	5	6	7	8	9	10
y_i	8,98	8,82	9,09	11,94	24,63	14,06	14,00	24,93	33,22	15,7	35,92

Рассмотрим в качестве *допустимой модели регрессии* функцию

$$f_a(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

и найдем МНК-оценки неизвестных параметров модели регрессии: $\widehat{\beta}_0 = 6{,}92$; $\widehat{\beta}_1 = 2{,}27$; $\widehat{\beta}_2 = 0{,}08$. Таким образом, имеем

$$\widehat{y}(x) = 6{,}92 + 2{,}27x + 0{,}08x^2.$$

Есть основания предполагать, что $\beta_2 = 0$. Для проверки гипотезы H_0: $\beta_2 = 0$ (значимости коэффициента β_2) против альтернативной гипотезы H_1: $\beta_2 \neq 0$ находим значение $t_2 = 0{,}20$ статистики T_2 (7.25).

Воспользовавшись таблицей квантилей распределения Стьюдента (см. табл. П.4), на уровне значимости $\alpha = 0{,}1$ находим $t_{\text{кр}} = t_{1-\alpha/2}(n-m) = t_{0,95}(8) = 2{,}31$. Коэффициент β_2 незначим, так как $t_2 = 0{,}20 < t_{\text{кр}} = 2{,}31$.

Значение оценки коэффициента детерминации

$$\widehat{R}^2 = 1 - \frac{Q_l}{Q_y} = 1 - \frac{1060{,}51}{2214{,}24} \approx 0{,}52.$$

Полученный результат указывает на 52%-ный разброс результатов наблюдений относительно горизонтальной прямой $\overline{y} = 18{,}29$.

Анализ точности результатов, полученных с использованием регрессионной модели. Если модель регрессии прошла проверку на значимость, то ее можно использовать для

решения различных практических задач. Основными из них являются:

– определение значения отклика Y в той части факторного пространства, где эксперимент не проводился, т.е. либо интерполяция, либо экстраполяция (прогнозирование) отклика;

– определение экстремальных условий протекания процесса, модель которого построена, т.е. отыскание такой точки $x^* = (x_1^*, \ldots, x_n^*)$, в которой $\widehat{y}(x)$ имеет экстремум; эту задачу решают методами математического анализа [V].

В обоих случаях с помощью построенной модели

$$\widehat{Y}(x) = \sum_{k=0}^{m-1} \widehat{\beta}_k \psi_k(x)$$

требуется оценить точность предсказания в рассматриваемой точке $x = x_0$ либо среднего значения отклика $\mathrm{M}(Y \mid x) = \overline{y}(x)$, либо ожидаемого значения отклика $Y = Y_0$.

Для решения первой задачи нужно для величины $\overline{y}(x)$ построить доверительный интервал J_γ с заданным уровнем доверия γ, а для решения второй — так называемый прогнозирующий интервал $\widetilde{J}_\gamma$, в который случайная величина Y при $x = x^0$ попадает с заданной доверительной вероятностью γ.

При нахождении доверительного интервала J_γ важно то, что МНК-оценки $\widehat{\beta}_k(\vec{Y}_n)$ имеют нормальный закон распределения, а следовательно [XVI], оценка $\widehat{Y}(x)$ также распределена по нормальному закону со средним $\mathrm{M}\widehat{Y}(x) = \overline{y}(x)$ и дисперсией (см. (7.15))

$$\mathbf{D}\widehat{Y}(x) = \sigma^2 \psi^{\mathrm{T}}(x) C \psi(x).$$

Значит,

$$Z = \frac{\widehat{Y}(x) - \overline{y}(x)}{\sigma \sqrt{\psi^{\mathrm{T}}(x) C \psi(x)}} \sim N(0,1).$$

С другой стороны, несмещенная оценка дисперсии отклика σ^2, определяемая по формуле (7.17), не зависит от Z и

$$V = \frac{Q_l}{\sigma^2} = \frac{(n-m)S_y^2(\vec{Y}_n)}{\sigma^2} \sim \chi^2(n-m),$$

т.е. имеет χ^2-распределение с числом степеней свободы $n-m$.

Отсюда следует, что статистика $Z/\sqrt{V/(n-m)}$ распределена по закону Стьюдента с числом степеней свободы $n-m$ (см. **Д.3.1**):

$$\frac{Z}{\sqrt{V/(n-m)}} = \frac{\widehat{Y}(x) - \overline{y}(x)}{S_y(\vec{Y}_n)\sqrt{\psi^{\mathrm{T}}(x)C\psi(x)}} \sim S(n-m).$$

Таким образом, с вероятностью $\gamma = 1-\alpha$ выполняется неравенство

$$\left|\frac{\widehat{Y}(x) - \overline{y}(x)}{S_y(\vec{Y}_n)\sqrt{\psi^{\mathrm{T}}(x)C\psi(x)}}\right| < t_{1-\alpha/2}(n-m),$$

где $t_{1-\alpha/2}(n-m)$ — квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $n-m$.

Это равенство дает границы доверительного интервала с уровнем доверия γ для среднего значения отклика $\overline{y}(x)$ в произвольной точке x факторного пространства в виде

$$\widehat{y}(x) \pm t_{1-\alpha/2}(n-m)S_y(\vec{Y}_n)\sqrt{\psi^{\mathrm{T}}(x)C\psi(x)}, \qquad (7.27)$$

где, напомним, $C = (F^{\mathrm{T}}F)^{-1}$.

В частном случае простой линейной регрессии

$$\widehat{y}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

дисперсию $\widehat{Y}(x)$ вычисляют по формуле

$$\mathbf{D}\widehat{Y}(x) = \sigma^2\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right),$$

и формула (7.27) принимает следующий вид:

$$\widehat{y}(x) \pm t_{1-\alpha/2}(n-m) S_y(\vec{Y}_n) \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}. \qquad (7.28)$$

Из выражения (7.28) видно, что наиболее узким интервал J_γ будет в точке $x = \overline{x}$, и по мере удаления x от $\overline{x}$ точность уменьшается (рис. 7.6).

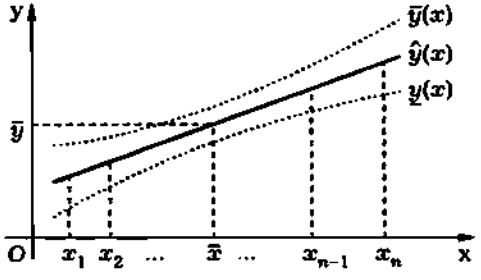


Рис. 7.6

Для отыскания прогнозирующего интервала $\widetilde{J}_\gamma$ с уровнем доверия γ используют тот факт, что разность между откликом Y и оценкой его среднего значения $\widehat{Y}(x)$ в любой точке x имеет нормальный закон распределения со средним значением $\mathbf{M}\big(Y - \widehat{Y}(x)\big) = 0$ и дисперсией (в силу независимости Y и $\widehat{Y}(x)$)

$$\mathbf{D}\big(Y - \widehat{Y}(x)\big) = \mathbf{D}Y + \mathbf{D}\widehat{Y}(x) = \sigma^2 + \mathbf{D}\widehat{Y}(x) = \sigma^2\big(1 + \psi^{\mathrm{T}}(x) C \psi(x)\big),$$

т.е. к дисперсии $\widehat{Y}(x)$ добавляется дисперсия отклика Y.

Повторяя предыдущие рассуждения при построении доверительного интервала, вместо (7.27) получаем окончательный результат в виде

$$\widehat{y}(x) \pm t_{1-\alpha/2}(n-m) S_y(\vec{Y}_n) \sqrt{1 + \psi^{\mathrm{T}}(x) C \psi(x)}. \qquad (7.29)$$

7.4. О выборе допустимой модели регрессии

Как уже отмечалось выше, при решении задач *регрессионного анализа* исследователь в первую очередь сталкивается с необходимостью выбора класса $\mathcal{F}$ *допустимых моделей регрессии.* Мы не останавливаемся на этой проблеме* и еще раз отметим, что при ее решении, как правило, исследователь исходит из преследуемых целей, собственного опыта, результатов предварительного анализа, имеющегося экспериментального материала и т.д.

Если класс $\mathcal{F}$ содержит, например, две допустимые модели регрессии, то возникает проблема выбора наилучшей (в каком-то смысле) *допустимой модели регрессии.* Обсуждение этой проблемы можно найти в специальной литературе**, а мы ограничимся рассмотрением *линейной регрессионной модели* (см. (7.6)). При этом будем предполагать, что выполнены основные допущения регрессионного анализа: независимость и нормальное распределение случайных величин ε_i, $i = \overline{1,n}$ (см. (7.4)).

Пусть имеем две допустимые модели регрессии

$$\sum_{k=0}^{m_1-1} \beta_k \psi_k(\vec{x}) \quad \text{и} \quad \sum_{k=0}^{m_2-1} \beta_k \psi_k(\vec{x}), \tag{7.30}$$

где $m_2 > m_1$ и объем выборки равен n. Проверим *гипотезу*

$$H_0\colon\ \beta_{m_1} = \beta_{m_1+1} = \ldots = \beta_{m_2-1} = 0$$

против *альтернативной гипотезы*

$$H_1\colon\ \sum_{k=m_1}^{m_2-1} \beta_k^2 \neq 0.$$

*См.: *Кашьяп Р.Л., Рао А.Р.*

**См. там же.

Для проверки гипотезы H_0 можно применить *статистику*

$$F = \frac{Q_{l1}(\vec{Y}_n) - Q_{l2}(\vec{Y}_n)}{Q_{l2}(\vec{Y}_n)} \frac{n - m_2}{m_2 - m_1}, \tag{7.31}$$

где $Q_{l1}(\vec{Y}_n)$ и $Q_{l2}(\vec{Y}_n)$ — *остаточные суммы квадратов* соответственно для первой и второй моделей (7.30). Статистика F имеет *распределение Фишера* с числом степеней свободы $m_2 - m_1$ и $n - m_1 - m_2$.

Гипотезу H_0 следует принять на *уровне значимости* α (принять модель $\sum\limits_{k=0}^{m_1-1} \beta_k \psi(\vec{x})$), если значение $f_в$ статистики F, рассчитанное по результатам наблюдений, не превышает $f_{кр} = f_{1-\alpha}(m_2 - m_1, n - m_1 - m_2)$.

Заметим, что при $\widehat{Q}_{l2} > \widehat{Q}_{l1}$ всегда следует выбирать модель $\sum\limits_{k=0}^{m_1-1} \beta_k \psi(\vec{x})$.

Рассмотренный критерий называют ***критерием отношения остаточных дисперсий***. Смысл его прозрачен: усложнение допустимой модели регрессии статистически оправдано, если это приводит к значимому (на уровне значимости α) уменьшению значения оценки *остаточной дисперсии*.

Пример 7.7. Вернемся к примеру 7.6. Результаты наблюдений дают основание утверждать, что допустимыми моделями регрессии являются

$$\sum_{k=0}^{1} \beta_k \psi_k(\vec{x}) \quad \text{и} \quad \sum_{k=0}^{2} \beta_k \psi_k(\vec{x}).$$

С помощью *метода наименьших квадратов* находим значения оценок для параметров β_k, $k = \overline{0,1}$, первой модели регрессии. Для второй модели оценки параметров найдены в примере 7.6. Имеем

$$\widehat{y}_1(x) = 6{,}92 + 2{,}27x \quad \text{и} \quad \widehat{y}_2(x) = 6{,}92 + 2{,}27x + 0{,}08x^2.$$

Коэффициент β_2 во второй модели незначим (см. пример 7.6).

Применяя статистику (7.31), проверим гипотезу H_0: $\beta_2 = 0$ против альтернативной гипотезы H_1: $\beta_2 \neq 0$.

В нашем случае $n = 11$, $m_1 = 2{,}28$, $m_2 = 0{,}08$. Рассчитываем остаточные суммы квадратов $Q_{l1} = 393{,}84$ и $Q_{l2} = 455{,}21$. Значения оценок *остаточных дисперсий* соответственно равны 43,76 и 56,90. Поскольку $56{,}90 > 43{,}76$, то следует выбрать модель $\widehat{y}_1(x) = 6{,}91 + 2{,}28x$.

7.5. Решение типовых примеров

Пример 7.8. По заданной *выборке* (табл. 7.5) найдем *оценки* параметров *простой линейной регрессии* y на x: $y = \beta_0 + \beta_1 x$.

Таблица 7.5

x_i	2,7	4,6	6,3	7,8	9,2	10,6	12,0	13,4	14,7
y_i	17,0	16,2	13,3	13,0	9,7	9,9	6,2	5,8	5,7

В данном случае $\psi_0(x) = 1$, $\psi_1(x) = x$, матрицы F и Y имеют вид

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2{,}7 & 4{,}6 & 6{,}3 & 7{,}8 & 9{,}2 & 10{,}6 & 12{,}0 & 13{,}4 & 14{,}7 \end{pmatrix}^{\mathrm{T}},$$

$$Y = (17{,}0\ \ 16{,}2\ \ 13{,}3\ \ 13{,}0\ \ 9{,}7\ \ 9{,}9\ \ 6{,}2\ \ 5{,}8\ \ 5{,}7)^{\mathrm{T}}.$$

Находим матрицы

$$M = F^{\mathrm{T}}F = \begin{pmatrix} 9 & 81{,}4 \\ 81{,}4 & 865{,}63 \end{pmatrix}, \quad M^{-1} = \begin{pmatrix} 0{,}74322 & -0{,}06989 \\ -0{,}06989 & 0{,}00773 \end{pmatrix}.$$

В результате получаем

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = M^{-1}F^{\mathrm{T}}Y \approx \begin{pmatrix} 20{,}53 \\ -1{,}08 \end{pmatrix}.$$

Следовательно, $\widehat{y}(x) = 20{,}53 - 1{,}08x$.

Пример 7.9. Функциональная зависимость удельного сопротивления ρ кристаллического кварца от его температуры T имеет вид $\rho = 10^{a/T+b}$. Используя опытные данные (табл. 7.6), найдем оценки параметров a и b.

Таблица 7.6

T	335	365	400	445	500	570	670
ρ	$5\cdot10^{16}$	$4\cdot10^{15}$	$3\cdot10^{14}$	$2\cdot10^{13}$	$2\cdot10^{12}$	$1{,}5\cdot10^{11}$	10^{10}

Для решения задачи нелинейную *модель* преобразуем в *линейную по параметрам.* Для этого прологарифмируем левую и правую части: $\lg\rho = a/T + b$. Обозначим $x = 1000/T$ и $y = \lg\rho$. В результате приходим к задаче нахождения параметров простой линейной регрессии $y = ax + b$. Пересчитаем опытные данные в переменных x и y (табл. 7.7).

Таблица 7.7

x	2,985	2,740	2,500	2,247	2,000	1,754	1,493
y	16,699	15,602	14,477	13,301	12,301	11,176	10,000

Составляем матрицы

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2{,}985 & 2{,}740 & 2{,}500 & 2{,}247 & 2{,}000 & 1{,}754 & 1{,}493 \end{pmatrix}^{\mathrm{T}}$$

и

$$Y = (16{,}699\ \ 15{,}602\ \ 14{,}477\ \ 13{,}301\ \ 12{,}301\ \ 11{,}176\ \ 10{,}000)^{\mathrm{T}}.$$

Далее вычисляем матрицы

$$M = F^{\mathrm{T}}F = \begin{pmatrix} 7 & 15{,}719 \\ 15{,}719 & 37{,}022 \end{pmatrix}, \quad M^{-1} = \begin{pmatrix} 3{,}067 & -1{,}302 \\ -1{,}302 & 0{,}580 \end{pmatrix}.$$

Наконец, находим вектор-столбец параметров

$$\begin{pmatrix} a \\ b \end{pmatrix} = M^{-1}F^{\mathrm{T}}Y^{\mathrm{T}} = \begin{pmatrix} 3{,}306 \\ 4{,}480 \end{pmatrix}.$$

Итак, регрессионная модель в переменных x и y имеет вид $\widehat{y} = 3{,}306x + 4{,}480$. Следовательно, $\widehat{\rho} = 10^{4480/T+3{,}306}$.

Пример 7.10. В условиях примера 7.8 проверим значимость *коэффициента регрессии* y на x на *уровне значимости* $\alpha = 0{,}1$ и найдем значение *оценки коэффициента корреляции* ρ_{xy}.

Проверка значимости коэффициента регрессии в данном случае означает проверку *гипотезы* H_0: $\beta_1 = 0$ против *альтернативной гипотезы* H_1: $\beta_1 \neq 0$. Воспользуемся *статистикой*

$$F = \frac{Q_f(\vec{Y}_n)}{m-1} \frac{n-m}{Q_l(\vec{Y}_n)}.$$

Значения показателей Q_f и Q_l найдем по результатам наблюдений. Соответствующие вычисления сведем в таблицу (табл. 7.8), в которой $\widehat{y}_i = \widehat{y}(x_i)$ и $\overline{y}$ — среднее выборочное, равное

$$\overline{y} = \frac{1}{9}\sum_{i=1}^{9} y_i = 10{,}756.$$

Таблица 7.8

x_i	y_i	$\widehat{y}_i$	$y_i - \widehat{y}_i$	$(y_i - \widehat{y}_i)^2$	$\widehat{y}_i - \overline{y}$	$(\widehat{y}_i - \overline{y})^2$
2,7	17,0	17,614	−0,614	0,376996	16,858	47,032164
4,6	16,2	15,562	0,638	0,405044	4,806	23,097639
6,3	13,3	13,726	−0,426	0,181476	2,970	8,820900
7,8	13,0	12,106	0,894	0,799236	1,350	1,822500
9,2	9,7	10,594	−0,844	0,712336	−0,162	0,026244
10,6	9,9	9,082	0,818	0,669124	−1,674	2,802276
12,0	6,2	7,570	−1,370	1,876900	−3,186	10,150596
13,4	5,8	6,058	−0,258	0,066564	−4,698	22,071204
14,7	5,7	4,654	1,046	1,094116	−6,102	37,234404

Из табл. 7.8 получаем

$$Q_f = \sum_{i=1}^{9}(\widehat{y}_i - \overline{y})^2 = 153{,}0579, \qquad Q_l = \sum_{i=1}^{9}(y_i - \widehat{y})^2 = 6{,}1837$$

и вычисляем

$$f_в = \frac{153{,}0579 \cdot 7}{6{,}1837} \approx 172.$$

По таблице квантилей распределения Фишера с числом степеней свободы $m - 1 = 2 - 1 = 1$ и $n - m = 9 - 2 = 7$ (см. табл. П.5) находим $f_{кр} = f_{1-\alpha/2} = f_{0,95} = 5{,}59$. Из неравенства $f_в = 172 > f_{кр} = 5{,}59$ следует, что регрессия значима.

Чтобы найти коэффициент корреляции ρ_{xy}, воспользуемся равенством $\rho_{xy} = (\operatorname{sgn}\beta_1)\widehat{R}$, где $\widehat{R}^2$ — значение *оценки коэффициента детерминации*, равный

$$\widehat{R}^2 = \frac{Q_f}{Q_y} = \frac{153{,}0579}{153{,}0579 + 6{,}1837} = \frac{153{,}0579}{159{,}2416} \approx 0{,}9612.$$

В результате $\rho_{xy} \approx -0{,}96$.

Пример 7.11. Считая, что зависимость между x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$, найдем значения оценок параметров и проверим значимость *модели регрессии* на уровне $\alpha = 0{,}1$ по выборке, представленной в табл. 7.9.

Таблица 7.9

x_i	26	30	34	38	42	46	50
y_i	3,94	4,60	5,67	6,93	7,73	8,25	9,56

По данным выборки запишем матрицы

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 26 & 30 & 34 & 38 & 42 & 46 & 50 \\ 676 & 900 & 1156 & 1444 & 1764 & 2116 & 2500 \end{pmatrix}^{\mathrm{T}}$$

и

$$Y = (3{,}94 \;\; 4{,}60 \;\; 5{,}67 \;\; 6{,}93 \;\; 7{,}73 \;\; 8{,}25 \;\; 9{,}56)^{\mathrm{T}}.$$

Используя матрицу F, находим

$$M = F^{\mathrm{T}}F = \begin{pmatrix} 7 & 266 & 10556 \\ 266 & 10556 & 435176 \\ 10556 & 435176 & 18527600 \end{pmatrix},$$

$$M^{-1} = \begin{pmatrix} 91{,}926 & -4{,}962 & 0{,}064 \\ -4{,}962 & 0{,}271 & -3{,}534 \cdot 10^{-3} \\ 0{,}064 & -3{,}534 \cdot 10^{-3} & 4{,}65 \cdot 10^{-5} \end{pmatrix}.$$

Теперь вычисляем вектор-столбец параметров

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = M^{-1}F^{\mathrm{T}}Y == \begin{pmatrix} -2{,}6589 \\ 0{,}2579 \\ -0{,}0003 \end{pmatrix}.$$

Итак, $\widehat{y}(x) = -2{,}6589 + 0{,}2579x - 0{,}0003x^2$.

Проверим значимость модели регрессии на уровне $\alpha = 0{,}1$. Для этого составим таблицу (табл. 7.10), в которой $\widehat{y}_i = \widehat{y}(x_i)$ и $\overline{y}$ — выборочное среднее показателя y (среднее значение второго столбца табл. 7.9), равное

$$\overline{y} = \frac{1}{7}\Big(3{,}94 + 4{,}6 + 5{,}67 + 6{,}93 + 7{,}73 + 8{,}25 + 9{,}56\Big) = 6{,}67.$$

Таблица 7.10

x_i	y_i	$\widehat{y}_i$	$y_i - \widehat{y}_i$	$(y_i - \widehat{y}_i)^2$	$\widehat{y}_i - \overline{y}$	$(\widehat{y}_i - \overline{y})^2$
26	3,94	3,8343	0,11	0,0121	−2,84	8,0656
30	4,60	4,7958	−0,20	0,0400	−1,87	3,4969
34	5,67	5,7472	−0,08	0,0064	−0,92	0,8464
38	6,93	6,6886	0,24	0,0576	0,02	0,0004
42	7,73	7,6201	0,11	0,0121	0,95	0,9025
46	8,25	8,5415	−0,29	0,0841	1,87	3,4969
50	9,56	9,453	0,11	0,0121	2,78	7,7284

По данным составленной таблицы находим

$$Q_l = \sum_{i=1}^{7}(y_i - \widehat{y})^2 = 0{,}2244, \qquad Q_f = \sum_{i=1}^{7}(\widehat{y} - \overline{y})^2 = 24{,}5371,$$

$$f_{\text{в}} = \frac{24{,}5371 \cdot 4}{2 \cdot 0{,}2244} \approx 218{,}69.$$

По таблице квантилей распределения Фишера (см. табл. П.5) находим

$$f_{\text{кр}} = f_{1-\alpha/2}(m-1, n-m) = f_{0{,}95}(2, 4) = 6{,}94.$$

Из неравенства $f_{\text{в}} = 218{,}69 > f_{\text{кр}} = 6{,}94$, согласно критерию (7.24), приходим к заключению, что модель значима.

Пример 7.12. В условиях примера 7.11 проверим значимость коэффициента β_2, т.е. гипотезу H_{02}: $\beta_2 = 0$ при альтернативной гипотезе $\beta_2 \neq 0$ с уровнем значимости $\alpha = 0{,}1$.

Для решения поставленной задачи используем статистику

$$T = \frac{\widehat{\beta}_2(\vec{Y}_n)}{S_y(\vec{Y}_n)\sqrt{c_{22}}} \sim S(4)$$

(см. (7.25)). Ее выборочное значение равно

$$|t_{\text{в}}| = \frac{0{,}0003 \cdot 4}{0{,}2244\sqrt{\frac{1}{21504}}} \approx 0{,}784.$$

По таблице квантилей распределения Стьюдента (см. табл. П.4) находим

$$t_{\text{кр}} = t_{1-\alpha/2}(4) = t_{0{,}95}(4) = 2{,}132.$$

Так как $|t_{\text{в}}| < t_{\text{кр}}$, гипотезу H_{02} принимаем, т.е. коэффициент β_2 не является значимым.

Пример 7.13. По данным наблюдений (табл. 7.11) найдем оценки параметров модели регрессии $y = \beta_0 + \beta_1 x + \beta_2 x^2$ и проверим адекватность этой модели на уровне значимости $\alpha = 0{,}01$.

Таблица 7.11

x_i	0	0	0	1	2	2	3	3	4	4
y_i	22,8	21,9	22,1	24,5	26,0	26,1	26,8	27,3	28,2	28,5
x_i	5	6	6	6	7	8	8	9	10	
y_i	28,9	30,0	30,3	29,8	30,4	31,4	31,5	31,8	33,1	

По данным из табл. 7.11 запишем матрицы

$$F = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 6 & 36 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \\ 1 & 8 & 64 \\ 1 & 8 & 64 \\ 1 & 9 & 81 \\ 1 & 10 & 100 \end{pmatrix}, \quad Y = \begin{pmatrix} 22{,}8 \\ 21{,}9 \\ 22{,}1 \\ 24{,}5 \\ 26{,}0 \\ 26{,}1 \\ 26{,}8 \\ 27{,}3 \\ 28{,}2 \\ 28{,}5 \\ 28{,}9 \\ 30{,}0 \\ 30{,}3 \\ 29{,}8 \\ 30{,}4 \\ 31{,}4 \\ 31{,}5 \\ 31{,}8 \\ 33{,}1 \end{pmatrix}.$$

Найдя матрицы

$$M = F^{\mathrm{T}}F = \begin{pmatrix} 19 & 84 & 550 \\ 84 & 550 & 3018 \\ 550 & 3018 & 30274 \end{pmatrix},$$

$$M^{-1} = \begin{pmatrix} 0{,}253 & -0{,}097 & 7{,}903 \cdot 10^{-3} \\ -0{,}097 & 0{,}063 & -6{,}266 \cdot 10^{-3} \\ 7{,}903 \cdot 10^{-3} & -6{,}266 \cdot 10^{-3} & 6{,}84 \cdot 10^{-4} \end{pmatrix},$$

вычисляем вектор-столбец параметров

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = M^{-1}F^{\mathrm{T}}Y = \begin{pmatrix} 22{,}561 \\ 1{,}668 \\ -0{,}068 \end{pmatrix}.$$

Таким образом, $\widehat{y} = 22{,}561 + 1{,}668x - 0{,}068x^2$.

Для проверки адекватности найденной модели воспользуемся статистикой $F = S^2_{\text{ад}}(\vec{Y}_N)/S^2_y(\vec{Y}_N)$, которая имеет закон распределения Фишера с числом степеней свободы $r_n = n - m$ и $r_p = \sum\limits_{i=1}^{n}(r_i - 1)$. Для вычисления этой статистики сведем промежуточные данные в таблицу (табл. 7.12).

Таблица 7.12

x_i	r_i	$\overline{y}_i$	$\widehat{y}_i$	$\overline{y}_i - \widehat{y}_i$	$r_i(\overline{y}_i - \widehat{y}_i)^2$
0	3	22,267	22,561	−0,294	0,2593
1	1	24,500	24,161	0,339	0,1149
2	2	26,050	25,625	0,425	0,3612
3	2	27,050	26,953	0,097	0,0188
4	2	28,350	28,145	0,205	0,0840
5	1	28,900	29,201	−0,301	0,0906
6	3	30,033	30,121	−0,088	0,0232
7	1	30,400	30,905	−0,505	0,2550
8	2	31,450	31,553	−0,103	0,0212
9	1	31,800	32,012	−0,212	0,0449
10	1	33,100	32,441	0,659	0,4343

По данным таблицы находим

$$Q_n = \sum_{i=1}^{n} r_i \left(\overline{y}_i - \widehat{y}_i\right)^2 \approx 1{,}708$$

и

$$Q_p = \sum_{i=1}^{n}\sum_{j=1}^{n} \left(y_{ij} - \overline{y}\right)^2 = 0{,}753.$$

Теперь можно определить выборочное значение статистики:

$$f_{\text{в}} = \frac{753 \cdot (11-3)}{8 \cdot 1{,}708} = \frac{0{,}753}{1{,}708} \approx 0{,}441.$$

По таблице квантилей распределения Фишера (см. табл. П.5) находим критическое значение $f_{\text{кр}} = f_{0{,}95}(8,8) = 7{,}50$. Поскольку $f_{\text{в}} = 0{,}441 < f_{\text{кр}} = 7{,}50$, то найденная модель регрессии адекватна результатам наблюдений.

Пример 7.14. В условиях примера 7.5 построим: а) *доверительный интервал* для среднего значения отклика в точке $x = 10$; б) *прогнозирующий доверительный интервал. Доверительную вероятность* выберем $\gamma = 0{,}99$.

а. Границы доверительного интервала для среднего значения отклика в соответствии с (7.27) равны

$$\widehat{y}(x) \mp t_{1-(1-\gamma)/2}(n-m) S_y \sqrt{\psi^{\mathrm{T}}(x) C \psi(x)}.$$

Находим значения $\widehat{y}(x)$ и $\psi(x)^{\mathrm{T}} C \psi(x)$ в точке $x = 10$:

$$\widehat{y}(10) = 20{,}53 - 1{,}08 \cdot 10 = 9{,}73,$$

$$\psi(10)^{\mathrm{T}} C \psi(10) = (1 \;\; 10) \begin{pmatrix} 0{,}74322 & -0{,}06989 \\ -0{,}06989 & 0{,}00773 \end{pmatrix} \begin{pmatrix} 1 \\ 10 \end{pmatrix} = 0{,}11842.$$

Далее определяем выборочное значение S_y^2:

$$S_y^2 = \frac{1}{n-m}\sum_{i=1}^{n}\bigl(y_i - \widehat{y}(x_i)\bigr)^2 =$$

$$= \frac{1}{9-2}\Bigl(0{,}376996 + 0{,}405044 + 0{,}181476 + 0{,}799236 + 0{,}712336 +$$

$$+ 0{,}669124 + 1{,}876900 + 0{,}066564 + 1{,}094116\Bigr) = 0{,}8831.$$

По таблице квантилей распределения Стьюдента (см. табл. П.4) находим квантиль $t_{0{,}995}(7) = 3{,}499$. В результате получаем доверительный интервал для среднего значения отклика в точке $x = 10$: (8,60, 10,86).

6. Границы прогнозирующего доверительного интервала равны

$$\widehat{y}(x) \mp t_{1-(1-\gamma)/2} S_y \sqrt{1 + \psi(x)^{\mathrm{T}} C \psi(x)}.$$

По таблице квантилей распределения Стьюдента определяем $t_{1-(1-\gamma)/2} = t_{0{,}995} = 2{,}576$. В результате для границ интервала получаем

$$9{,}73 \pm 2{,}576\sqrt{0{,}8831}\sqrt{1 + 0{,}11842}.$$

После упрощений окончательно находим $\widetilde{y}_{0{,}99}(y) = (7{,}17,\ 12{,}29)$.

Вопросы и задачи

7.1. Какую функцию называют функцией регрессии?

7.2. Какие переменные называют входными (факторами), выходными (откликами)?

7.3. Что называют планом эксперимента?

7.4. Какую регрессионную модель называют линейной?

7.5. Сформулируйте исходные предположения метода наименьших квадратов.

7.6. В чем состоит метод наименьших квадратов нахождения параметров линейной регрессионной модели? Запишите формулу для оценок неизвестных параметров.

7.7. Запишите дисперсионную матрицу Фишера. Какой смысл имеют ее элементы?

7.8. В чем состоит анализ регрессионной модели? При каких предположениях его проводят?

7.9. Какую статистику используют для проверки значимости модели регрессии?

7.10. Какую линейную регрессионную модель называют адекватной? Сформулируйте правило проверки адекватности модели.

7.11. Запишите формулу для вычисления несмещенной оценки дисперсии отклика в случае адекватной регрессионной модели.

7.12. По данным эксперимента (табл. 7.13) с помощью метода наименьших квадратов найдите значения оценок параметров модели $y = a + b\ln x$.

Таблица 7.13

x_i	2,4	2,7	3,0	3,3	3,6	4,9	4,2
y_i	5,36	5,45	5,52	5,53	5,57	5,63	5,54

Ответ: $\widehat{y} = 4{,}97 + 0{,}47\ln x$.

7.13. Считая, что переменные x и y связаны зависимостью $y = \beta_0 e^{\beta_1 x}$, по выборке (1, 10), (2, 5), (3, 3), (4, 1) найдите значения оценок параметров β_0 и β_1.

Указание: используйте результаты примера 7.9.

Ответ: $\widehat{y} = 22{,}64e^{-0{,}74x}$.

7.14. Результаты эксперимента представлены таблицей

Значение x	Значения y
1	1; 1; 2
2	1; 2; 2; 3; 3; 3; 4
3	3; 4; 4; 4; 5; 5
4	4; 5; 5; 6
5	5; 5; 6
6	5; 6

Полагая, что переменные y и x связаны линейной зависимостью, найдите значения оценок параметров.

Ответ: $\hat{y} = 0{,}932 + 0{,}906x$.

7.15. Для модели регрессии, построенной в примере 7.3, проверьте ее значимость на уровне значимости $\alpha = 0{,}05$ и значимость ее коэффициентов β_0 и β_1 на уровне значимости $\alpha = 0{,}1$.

Ответ: модель значима; оба коэффициента значимы.

7.16. Зависимость между переменными x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$. По данным выборки

(0,07, 1,34); (0,31, 1,08); (0,61, 0,94); (0,99, 1,06);
(1,29, 1,25); (1,78, 2,01); (2,09, 2,60)

выполните следующее:

а) найдите значения оценок параметров модели регрессии;

б) проверьте значимость модели регрессии на уровне значимости $\alpha = 0{,}05$.

Ответ: а) $\hat{y} = 1{,}40 - 1{,}22x + 0{,}87x^2$; б) модель значима.

7.17. Зависимость между переменными x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$. По данным выборки (табл. 7.14) выполните следующее:

а) найдите значения оценок параметров модели регрессии;

б) проверьте значимость модели регрессии на уровне значимости $\alpha = 0{,}01$.

Таблица 7.14

x_i	26	30	34	38	42	46	50
y_i	3,94	4,60	5,67	6,93	8,25	7,73	10,55

Ответ: а) $\widehat{y} = 0{,}175 + 0{,}085x + 0{,}002x^2$; б) модель не является значимой.

7.18. Проведены равноточные измерения некоторой величины y через равные интервалы аргумента x (табл. 7.15). Считая, что зависимость между x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$, выполните следующее:

а) найдите значения оценок параметров модели регрессии;

б) проверьте значимость модели на уровне значимости $\alpha = 0{,}01$;

в) проверьте значимость коэффициентов β_1 и β_2 на уровне значимости $\alpha = 0{,}01$.

Таблица 7.15

x_i	−3	−2	−1	0	1	2	3
y_i	−0,71	−0,01	0,51	0,82	0,88	0,81	0,49

Ответ: а) $\widehat{y} = 0{,}200x - 0{,}102x^2$; б) модель значима; в) коэффициенты β_1 и β_2 значимы.

7.19. В условиях задачи 7.14 проверьте адекватность простой линейной модели.

Ответ: модель адекватна.

7.20. В условиях задачи 7.19 постройте: а) доверительный интервал для среднего значения отклика в точке $x = 5$ с доверительной вероятностью $\gamma = 0{,}9$; б) прогнозирующий интервал с доверительной вероятностью $\gamma = 0{,}9$.

Ответ: (−0,814, −0,674); (−0,802, −0,686).

8. ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА

8.1. Исходные понятия

Объектами исследования ***дисперсионного анализа*** являются *стохастические связи* между *откликом* и *факторами*, когда последние носят не количественный, а качественный характер. Примерами таких факторов могут служить:

- способ крепления детали при ее обработке;
- режим функционирования прибора;
- уровень квалификации оператора;
- методика обучения (или лечения) и т.д.

Чтобы подчеркнуть качественный характер факторов, будем их обозначать через $A, B, C, \ldots$, а отклик при этом — через X. Каждый из факторов имеет несколько ***уровней***, или градаций. Так, например, если X — это степень износа покрышки на колесе автомобиля, а выбранные факторы A и B — это тип дороги и тип рисунка протектора, то различные уровни фактора A — различные типы дорог, различные уровни фактора B — различные рисунки протектора.

Пусть наблюдаемый объект обладает таким свойством, которое характеризуется переменным (откликом) X и подвержено влиянию некоторых учитываемых факторов A, B и других, не контролируемых в данном эксперименте факторов. Задача дисперсионного анализа состоит в том, чтобы по результатам наблюдений за этим объектом дать ответ на вопрос: следует ли считать действие факторов A и B существенным (значимым) на фоне остальных (неучтенных) факторов или нет?

Формулировка и проверка соответствующих *статистических гипотез* для ответа на этот вопрос и является содержанием дисперсионного анализа.

В зависимости от числа анализируемых факторов различают ***однофакторный***, ***двухфакторный*** и т.д. ***дисперсионный анализ***. Мы здесь ограничимся рассмотрением однофакторного и двухфакторного дисперсионного анализа с постоянными (неслучайными) факторами. Подробное изложение предмета можно найти в литературе*.

8.2. Однофакторный дисперсионный анализ

Будем предполагать, что исследователя интересует степень влияния *фактора* A на *отклик* X. Для конкретности, пусть X — долговечность покрышки на колесе автомобиля, а фактор A — тип дорожного покрытия, который имеет l уровней (l — целое число).

Пусть $\mu_0 = \mathrm{M}X$ — среднее значение случайной величины X и пусть x_{ik} — значение X в i-м эксперименте, $i = \overline{1, n_k}$, соответствующем k-му *уровню фактора* A, $k = \overline{1, l}$. Тогда математическую модель *однофакторного дисперсионного анализа* можно представить в виде** (***линейная модель дисперсионного анализа***)

$$X_{ik} = \mu_0 + \alpha_k + \varepsilon_{ik}, \quad i = \overline{1, n_k}, \tag{8.1}$$

где α_k — вклад в величину X_{ik}, обусловленный действием фактора A (α_k — неслучайная величина); ε_{ik} — вклад в X_{ik}, обусловленный действием неучтенных факторов (случайные ошибки эксперимента, т.е. ε_{ik} — случайные величины). При этом $\sum\limits_{k=1}^{n} \alpha_k = 0$.

*См., например: *Шеффе Г.*

**См.: *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.*, 1985.

Относительно случайных величин ε_{ik} сделаем те же предположения, что и в *регрессионном анализе* (см. **7.1**, **7.3**):

– систематическая ошибка отсутствует, т.е. $\mathrm{M}\varepsilon_{ik} = 0$ для любых i и k;

– случайные ошибки эксперимента ε_{ik} не коррелированы между собой и имеют одинаковую (неизвестную) дисперсию, т.е.

$$\mathrm{M}(\varepsilon_{ik}\varepsilon_{jm}) = \begin{cases} \sigma^2, & i = j \text{ и } k = m); \\ 0, & i \neq j \text{ или } k \neq m; \end{cases}$$

– случайные ошибки эксперимента ε_{ik} имеют нормальный закон распределения с нулевым средним и неизвестной дисперсией σ^2, т.е.

$$\varepsilon_{ik} \sim N(0, \sigma^2).$$

Именно последнее допущение позволит нам проводить проверку *статистических гипотез*, используя уже известные критерии, основанные на нормальном законе распределения наблюдаемых в эксперименте случайных величин. Разумеется, принятые допущения требуют последующей проверки. Однако на первом этапе исследования они являются вполне естественными.

С учетом принятых допущений о случайных ошибках эксперимента и на основании принятой модели (8.1) делаем заключение, что случайные величины X_{ik} имеют нормальный закон распределения со средним значением $\mathrm{M}X_{ik} = \mu_0 + \alpha_k$ и дисперсией $\mathrm{D}X_{ik} = \sigma^2$, $k = \overline{1, l}$.

Таким образом, действие фактора A проявляется в том, что для каждого его уровня k $(k = \overline{1, l})$ результаты наблюдений над случайной величиной (откликом) X можно рассматривать как *случайную выборку* $X_{1k}, X_{2k}, \ldots, X_{n_k k}$ *объема* n_k *из генеральной совокупности* X_k, причем каждая случайная величина X_k, $k = \overline{1, l}$, нормально распределена со средним значением $\mu_k = \mu_0 + \alpha_k$ и дисперсией σ^2.

Отсюда следует, что статистическая гипотеза H_0, предполагающая отсутствие влияния фактора A на отклик X, означает, что $\mu_k = \mu_0 + \alpha_k = \mu_0$, или $\alpha_k = 0$, $k = \overline{1,l}$. В качестве *альтернативной гипотезы* H_1 могут выступать различные предположения о значениях величин α_k или их некоторых линейных комбинаций — далее этот вопрос рассмотрен подробно.

Итак, задача проверки влияния фактора A на отклик X по результатам эксперимента сводится к следующей формализованной постановке, если принята модель наблюдений (8.1) и сформулированные выше предположения о случайных ошибках эксперимента.

Пусть $X_1, \ldots, X_l$ — независимые случайные величины и $X_k \sim N(\mu_k, \sigma^2)$, $k = \overline{1,l}$. Пусть для каждого $k = \overline{1,l}$ дана случайная выборка $X_{1k}, \ldots, X_{n_k k}$ из генеральной совокупности случайной величины X_k, которую далее мы будем называть ***k*-й *случайной выборкой***.

Требуется по этим данным проверить на заданном *уровне значимости* α гипотезу H_0: $\mu_1 = \mu_2 = \ldots = \mu_l = \mu_0$ (или, что то же самое, H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_l = 0$, если $\mu_k = \mu_0 + \alpha_k$, $k = \overline{1,l}$).

Для нашей интерпретации отклика X (долговечность покрышки) и фактора A (тип дорожного покрытия) каждая случайная величина X_k, $k = \overline{1,l}$, характеризует долговечность покрышки на дорогах с k-м типом покрытия. Отсутствие влияния фактора A, т.е. выполнение гипотезы H_0, означает, что на дорогах с любым типом покрытия средняя долговечность одна и та же. Если гипотеза H_0 неверна, то тип покрытия (фактор A) влияет на долговечность покрышки.

Заметим, что при наличии у фактора A только двух уровней ($l = 2$) наша задача сводится к проверке стандартной гипотезы о равенстве двух средних значений нормальных совокупностей (см. **4.2**). Если фактор A имеет более двух уровней ($l > 2$), то для проверки гипотезы о равенстве l средних применяют *однофакторный дисперсионный анализ*, суть которого состоит в следующем.

Пусть X_{ik} — *i-й элемент k-й случайной выборки*, $i = \overline{1, n_k}$, $k = \overline{1, l}$, и $\overline{X}_k$ — *выборочное среднее k-й* выборки, т.е.

$$\overline{X}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ik} = \frac{1}{n_k} X_{\cdot k},$$

а $\overline{X}$ — общее выборочное среднее:

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{l}\sum_{i=1}^{n_k} X_{ik} = \frac{1}{n} X_{\cdot,\cdot},$$

где $n = n_1 + \ldots + n_l$ — общее число наблюдений.

Общая сумма квадратов отклонений наблюдений отклика от общего выборочного среднего $\overline{X}$ может быть представлена в следующем виде:

$$\sum_{k=1}^{l}\sum_{i=1}^{n_k}(X_{ik} - \overline{X})^2 = \sum_{k=1}^{l} n_k(\overline{X}_k - \overline{X})^2 + \sum_{k=1}^{l}\sum_{i=1}^{n_k}(X_{ik} - \overline{X}_k)^2. \quad (8.2)$$

Это основное тождество дисперсионного анализа, которое будем записывать кратко так:

$$Q(\vec{X}_n) = Q_A(\vec{X}_n) + Q_l(\vec{X}_n), \quad (8.3)$$

где $Q(\vec{X}_n)$ — общая сумма квадратов отклонений отклика от общего среднего; $Q_A(\vec{X}_n)$ — сумма квадратов отклонений, обусловленных отличием выборочных средних $\overline{X}_k$ по группам (уровням) от общего выборочного среднего $\overline{X}$ (среднее квадратичное отклонение между группами или между уровнями); $Q_l(\vec{X}_n)$ — сумма квадратов отклонений наблюдений от выборочных средних для каждого уровня (внутри групп).

Тождество (8.2) легко проверяется, для чего нужно возвести в квадрат и просуммировать по i и k очевидное равенство

$$X_{ik} - \overline{X} = (\overline{X}_k - \overline{X}) + (X_{ik} - \overline{X}_k)$$

и учесть, что

$$\sum_{k=1}^{l}\sum_{i=1}^{n_k}(X_{ik}-\overline{X}_k)(\overline{X}_k-\overline{X})=\sum_{k=1}^{l}(\overline{X}_k-\overline{X})\sum_{i=1}^{n_k}(X_{ik}-\overline{X}_k)=0$$

в силу определения выборочных средних $\overline{X}_k$ и $\overline{X}$. Действительно, внутренняя сумма

$$\sum_{i=1}^{n_k}(X_{ik}-\overline{X}_k)=\sum_{i=1}^{n_k}X_{ik}-n_k\overline{X}_k=n_k\overline{X}_k-n_k\overline{X}_k=0.$$

Можно показать*, что если гипотеза H_0: $\mu_1=\mu_2=\ldots=\mu_l$ верна, то *статистики* $Q_A(\vec{X}_n)/\sigma^2$ и $Q_l(\vec{X}_n)/\sigma^2$ независимы и имеют χ^2*-распределение* с числом степеней свободы соответственно $l-1$ и $n-l$, а статистики $S_A^2(\vec{X}_n)=Q_A(\vec{X}_n)/(l-1)$ и $S_l^2(\vec{X}_n)=Q_l(\vec{X}_n)/(n-l)$ являются *несмещенными оценками* неизвестной дисперсии σ^2.

Оценка $S_A^2(\vec{X}_n)$ характеризует рассеяние средних значений $\overline{X}_k$, а оценка $S_l^2(\vec{X}_n)$ — рассеяние выборочных значений X_{ik} внутри групп, которое обусловлено действием неучтенных факторов. Значительное превышение величины $S_A^2(\vec{X}_n)$ над значением величины $S_l^2(\vec{X}_n)$ можно объяснить различием средних значений m_k, $k=\overline{1,l}$, в группах (для различных уровней фактора A), т.е. существенным влиянием фактора A.

Таким образом, если гипотеза H_0: $\mu_1=\mu_2=\ldots=\mu_l$ верна, то

$$F=\frac{Q_A(\vec{X}_n)/(l-1)}{Q_l(\vec{X}_n)/(n-l)}=\frac{S_A^2(\vec{X}_n)}{S_l^2(\vec{X}_n)}\sim F(r_A,r_l), \qquad (8.4)$$

т.е. статистика F имеет распределение Фишера с числом степеней свободы $r_A=l-1$ и $r_l=n-l$ (см. **Д.3.1**).

Статистику F используют для проверки гипотезы H_0: $\mu_1=\ldots=\mu_l=\mu_0$. Гипотеза H_0 не противоречит результатам

*См.: *Крамер Г.*

наблюдений, если выборочное значение $F_в$ статистики (8.4) меньше ее критического уровня $F_{кр} = F_{1-\alpha}(r_A, r_l)$, т.е. если

$$F_в \leqslant F_{кр} = F_{1-\alpha}(r_A, r_l).$$

Если же

$$F_в > F_{кр} = F_{1-\alpha}(r_A, r_l),$$

то гипотеза H_0 отклоняется и следует считать, что среди средних значений $\mu_1, \ldots, \mu_l$ имеются хотя бы два, не равных друг другу.

В случае принятия гипотезы H_0 в качестве несмещенных оценок параметров μ_0 и σ^2 можно взять соответственно $\overline{X}$ и $S_l^2 = Q_l(\vec{X}_n)/(n-l)$.

Результаты проверки гипотезы H_0 принято оформлять в виде так называемой ***таблицы дисперсионного анализа*** (табл. 8.1).

Таблица 8.1

Источник изменчивости	Сумма квадратов (СК)	Степени свободы	Средняя сумма квадратов	Статистика F
Между группами (фактор A)	$Q_A(\vec{X}_n) = \sum_{k=1}^{l} n_k(\overline{X}_k - \overline{X})^2$	$l-1$	$S_A^2(\vec{X}_n) = \frac{Q_A(\vec{X}_n)}{r_A}$	$F_в = S_A^2/S_l^2$
Внутри групп (ошибки)	$Q_l(\vec{X}_n) = \sum_{k=1}^{l}\sum_{i=1}^{n_k}(X_{ik} - \overline{X}_k)^2$	$n-l$	$S_l^2(\vec{X}_n) = \frac{Q_l(\vec{X}_n)}{r_l}$	
Общая сумма квадратов	$Q(\vec{X}_n) = \sum_{k=1}^{l}\sum_{i=1}^{n_k}(X_{ik} - \overline{X})^2$	$n-1$		$F_{кр} = F_{1-\alpha}(r_A, r_l)$

Пример 8.1. Три группы операторов ЭВМ обучались по трем различным методикам. После окончания срока обучения

был проведен тестовый контроль случайно отобранных операторов из каждой группы. Получены следующие результаты (табл. 8.2).

Таблица 8.2

Номер группы k	Число ошибок, допущенных операторами, X_{ik}	Сумма $\sum_{i=1}^{n_k} X_{ik}$	Число контролируемых операторов n_k
1	1, 3, 2, 1, 0, 2, 1	10	7
2	2, 3, 2, 1, 4, –, –	12	5
3	4, 5, 3, –, –, –, –	12	3

Требуется на уровне значимости $\alpha = 0{,}05$ проверить гипотезу об отсутствии влияния различных методик обучения на результаты тестового контроля операторов. Предполагается, что выборки получены из независимых нормально распределенных совокупностей с одной и той же дисперсией.

В данном случае фактор A — это тип методики обучения, имеющий $l = 3$ уровня. Объем наблюдений $n = n_1 + n_2 + n_3 = 15$. Проверяется гипотеза H_0: $\mu_1 = \mu_2 = \mu_3$, где μ_k — математическое ожидание числа ошибок, допущенных операторами k-й группы.

Сперва вычисляем суммы

$$x_{\cdot,\cdot} = \sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik} = 10 + 12 + 12 = 34, \qquad \sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik}^2 = 104.$$

Затем, используя (8.2) и (8.3), находим

$$Q = \sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik}^2 - \frac{1}{n}x_{\cdot,\cdot}^2 = 104 - \frac{1}{15}\cdot 34^2 \approx 26{,}93,$$

$$Q_A = \sum_{k=1}^{l}\frac{1}{n_k}x_{\cdot k}^2 - \frac{1}{n}x_{\cdot,\cdot}^2 = 91{,}08 - \frac{1}{15}\cdot 34^2 \approx 14{,}02,$$

$$Q_l = Q - Q_A = 26{,}93 - 14{,}02 = 12{,}91.$$

Теперь вычисляем выборочное значение статистики (8.4):

$$f_в = \frac{Q_A/(l-1)}{Q_l/(n-l)} = \frac{14{,}02/2}{12{,}91} \approx 6{,}52.$$

Из таблицы квантилей распределения Фишера (см. табл. П.5) для уровня значимости $\alpha = 0{,}05$ и степеней свободы $r_A = l-1 = 2$, $r_l = n - l = 12$ находим $F_{кр} = F_{0{,}95}(2{,}12) = 3{,}89$. Так как $F_в = 6{,}52 > F_{кр}$, то гипотеза H_0 о равенстве средних отклоняется. Это означает, что исследуемые методики обучения операторов дают значимо различные результаты тестового контроля.

8.3. Понятие линейных контрастов

Если *гипотеза* H_0 о равенстве средних значений l нормальных *генеральных совокупностей* отклоняется (т.е. хотя бы в какой-то паре групп средние отличаются друг от друга), то требуется определить, какие именно группы имеют значимое различие средних. Для этой цели используются так называемые ***линейные контрасты***. Линейный контраст L определяется как линейная комбинация

$$L = \sum_{k=1}^{l} c_k \mu_k, \tag{8.5}$$

где c_k — постоянные, однозначно определяемые из формулировки проверяемых гипотез, причем $c_1 + \ldots + c_l = 0$.

Примерами линейных контрастов являются:

$L^{(1)} = \mu_1 - \mu_2$; здесь $c_1 = 1$, $c_2 = -1$, $c_3 = 0$, а выдвигаемая гипотеза $H_0^{(1)}$: $\mu_1 - \mu_2 = 0$;

$L^{(2)} = 0{,}5(\mu_1 + \mu_3) - \mu_2$; здесь $c_1 = c_3 = 0{,}5$, $c_2 = -1$, а выдвигаемая гипотеза $H_0^{(2)}$: $0{,}5(\mu_1 + \mu_3) - \mu_2 = 0$.

Таким образом, если гипотеза H_0: $\mu_1 = \mu_2 = \ldots = \mu_l$ отклоняется, то с помощью линейного контраста можно выдвинуть вспомогательные нулевые гипотезы относительно различных

линейных комбинаций средних значений $\mu_1, \ldots, \mu_l$, образующих линейный контраст.

Любая такая гипотеза имеет вид H_0': $L = c_1\mu_1 + \ldots + c_l\mu_l$ при некотором заданном наборе постоянных c_k, для которых $c_1 + \ldots + c_l = 0$.

Нетрудно увидеть, что *несмещенной оценкой* линейного контраста L (при сделанных выше предположениях о *случайных ошибках* эксперимента ε_{ik}) является *оценка*

$$\widehat{L}(\vec{X}_n) = \sum_{k=1}^{l} c_k \overline{X}_k, \tag{8.6}$$

дисперсия которой (с учетом того, что $\mathbf{D}\overline{X}_k = \sigma^2/n_k$ и $\overline{X}_k$ — независимые случайные величины) равна

$$\mathbf{D}\,\widehat{L}(\vec{X}_n) = \sigma^2 \sum_{k=1}^{l} \frac{c_k^2}{n_k}. \tag{8.7}$$

При этом *статистика* $\widehat{L}(\vec{X}_n)$ имеет нормальный закон распределения со средним $L = c_1\mu_1 + \ldots + c_l\mu_l$ и дисперсией $\mathbf{D}\,\widehat{L}(\vec{X}_n)$, т.е.

$$\widehat{L}(\vec{X}_n) \sim N\big(L, \mathbf{D}\,\widehat{L}(\vec{X}_n)\big), \tag{8.8}$$

Следовательно,

$$T = \frac{\widehat{L}(\vec{X}_n) - L}{\sqrt{\mathbf{D}\,\widehat{L}(\vec{X}_n)}} \sim N(0,1), \tag{8.9}$$

т.е. статистика T имеет стандартное нормальное распределение.

Последнее утверждение следует из того, что *выборочные средние* $\overline{X}_k$ имеют нормальное распределение, $X_k \sim N(\mu_k, \sigma^2)$, $k = \overline{1,l}$, а линейная комбинация $\widehat{L} = c_1\overline{X}_1 + \ldots + c_l\overline{X}_l$ нормально распределенных случайных величин также распределена по нормальному закону с параметрами $\mathbf{M}\,\widehat{L}(\vec{X}_n) = L$ и $\mathbf{D}\,\widehat{L}(\vec{X}_n) = \sigma^2 c_1^2/n_1 + \ldots + c_l^2/n_l$. Кроме того, статистика $Q_l(\vec{X}_n)/\sigma^2$ име-

ет χ^2-*распределение* с числом степеней свободы $r_l = n - l$, т.е.

$$V = \frac{(n-l)S_l^2(\vec{X}_n)}{\sigma^2} \sim \chi^2(n-l), \tag{8.10}$$

и можно показать, что V и T — независимые случайные величины. На основании (8.9) и (8.10) приходим к следующему *критерию* проверки гипотезы H_0': $L = c_1\mu_1 + \ldots + c_l\mu_l = 0$.

Если гипотеза H_0' верна, то статистика $t = T/\sqrt{V(n-l)}$ имеет распределение Стьюдента с числом степеней свободы $n-l$, т.е.

$$t = \frac{\sum\limits_{k=1}^{l} c_k\overline{X}_k}{S_l(\vec{X}_n)\sqrt{\sum\limits_{k=1}^{l}\frac{c_k^2}{n_k}}} \sim S(n-l). \tag{8.11}$$

Таким образом, гипотезу H_0 следует отклонить на *уровне значимости* α (т.е. считать значимым отличие от нуля выбранной линейной комбинации средних $\mu_1, \mu_2, \ldots, \mu_l$), если *выборочное значение* $t_в$ статистики (8.11) по абсолютной величине превышает $t_{кр} = t_{1-\alpha/2}(n-l)$:

$$|t_в| > t_{кр} = t_{1-\alpha/2}(n-l).$$

Пример 8.2. В условиях примера 8.1 при двусторонних *альтернативных гипотезах* проверим гипотезы $H_0^{(1)}$: $\mu_1 = \mu_2$, $H_0^{(2)}$: $\mu_1 = \mu_3$, $H_0^{(3)}$: $\mu_2 = \mu_3$, $H_0^{(4)}$: $\frac{1}{2}(\mu_1 + \mu_3) = \mu_2$.

В соответствии с проверяемыми гипотезами $H_0^{(i)}$, $i = \overline{1,4}$, определим линейные контрасты

$$L_1 = \mu_1 - \mu_2 \quad (c_1 = 1,\ c_2 = -1,\ c_3 = 0);$$
$$L_2 = \mu_1 - \mu_3 \quad (c_1 = 1,\ c_2 = 0,\ c_3 = -1);$$
$$L_3 = \mu_2 - \mu_3 \quad (c_1 = 0,\ c_2 = 1,\ c_3 = -1);$$
$$L_4 = \frac{1}{2}(\mu_1 + \mu_2) - \mu_3 \quad \left(c_1 = \frac{1}{2},\ c_2 = \frac{1}{2},\ c_3 = -1\right).$$

Предварительно вычислим значения оценок линейных контрастов L_i, $i=\overline{1,4}$, и их дисперсий. Выборочные средние $\overline{x}_1=1{,}43$, $\overline{x}_2=2{,}4$, $\overline{x}_3=4$. Значение оценки дисперсии

$$S_l^2=\frac{Q_l}{n-l}=\frac{12{,}91}{15-3}\approx 1{,}08.$$

Значения оценок контрастов и их дисперсий равны:

$$\widehat{L}_1=1{,}43-2{,}4=-0{,}97,\qquad \mathbf{D}\,\widehat{L}_1=1{,}08\Big(\frac{1}{7}+\frac{1}{5}\Big)\approx 0{,}37;$$

$$\widehat{L}_2=1{,}43-4=-2{,}57,\qquad \mathbf{D}\,\widehat{L}_2=1{,}08\Big(\frac{1}{7}+\frac{1}{3}\Big)\approx 0{,}51;$$

$$\widehat{L}_3=2{,}4-4=-1{,}60,\qquad \mathbf{D}\,\widehat{L}_3=1{,}08\Big(\frac{1}{5}+\frac{1}{3}\Big)\approx 0{,}58;$$

$$\widehat{L}_4=\frac{1}{2}(1{,}43+2{,}4)-4=-2{,}08,$$

$$\mathbf{D}\,\widehat{L}_4=1{,}08\Big(\frac{(1/2)^2}{7}+\frac{(1/2)^2}{5}+\frac{1}{3}\Big)\approx 0{,}45.$$

Следовательно, выборочные значения $|t_{\text{в}}^{(i)}|$ статистики (8.11) равны:

- для гипотезы $H_0^{(1)}$: $|t_{\text{в}}^{(1)}|=\Big|\dfrac{\widehat{L}_1}{\sqrt{\mathbf{D}\,\widehat{L}_1}}\Big|=\dfrac{0{,}97}{\sqrt{0{,}37}}\approx 1{,}595$;
- для гипотезы $H_0^{(2)}$: $|t_{\text{в}}^{(2)}|=\Big|\dfrac{\widehat{L}_2}{\sqrt{\mathbf{D}\,\widehat{L}_2}}\Big|=\dfrac{2{,}57}{\sqrt{0{,}51}}\approx 3{,}598$;
- для гипотезы $H_0^{(3)}$: $|t_{\text{в}}^{(3)}|=\Big|\dfrac{\widehat{L}_3}{\sqrt{\mathbf{D}\,\widehat{L}_3}}\Big|=\dfrac{1{,}60}{\sqrt{0{,}58}}\approx 2{,}101$;
- для гипотезы $H_0^{(4)}$: $|t_{\text{в}}^{(4)}|=\Big|\dfrac{\widehat{L}_4}{\sqrt{\mathbf{D}\,\widehat{L}_4}}\Big|=\dfrac{2{,}08}{\sqrt{0{,}45}}\approx 3{,}002$.

Критическое значение $t_{\text{кр}}=t_{0{,}975}(12)=2{,}179$. Так как $|t_{\text{в}}^{(1)}|<t_{\text{кр}}$ и $|t_{\text{в}}^{(3)}|<t_{\text{кр}}$, то гипотезы $H_0^{(1)}$ и $H_0^{(3)}$ принимаются. Гипотезы $H_0^{(2)}$ и $H_0^{(4)}$ отклоняются, ибо $|t_{\text{в}}^{(2)}|>t_{\text{кр}}$ и $|t_{\text{в}}^{(4)}|>t_{\text{кр}}$.

Таким образом, значимо различны средние первой и третьей группы, а также среднее арифметическое средних для первых двух групп и среднее третьей группы.

8.4. Двухфакторный дисперсионный анализ

Рассмотрим случай влияния двух *факторов* на *отклик* X. В этом случае *дисперсионный анализ* основывается на результатах эксперимента, проводимого на различных *уровнях* каждого из факторов.

Будем предполагать, что взаимосвязь между факторами отсутствует*. Для простоты изложения ограничимся случаем, когда для каждой пары уровней рассматриваемых факторов проводится по одному наблюдению. Через l_A обозначим число уровней фактора A, а через l_B — число уровней фактора B. Тогда общее число наблюдений для всех возможных пар уровней факторов A и B равно $n = l_A l_B$.

Математическую модель двухфакторного дисперсионного анализа в этом случае можно представить в виде

$$X_{ij} = \mu_0 + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = \overline{1, l_A}, \quad j = \overline{1, l_B}, \tag{8.12}$$

где X_{ij} — отклик X на i-м уровне фактора A и j-м уровне фактора B; $\mu_0 = \mathrm{M}X$; α_i, β_j — неслучайные величины, характеризующие вклады в X_{ij}, обусловленные действием соответствующих факторов A и B; ε_{ij} — случайная величина, характеризующая вклад в X_{ij}, обусловленный действием неучтенных факторов.

Предположения, сделанные в 8.2 относительно случайных величин ε_{ij}, остаются в силе. При этом

$$\mathrm{M}X_{ij} = m_0 + \alpha_i + \beta_j$$

и $\alpha_1 + \ldots + \alpha_{l_A} = \beta_1 + \ldots + \beta_{l_B} = 0$, что и означает независимость факторов A и B.

Поскольку в модели (8.12) взаимодействие факторов отсутствует, *проверка гипотез* о влиянии факторов A и B на отклик X проводится отдельно для каждого фактора. Рассмотрим

*См.: *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.*, 1985.

критерии для проверки гипотез о влиянии фактора A (фактора B) на отклик X. Введем обозначения

$$\overline{X}_{i\cdot} = \frac{1}{l_B}\sum_{j=1}^{l_B} X_{ij}, \quad \overline{X}_{\cdot j} = \frac{1}{l_A}\sum_{i=1}^{l_A} X_{ij}, \quad \overline{X} = \frac{1}{l_A l_B}\sum_{i=1}^{l_A}\sum_{j=1}^{l_B} X_{ij}.$$

Общая сумма квадратов отклонений X_{ij} от выборочного среднего $\overline{X}$ может быть представлена в виде

$$Q(\vec{X}_n) = \sum_{i=1}^{l_A}\sum_{j=1}^{l_B}(X_{ij} - \overline{X})^2 = l_B\sum_{i=1}^{l_A}(\overline{X}_{i\cdot} - \overline{X})^2 + \\ + l_A\sum_{i=1}^{l_B}(\overline{X}_{\cdot j} - \overline{X})^2 + \sum_{i=1}^{l_A}\sum_{j=1}^{l_B}(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X})^2$$

(в этом можно убедиться с помощью рассуждений, аналогичных приведенным в 8.2). Отсюда вытекает равенство

$$Q(\vec{X}_n) = Q_A(\vec{X}_n) + Q_B(\vec{X}_n) + Q_0(\vec{X}_n), \qquad (8.13)$$

где слагаемое

$$Q_A(\vec{X}_n) = l_B\sum_{i=1}^{l_A}(\overline{X}_{i\cdot} - \overline{X})^2$$

обусловлено отличием *выборочных средних* $\overline{X}_{i\cdot}$ и $\overline{X}$, т.е. влиянием фактора A на отклик X; слагаемое

$$Q_B(\vec{X}_n) = l_A\sum_{i=1}^{l_B}(\overline{X}_{\cdot j} - \overline{X})^2$$

обусловлено отличием выборочных средних $\overline{X}_{\cdot j}$ и $\overline{X}$, т.е. влиянием фактора B на отклик X; слагаемое

$$Q_0(\vec{X}_n) = \sum_{i=1}^{l_A}\sum_{j=1}^{l_B}(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X})^2$$

учитывает влияние всех факторов, в том числе и неучтенных.

Проверка гипотез о влиянии факторов A и B на отклик X основана на сравнении *статистик* $Q_A(\vec{X}_n)$ и $Q_B(\vec{X}_n)$ с $Q_0(\vec{X}_n)$.

Проверим, например, гипотезу H_0 о том, что фактор A не влияет на отклик X, т.е. $\alpha_i = 0$, $i = \overline{1, l_A}$.

Если гипотеза H_0 верна, то при сделанных выше предположениях относительно ε_{ij}, $i = \overline{1, l_A}$, $j = \overline{1, l_B}$, статистики $Q_A(\vec{X}_n)/\sigma^2$ и $Q_0(\vec{X}_n)/\sigma^2$ независимы и имеют *χ^2-распределение* с числом степеней свободы $l_A - 1$ и $(l_A - 1)(l_B - 1)$ соответственно, а статистики

$$S_A^2(\vec{X}_n) = \frac{Q_A(\vec{X}_n)}{l_A - 1} \quad \text{и} \quad S_0^2(\vec{X}_n) = \frac{Q_0(\vec{X}_n)}{(l_A - 1)(l_B - 1)} \tag{8.14}$$

являются *несмещенными оценками* дисперсии σ^2 отклика* X. Отсюда следует (см. **Д.3.1**), что

$$F = \frac{S_A^2(\vec{X}_n)}{S_0^2(\vec{X}_n)} \sim F\big(l_A - 1,\, (l_A - 1)(l_B - 1)\big). \tag{8.15}$$

Гипотеза H_0 не противоречит результатам наблюдений, если выборочное значение $f_в$ статистики $S_A^2(\vec{X}_n)/S_0^2(\vec{X}_n)$ не превосходит $f_{кр} = f_{1-\alpha}\big(l_A - 1,\, (l_A - 1)(l_B - 1)\big)$ для заданного *уровня значимости* α. В противном случае, т.е. если

$$f_в > f_{кр},$$

гипотезу H_0 отклоняют.

Если приходится отвергать гипотезу H_0, то может возникнуть необходимость в проверке одной из гипотез $H_0^{(i)}$, согласно которой влияние на отклик оказывает i-й уровень фактора A, т.е. проверяют гипотезу

$$H_0^i\colon\ \alpha_1 = \ldots = \alpha_{i-1} = \alpha_{i+1} = \ldots = \alpha_{l_A} = 0,\ \ \alpha_i \neq 0.$$

*См.: *Крамер Г.*

Пусть $i = 1$, а

$$\overline{X}_{\cdot(2\ldots l_A)} = \frac{1}{(l_A - 1)l_B} \sum_{i=2}^{l_A} \sum_{j=1}^{l_B} X_{ij}.$$

Тогда сумма квадратов $Q_A(\vec{X}_n)$ может быть представлена в виде

$$Q_A(\vec{X}_n) = Q'_A(\vec{X}_n) + Q''_A(\vec{X}_n), \tag{8.16}$$

где

$$Q'_A(\vec{X}_n) = \frac{(l_A - 1)l_B}{l_A} \left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2,$$

$$Q''_A(\vec{X}_n) = l_B \sum_{i=2}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2.$$

Действительно, учитывая равенства

$$\overline{X} = \frac{1}{l_A l_B} \sum_{i=1}^{l_A} \sum_{j=1}^{l_B} X_{ij} = \frac{1}{l_A l_B} \sum_{i=2}^{l_A} \sum_{j=1}^{l_B} X_{ij} + \frac{1}{l_A l_B} \sum_{j=1}^{l_B} X_{1j} =$$

$$= \frac{l_A - 1}{l_A} \overline{X}_{\cdot(2\ldots l_A)} + \frac{1}{l_A} \sum_{j=1}^{L_B} X_{ij},$$

находим

$$Q_A(\vec{X}_n) = l_B \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}\right)^2 =$$

$$= l_B \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \frac{l_A - 1}{l_A} \overline{X}_{\cdot(2\ldots l_A)} - \frac{1}{l_A} \overline{X}_{1\cdot}\right)^2 =$$

$$= l_B \sum_{i=1}^{l_A} \left(\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right) + \frac{1}{l_A} \left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right) \right)^2.$$

В полученной сумме преобразуем каждое слагаемое по формуле квадрата суммы. В результате находим

$$Q_A(\vec{X}_n) = l_B \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 + \frac{l_B}{l_A^2} \sum_{i=1}^{l_A} \left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right)^2 +$$

$$+ 2\frac{l_B}{l_A} \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)\left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right) =$$

$$= l_B \sum_{i=2}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 + l_B \left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 +$$

$$+ \frac{l_B}{l_A} \left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right)^2 + 2\frac{l_B}{l_A} \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)\left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right).$$

Так как в силу определения величин $\overline{X}_{i\cdot}$ и $\overline{X}_{\cdot(2\ldots l_A)}$

$$\sum_{i=2}^{l_A} \overline{X}_{i\cdot} = (l_A - 1)\overline{X}_{\cdot(2\ldots l_A)},$$

то

$$\sum_{i=2}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right) = \sum_{i=2}^{l_A} \overline{X}_{i\cdot} - (l_A - 1)\overline{X}_{\cdot(2\ldots l_A)} = 0.$$

Поэтому

$$\sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)\left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right) =$$

$$= -\left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 + \left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right) \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right) =$$

$$= -\left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2.$$

Собирая теперь все слагаемые, получаем

$$Q_A(\vec{X}_n) = l_B \sum_{i=2}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 + \\ + l_B\left(1 - \frac{2}{l_A} + \frac{1}{l_A}\right)\left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2,$$

что равносильно (8.16).

Для проверки гипотезы $H_0^{(i)}$ по результатам наблюдений используют статистику

$$F = \frac{S''_A(\vec{X}_n)^2}{S_0^2(\vec{X}_n)},$$

где

$$S''_A(\vec{X}_n)^2 = \frac{Q''_A(\vec{X}_n)}{l_A - 2}.$$

Эта статистика имеет *распределение Фишера* с числом степеней свободы $l_A - 2$ и $(l_A - 1)(l_B - 1)$, если гипотеза $H_0^{(i)}$ верна*.

Аналогично строятся критерии для проверки влияния фактора B на отклик X.

Порядок проведения двухфакторного анализа представим в виде таблицы (табл. 8.3).

8.5. Решение типовых примеров

Пример 8.3. Результаты измерений продолжительности (в секундах) химической реакции при различном содержании катализатора даны в табл. 8.4. Проверим *гипотезу* H_0 о том, что время химической реакции не зависит от процентного содержания катализатора на *уровне значимости* $\alpha = 0{,}01$.

*См.: *Крамер Г.*

Таблица 8.3

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средняя сумма квадратов	Статистика
Фактор A	$Q_A(\vec{X}_n) = l_B \sum_{i=1}^{l_A} (\overline{X}_{i\cdot} - \overline{X})^2$	$l_A - 1$	$S_A^2(\vec{X}_n) = \dfrac{Q_A(\vec{X}_n)}{l_A - 1}$	$F = \dfrac{S_A^2(\vec{X}_n)}{S_0^2(\vec{X}_n)}$
Фактор B	$Q_B(\vec{X}_n) = l_A \sum_{j=1}^{l_A} (\overline{X}_{\cdot j} - \overline{X})^2$	$l_B - 1$	$S_B^2(\vec{X}_n) = \dfrac{Q_B(\vec{X}_n)}{l_B - 1}$	$F = \dfrac{S_B^2(\vec{X}_n)}{S_0^2(\vec{X}_n)}$
Ошибки	$Q(\vec{X}_n) = \sum_{i=1}^{l_A} \sum_{j=1}^{l_B} (X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X})^2$	$(l_A - 1) \times (l_B - 1)$	$S_0^2(\vec{X}_n) = \dfrac{Q_0(\vec{X}_n)}{(l_A - 1)(l_B - 1)}$	
Сумма	$Q_0(\vec{X}_n) = \sum_{i=1}^{l_A} \sum_{j=1}^{l_B} (X_{ij} - \overline{X})^2$	$l_A l_B - 1$		

Таблица 8.4

Содержание катализатора, %	Номер эксперимента												Сумма по строкам
	1	2	3	4	5	6	7	8	9	10	11	12	
5	5,9	6,0	7,0	6,5	5,5	7,0	8,1	7,5	6,2	6,4	7,1	6,9	80,1
10	4,0	5,1	6,2	5,3	4,5	4,4	5,3	5,4	5,6	5,2			51,0
15	8,2	6,8	8,0	7,5	7,0	7,2	7,9	8,1	8,5	7,8	8,1		85,1

В этой задаче *фактор A* — процентное содержание катализатора, а случайная величина X (*отклик*) — время химической реакции.

Для проверки гипотезы H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_0$ о равенстве средних значений μ_i, $i = 1,2,3$, времени химической реакции при различных *уровнях фактора A* (5, 10, 15% содержания

катализатора) используем *статистику* (8.4)

$$F = \frac{Q_A(\vec{X}_n)/(l-1)}{Q_l(\vec{X}_n)/(n-l)}.$$

Находим *выборочное значение* статистики $F_в$ по результатам эксперимента. Используя (8.2) и (8.3), вычисляем величины

$$Q = \sum_{k=1}^{3}\sum_{i=1}^{n_k}(x_{ik}-\overline{x})^2 = \sum_{k=1}^{3}\sum_{i=1}^{n_k}x_{ik}^2 - \frac{1}{33}\Big(\sum_{k=1}^{3}\sum_{i=1}^{n_k}x_{ik}\Big)^2 =$$

$$= 1465{,}68 - \frac{1}{33}(216{,}2)^2 = 1465{,}68 - 1416{,}44 = 49{,}24$$

и

$$Q_A = \sum_{k=1}^{3} n_k(\overline{x}_k - \overline{x})^2 = \sum_{k=1}^{3}\frac{1}{n_k}\Big(\sum_{i=1}^{n_k}x_{ik}\Big)^2 - \frac{1}{33}\Big(\sum_{k=1}^{3}\sum_{i=1}^{n_k}x_{ik}\Big)^2 =$$

$$= \frac{1}{12}\cdot 80{,}1^2 + \frac{1}{10}\cdot 51^2 + \frac{1}{11}\cdot 85{,}1^2 - 1416{,}44 =$$

$$= 1453{,}132 - 1416{,}44 = 36{,}692.$$

Теперь определяем разность этих величин

$$Q_l = Q - Q_A = 49{,}24 - 36{,}602 = 12{,}638$$

и выборочное значение статистики

$$f_в = \frac{36{,}692/2}{12{,}638/30} = \frac{18{,}346}{0{,}421} = 43{,}58.$$

По таблице квантилей распределения Фишера (см. табл. П.5) находим $f_{кр} = f_{0{,}99}(2, 30) = 3{,}25$. Так как

$$f_в = 43{,}47 > f_{кр} = 3{,}25,$$

гипотезу H_0 следует отклонить.

Пример 8.4. В условиях примера 8.3 проверим гипотезы: а) $H_0^{(1)}$: $\mu_1 = \mu_2$; б) $H_0^{(2)}$: $\mu_2 = \mu_3$.

а. Чтобы проверить гипотезу $H_0^{(1)}$, воспользуемся статистикой (8.11)

$$t = \frac{\sum\limits_{k=1}^{l} c_k \overline{X}_k}{S_l \sqrt{\sum\limits_{k=1}^{l} \frac{c_k^2}{n_k}}}.$$

В данном случае $c_1 = 1$, $c_2 = -1$, $c_3 = 0$ (см. **8.3**). Найдем значения выборочных средних $\overline{x}_k$, $k = 1, 2, 3$:

$$\overline{x}_1 = \frac{80{,}1}{12} = 6{,}675, \quad \overline{x}_2 = \frac{51{,}1}{10} = 5{,}1, \quad \overline{x}_3 = \frac{85{,}1}{11} = 7{,}736.$$

Затем определим значение

$$S_l = \sqrt{\frac{Q_l}{30}} = \sqrt{\frac{12{,}638}{30}} = 0{,}649.$$

Наконец, вычислим выборочное среднее значение статистики

$$t_в = \frac{6{,}675 - 5{,}1}{0{,}649\sqrt{\frac{1}{12} + \frac{1}{10}}} = \frac{1{,}175}{0{,}119} = 9{,}874.$$

По таблице квантилей распределения Стьюдента (см. табл. П.2) находим $t_{кр} = t_{0{,}995}(30) = 3{,}030$. Гипотезу $H_0^{(1)}$ отвергаем, поскольку

$$|t_в| = 9{,}874 > t_{кр} = 3{,}030.$$

б. В случае гипотезы $H_0^{(2)}$ имеем

$$t_в = \frac{5{,}1 - 7{,}736}{0{,}649\sqrt{\frac{1}{10} + \frac{1}{11}}} = -\frac{2{,}636}{0{,}124} = -21{,}258.$$

Так как

$$|t_{\text{в}}| = 21{,}258 > t_{\text{кр}} = 3{,}030,$$

то гипотезу $H_0^{(2)}$ отвергаем.

Пример 8.5. В табл. 8.5 приведены опытные данные спектрографического исследования с целью проверки влияния различных фотопленок (фактор A) и электродов (фактор B) на величину X (отклик), характеризующую интенсивность света.

Таблица 8.5

Уровни фактора $B(j)$	Уровни фактора $A(i)$				
	1	2	3	4	5
1	4	18	26	38	44
2	3	19	25	35	43
3	6	18	24	28	39
4	7	13	21	31	38

В данном случае фактор A имеет $l_A = 5$ уровней, фактор B — $l_B = 4$ уровня, число опытов равно $n = l_A l_B = 20$.

Проверим на уровне значимости $\alpha = 0{,}01$ гипотезы:

H_0^A — отсутствие влияния фактора A на величину X;

H_0^B — отсутствие влияния фактора B на величину X.

Для этого рассчитаем

$$\overline{x} = \frac{1}{l_A l_B} \sum_{i=1}^{5} \sum_{j=1}^{4} x_{ij} = \frac{1}{20}(20 + 68 + 96 + 132 + 164) = 24.$$

Значения статистик $\overline{X}_{i\cdot}$ и $\overline{X}_{\cdot j}$, вычисленные по формулам

$$\overline{x}_{i\cdot} = \frac{1}{l_B} \sum_{j=1}^{4} x_{ij} \qquad \text{и} \qquad \overline{x}_{\cdot j} = \frac{1}{l_A} \sum_{i=1}^{5} x_{ij},$$

приведены соответственно в табл. 8.6 и 8.7.

Таблица 8.6

i	1	2	3	4	5
$\overline{X}_{i\cdot}$	20	68	96	132	164

Таблица 8.7

i	1	2	3	4
$\overline{X}_{\cdot j}$	26	25	23	22

Далее вычисляем

$$Q_A = l_B \sum_{i=1}^{l_A} (\overline{x}_{i\cdot} - \overline{x})^2 = 4(361 + 49 + 81 + 289) = 3120,$$

$$Q_B = 5 \sum_{j=1}^{4} (\overline{x}_{\cdot j} - \overline{x})^2 = 5(4 + 1 + 1 + 4) = 50,$$

$$Q = l_B \sum_{i=1}^{5} \sum_{j=1}^{4} (x_{ij} - \overline{x})^2 =$$

$$= 400 + 36 + 4 + 196 + 400 + 441 + 25 + 1 + 100 + 361 +$$

$$+ 324 + 36 + 16 + 225 + 289 + 121 + 9 + 49 + 196 = 3229.$$

Таблица 8.8

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средняя сумма квадратов	Статистика
Фактор A	$Q_A = 3120$	$l_A - 1 = 4$	$S_A^2 = \dfrac{Q_A}{l_A - 1} = 780$	$F_в = \dfrac{S_A^2}{S_0^2} = 158{,}54$
Фактор B	$Q_B = 50$	$l_B - 1 = 3$	$S_B^2 = \dfrac{Q_B}{l_B - 1} = 16{,}67$	$F_в = \dfrac{S_B^2}{S_0^2} = 3{,}39$
Ошибки	$Q_0 = 59$	$(l_A - 1) \times (l_B - 1) = 12$	$S_0^2 = \dfrac{Q_0}{(l_A - 1)(l_B - 1)} = 4{,}92$	
Сумма	$Q = 3229$	$l_A l_B - 1 = 19$		

Находим разность вычисленных величин:

$$Q_0 = Q - Q_A - Q_B = 59.$$

Полученные результаты сведем в таблицу (табл. 8.8). Поскольку

$$f_в = 158{,}7 > f_{кр} = f_{0{,}01}(4, 12) = 5{,}41,$$

то гипотезу H_0^A следует отвергнуть. Гипотезу H_0^B следует принять, так как

$$f_в = 3{,}39 < f_{кр} = f_{0{,}01}(3, 12) = 5{,}92.$$

Вопросы и задачи

8.1. В каком случае дисперсионный анализ называют однофакторным, двухфакторным?

8.2. Какой вид имеет математическая модель (линейная модель) однофакторного дисперсионного анализа?

8.3. Запишите основное тождество дисперсионного анализа в случае: а) действия одного фактора; б) действия двух факторов.

8.4. Что такое линейный контраст?

8.5. Сформулируйте критерии для проверки статистической гипотезы об одинаковом действии фактора на всех уровнях в случае: а) однофакторного дисперсионного анализа; б) двухфакторного дисперсионного анализа. При каких предположениях относительно случайных ошибок эксперимента применяются эти критерии?

8.6. В табл. 8.9 представлены результаты наблюдений над откликом X на пяти уровнях. Проверьте гипотезу H_0 о равенстве средних на уровне значимости $\alpha = 0{,}05$.

О т в е т: гипотезу о равенстве средних следует отвергнуть.

Таблица 8.9

Уровень фактора A	Результаты наблюдений										
1	83	85									
2	84	85	85	86	86	87					
3	86	87	87	87	88	88	88	88	88	89	90
4	89	90	90	91							
5	90	92									

8.7. В трех магазинах, продающих товары одного вида, по данным товарооборота (в условных единицах) за 8 месяцев работы была составлена сводка (табл. 8.10). Проверьте на уровне значимости $\alpha = 0{,}01$ гипотезу H_0 о равенстве средних значений товарооборота для магазинов. Если гипотеза принимается, найдите несмещенные оценки для среднего и дисперсии товарооборота для всех трех магазинов.

Таблица 8.10

Магазин	Месяц							
	1	2	3	4	5	6	7	8
1	19	23	26	18	20	20	18	35
2	20	20	32	27	40	24	22	18
3	16	15	18	26	19	17	19	18

Ответ: гипотезу о равенстве средних значений товарооборота следует принять; $\overline{x} = 22{,}08$, $S_l^2 = \frac{Q_l}{n-l} = 32{,}64$.

8.8. В условиях задачи 8.6 проверьте гипотезы: а) $H_0^{(1)}$: $\mu_1 = \mu_2 = 0$; б) $H_0^{(2)}$: $\mu_4 = \mu_5 = 0$; в) $H_0^{(3)}$: $\mu_3 = \mu_4 = 0$.

Ответ: а) гипотезу следует принять; б) гипотеза отвергается; в) гипотеза отвергается.

8.9. В табл. 8.11 представлены результаты наблюдений над откликом X на пяти уровнях фактора A и трех уровнях фак-

тора B. На уровне значимости $\alpha = 0{,}05$ проверьте гипотезы: а) H_0^A — фактор A не оказывает влияния на отклик; б) H_0^B — фактор B не оказывает влияния на отклик.

Таблица 8.11

Уровни фактора B (j)	Уровни фактора A (i)				
	1	2	3	4	5
1	3	3	6	6	8
2	8	3	7	6	3
3	6	6	8	7	8

Ответ: а) гипотеза принимается; б) гипотеза принимается.

9. НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СТАТИСТИКИ

В предыдущих главах при решении задач математической статистики существенную роль играло предположение о виде (с точностью до параметров) закона распределения наблюдаемой случайной величины X. ***Методы*** математической статистики, основанные на этом предположении, называют ***параметрическими.*** Примерами параметрических методов являются методы нахождения точечных и *интервальных оценок* математического ожидания гауссовской (т.е. распределенной по нормальному закону) случайной величины X по данным *случайной выборки* из ее *генеральной совокупности.*

Однако у параметрических методов имеются существенные недостатки. Во-первых, на практике вид распределения наблюдаемой случайной величины очень часто неизвестен. Во-вторых, *экспериментальные данные* при сборе и обработке информации почти всегда искажаются, что меняет их вид распределения. Поэтому, применяя параметрические методы в условиях такой априорной стохастической неопределенности, необходимо ясно осознавать, что расхождение между *параметрической моделью* и реальной ситуацией может привести (и приводит) подчас к сильно искаженным или даже неверным результатам*.

Следовательно, возникает необходимость в разработке таких статистических процедур, которые, с одной стороны, в ситуации, наиболее благоприятной для параметрических методов, почти не уступали бы им в эффективности, а с другой

*См.: *Хьюбер Дж.П.*; а также: Робастность в статистике. Подход на основе функций влияния / *Хампель Ф.* и др.

стороны, были бы малочувствительны к нарушению предположений, лежащих в основе параметрической модели.

Такие методы существуют. Они получили название ***непараметрических методов***, так как не требуют знания закона распределения наблюдаемой случайной величины и используют лишь минимальную *априорную информацию* типа информации о непрерывности или симметрии функции распределения.

За последнее время непараметрические методы появились почти во всех разделах математической статистики. Они оказывают серьезную конкуренцию классическим процедурам, основанным, главным образом, на предположении о нормальном законе распределения наблюдаемых случайных величин. Причина этого в том, что непараметрический подход лишь незначительно уступает параметрическому по эффективности, если есть уверенность в истинности параметрической модели (например, в том, что наблюдаемая случайная величина имеет нормальный закон распределения). В то же время при нарушении исходных предположений о законе распределения непараметрические модели могут быть во много раз эффективнее параметрических.

Следует отметить, что непараметрические методы с вычислительной точки зрения более трудоемкие, чем параметрические, а иногда и очень сложные. Это сдерживало их применение, хотя многие из них появились еще в 1930–1940-е годы. Однако после появления компьютеров положение изменилось, и теперь во всех наиболее распространенных пакетах прикладных статистических программ реализованы и непараметрические процедуры.

9.1. Одновыборочная задача о сдвиге

Выше (см. 2) рассмотрена задача оценивания математического ожидания случайной величины $X \sim N(\mu, \sigma^2)$ по данным *случайной выборки* $X_1, \ldots, X_n$ из ее *генеральной совокупности*. Что делать, если предположение $X \sim N(\mu, \sigma^2)$ не выполняется?

Например, в примере 4.26 предполагается, что продолжительность времени работы лампы до отказа распределена по нормальному закону. Между тем „время жизни" различных технических устройств обычно описывается не нормальным, а другими распределениями, и прежде всего экспоненциальным*.

Можно привести ряд других примеров, когда возникающие при решении практических задач непрерывные случайные величины не имеют нормального распределения, и, следовательно, методы проверки *статистических гипотез* о математическом ожидании, изложенные в примерах 4.10–4.14, для них не применимы. Все эти задачи можно описать следующей схемой.

Пусть $\varepsilon_1, \ldots, \varepsilon_n$ — последовательность независимых одинаково распределенных с нулевым математическим ожиданием ненаблюдаемых случайных величин, которые можно интерпретировать как *случайные ошибки* наблюдений некоторой неслучайной величины θ. В этом случае простейшая *математическая модель* наблюдений может быть представлена в виде

$$X_i = \theta + \varepsilon_i, \qquad i = \overline{1,n}, \qquad \theta \in \mathbb{R}, \tag{9.1}$$

где случайные величины $X_1, \ldots, X_n$ являются независимыми и имеют один и тот же закон распределения, т.е. их совокупность можно рассматривать как случайную выборку из генеральной совокупности некоторой случайной величины X с математическим ожиданием θ (если оно существует).

Рассмотрим задачу проверки статистической гипотезы

$$H_0\colon\ \theta = \theta_0 \tag{9.2}$$

при одной из *альтернативных гипотез*

$$H_1\colon\ \theta < \theta_0, \qquad H_2\colon\ \theta > \theta_0, \qquad H_3\colon\ \theta \neq \theta_0 \tag{9.3}$$

по данным случайной выборки $X_1, \ldots, X_n$, где θ_0 — некоторое известное значение параметра θ. Предположим, что при различных значениях параметра θ функции распределения $F(x;\theta)$

*См.: *Гнеденко В.Б., Беляев Ю.К., Соловьев А.Д.*

и плотности распределения $p(x;\theta)$ каждого элемента X_i, $i=\overline{1,n}$, случайной выборки отличаются сдвигом на величину θ. Тем самым параметр θ, не изменяя формы графиков функций $F(x;\theta)$ и $p(x;\theta)$, определяет их положение на плоскости (рис. 9.1).

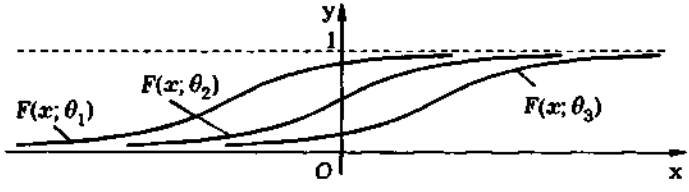


Рис. 9.1

Как правило, θ совпадает с математическим ожиданием случайной величины X_i, а при его отсутствии — с медианой или модой. Поэтому задачу (9.2)–(9.3) называют ***одновыборочной задачей о сдвиге.***

Если независимые случайные величины $\varepsilon_1, \ldots, \varepsilon_n$ распределены по нормальному закону с нулевым математическим ожиданием и неизвестной дисперсией σ^2, то, согласно (9.1), случайные величины $X_1, \ldots, X_n$ также являются независимыми, причем $X_i \sim N(\theta,\sigma^2)$, $i=\overline{1,n}$. Таким образом, закон распределения случайной выборки $X_1, \ldots, X_n$ известен с точностью до параметров, и *метод* проверки статистической гипотезы (9.2) будет *параметрическим* (см. пример 4.14).

Предположим теперь, что о плотности распределения вероятностей независимых одинаково распределенных случайных величин ε_i, $i=\overline{1,n}$, известно лишь то, что она является четной функцией. Оказывается, что даже в такой общей постановке существуют простые методы проверки статистических гипотез о параметре θ и оценивания этого параметра. Остановимся на двух наиболее распространенных из этих методов.

Критерий знаков. Обозначим через $\mathcal{K}_0$ множество функций распределения непрерывных случайных величин, имеющих единственную медиану, которая расположена в точке 0: За-

метим, что функция распределения случайной величины $X \sim N(0,\sigma^2)$ принадлежит множеству $\mathcal{K}_0$, и поэтому предлагаемый ниже *критерий знаков* применим и для решения задач, традиционно решаемых параметрическими методами.

Предположим, что $X_1, \ldots, X_n$ — случайная выборка из генеральной совокупности случайной величины X с функцией распределения $F(x;\theta) \equiv F(x-\theta)$, $F \in \mathcal{K}_0$. Рассмотрим задачу проверки статистической гипотезы H_0 (9.2) при альтернативной гипотезе H_A одного из видов (9.3).

Как и для проверки любой статистической гипотезы (см. 4) гипотезу H_0 естественно отклонить в пользу альтернативной гипотезы H_A, если в результате случайного эксперимента наблюдается некоторое случайное событие, появление которого „практически невозможно" при истинности H_0 и вероятность появления которого „достаточно велика", если верна H_A. Построение *статистического критерия* проверки H_0 при альтернативной гипотезе H_A и заключается в выборе такого события.

Одним из событий, появление которого „практически невозможно" при истинности H_0, является появление очень большого количества чисел одного знака в последовательности $X_1-\theta_0, \ldots, X_n-\theta_0$, или, что то же самое, в последовательности

$$X_{(1)}-\theta_0, \quad \ldots, \quad X_{(n)}-\theta_0, \tag{9.4}$$

где $X_{(1)}, \ldots, X_{(n)}$ — *вариационный ряд* случайной выборки $X_1, \ldots, X_n$.

Действительно, $F(x;\theta)$, как и всякая функция распределения, является неубывающей, а из единственности медианы следует, что в окрестности нуля она строго возрастает. Поэтому при $\theta > \theta_0$

$$\mathbf{P}\{X_i < \theta_0\} = F(\theta_0,\theta) \equiv F(\theta_0-\theta) < \frac{1}{2},$$

а при $\theta < \theta_0$

$$\mathbf{P}\{X_i > \theta_0\} = 1 - F(\theta_0-\theta) < \frac{1}{2}.$$

Отметим, что распределение случайных величин $X_1, \ldots, X_n$ определено функцией $F(x;\theta)$ и зависит от параметра θ. Во избежание путаницы здесь и в дальнейшем вероятность различных событий, порожденных случайной выборкой $X_1, \ldots, X_n$, будем обозначать символом $\mathbf{P}_\theta$, где индекс θ явно указывает на эту зависимость.

Итак, если верна альтернативная гипотеза H_A, то при $H_A = H_1$ большинство чисел последовательности (9.4) должны быть положительными, при $H_A = H_2$ — отрицательными, а при $H_A = H_0$ количество положительных и отрицательных чисел должно быть приблизительно равным, так как в этом случае

$$\mathbf{P}\{X_i > \theta_0\} = \mathbf{P}_\theta\{X_i < \theta_0\} = \frac{1}{2}.$$

Именно это свойство наблюдений и лежит в основе критерия знаков.

Определим случайную величину

$$S(\tau) = \sum_{i=1}^{n} \eta(X_i - \tau), \tag{9.5}$$

где $\eta(t)$ — функция Хевисайда, а $\tau \in \mathbb{R}$ — фиксированный параметр. Случайная величина $S(\tau)$ принимает свои значения $s(\tau)$ на множестве целых чисел в диапазоне от 0 до n. Очевидно, что ее закон распределения зависит и от τ, и от истинного значения параметра θ функции распределения $F(x;\theta)$ случайной величины X. Можно показать, что распределение случайной величины $X_i - \tau$, $i = \overline{1,n}$, зависит только от разности $\theta - \tau$. Поэтому от разности $\theta - \tau$ будет зависеть и распределение случайной величины $S(\tau)$. Следовательно, если θ — истинное значение параметра функции $F(x;\theta)$, то закон распределения *статистики* $S(\theta)$ не зависит от θ.

Обозначим символом s_γ квантиль уровня γ $(0 \leqslant \gamma \leqslant 1)$ распределения случайной величины $S(\theta_0)$ при условии, что верна

гипотеза H_0. Другими словами, s_γ определяется как решение уравнения

$$\mathbf{P}_{\theta_0}\{S(\theta_0) < s_\gamma\} = \gamma, \quad 0 \leqslant \gamma \leqslant 1. \tag{9.6}$$

Заметим, что случайная величина $S(\theta_0)$ дискретна, поэтому решение s_γ уравнения (9.6) для некоторых γ может не существовать.

Статистику $S(\theta_0)$ называют ***статистикой критерия знаков*** для задачи (9.2)–(9.3), а сам ***критерий знаков*** уровня α определяют следующим образом: гипотеза H_0 отклоняется в пользу альтернативной гипотезы H_A (это одна из гипотез (9.3)) на уровне значимости α, если:

а) $s(\theta_0) \geqslant s_{1-\alpha}$ в случае $H_A = H_1$;

б) $s(\theta_0) < s_\alpha$ в случае $H_A = H_2$;

в) $s(\theta_0) < s_{\alpha/2}$ или $s(\theta_0) \geqslant s_{1-\alpha/2}$ в случае $H_A = H_3$.

Смысл критерия знаков прозрачен. Из (9.5) следует, что значение $s(\theta_0)$ статистики $S(\theta_0)$ — это количество положительных чисел в (9.4). Если $s(\theta_0)$ приблизительно равно $n/2$, т.е. количество положительных чисел приблизительно равно количеству отрицательных, то разумно принять H_0. Если же $s(\theta_0)$ близко к n, т.е. почти все числа положительные, то H_0 естественно отклонить в пользу H_1. Малые значения $s(\theta_0)$ говорят о том, что, по-видимому, верна альтернативная гипотеза H_2. И наконец, если статистика $S(\theta_0)$ принимает значения, существенно отличающиеся от $n/2$, H_0 следует отклонить в пользу H_3.

Конечно же, для практического использования критерия необходимо уметь находить квантили s_α, т.е. знать распределение статистики $S(\theta_0)$ при истинности статистической гипотезы H_0. Ответ на этот вопрос дает следующая теорема.

Теорема 9.1. Пусть $X_1, \ldots, X_n$ — независимые одинаково распределенные случайные величины с функцией распределения $F(x;\theta) \equiv F(x-\theta)$, $F \in \mathcal{K}_0$, $\theta \in \mathbb{R}$. Тогда случайная величина

$S(\theta_0)$ имеет *биномиальное распределение* с параметром $p_\theta = 1 - F(\theta_0 - \theta)$:

$$\mathbf{P}_\theta\{S(\theta_0) = k\} = C_n^k p_\theta^k (1 - p_\theta)^{n-k}, \quad k = \overline{0, n}. \tag{9.7}$$

◀ Так как случайные величины $X_1, \ldots, X_n$ независимы, то независимы и случайные величины $\eta(X_1 - \theta_0), \ldots, \eta(X_n - \theta_0)$ как функции независимых случайных величин [XVI]. Кроме того,

$$\mathbf{P}_\theta\{\eta(X_i - \theta_0) = 0\} = \mathbf{P}_\theta\{X_i \leqslant \theta_0\} = F(\theta_0, \theta) \equiv F(\theta_0 - \theta),$$

$$\mathbf{P}\{\eta(X_i - \theta_0) = 1\} = 1 - \mathbf{P}_\theta\{\eta(X_i - \theta_0) = 0\} = 1 - F(\theta_0 - \theta).$$

Таким образом, $S(\theta_0)$ есть сумма независимых случайных величин, каждая из которых имеет биномиальное распределение с параметром $p_\theta = 1 - F(\theta_0 - \theta)$. Следовательно, $S(\theta_0)$ имеет биномиальное распределение с тем же параметром p_θ [XVI]. ▶

Следствие 9.1. При истинности статистической гипотезы H_0 случайная величина $S(\theta_0)$ имеет биномиальное распределение с параметром $p = 1/2$. При этом квантили s_α и $s_{1-\alpha}$ связаны равенством

$$s_\alpha = n - s_{1-\alpha} + 1, \quad 0 < \alpha < 1, \tag{9.8}$$

где n — объем случайной выборки $X_1, \ldots, X_n$ из генеральной совокупности X.

◀ При истинности статистической гипотезы H_0 имеем $\theta = \theta_0$. Поэтому

$$p_{\theta_0} = 1 - F(\theta_0 - \theta_0) = 1 - F(0) = 1 - \frac{1}{2} = \frac{1}{2}.$$

Так как $C_n^k = C_n^{n-k}$, $k = \overline{0, n}$, то для любого $k = \overline{0, n}$

$$\mathbf{P}_\theta\{S(\theta_0) = k\} = C_n^k \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{n-k} =$$

$$= C_n^k 2^{-n} = C_n^{n-k} 2^{-n} = \mathbf{P}_{\theta_0}\{S(\theta_0) = n - k\}.$$

Таким образом,

$$\alpha = \mathbf{P}_{\theta_0}\{S(\theta_0) < s_\alpha\} = \sum_{k=0}^{s_\alpha - 1} C_n^k 2^{-n} = \sum_{k=0}^{s_\alpha - 1} C_n^{n-k} 2^{-n} =$$

$$= \sum_{m=n}^{n-s_\alpha+1} C_n^m 2^{-n} = \mathbf{P}_{\theta_0}\{S(\theta_0) \geqslant n + 1 - s_\alpha\},$$

откуда следует (9.8). ▶

Итак, при истинности статистической гипотезы H_0 закон распределения случайной величины $S(\theta_0)$ не зависит от функции распределения F случайной величины X. Поэтому для практического применения критерия знаков нужны только таблицы биномиального распределения. Именно в этом смысле критерий проверки H_0, основанный на статистике $S(\theta_0)$, называется ***непараметрическим критерием***.

Конечно же, при истинности альтернативной гипотезы, например H_1, распределение случайной величины $S(\theta_0)$ зависит и от F, и от θ — это вытекает из равенства (9.7).

В практических задачах условие одинаковой распределенности случайных величин $X_1, \ldots, X_n$ может нарушаться. Например, если эти величины характеризуют измерения, которые проводились различными приборами и в различных условиях, то случайные величины $\varepsilon_i = X_i - \theta$, $i = \overline{1, n}$, могут иметь уже различные функции распределения $F_i(x)$, хотя по-прежнему из-за отсутствия систематической ошибки измерения $F_i(0) = 1/2$, $i = \overline{1, n}$.

Пример 9.1. Рассмотрим задачу, которая в математической статистике известна как ***задача парных наблюдений***. Пусть для двумерных случайных векторов (Y_i, Z_i), $i = \overline{1, n}$, верно представление

$$Z_i - Y_i = \theta + \varepsilon_i, \quad i = \overline{1, n},$$

где θ — скалярный параметр, а ε_i — независимые одинаково распределенные случайные величины с нулевым математическим ожиданием и непрерывной функцией распределения F. Независимость случайных величин Y_i и Z_i, $i = \overline{1,n}$, не предполагается, более того, на практике они, как правило, зависимы. Требуется проверить гипотезу (9.2) против одной из альтернативных гипотез (9.3). Эта задача сводится к одновыборочной задаче о сдвиге с моделью (9.1), в которой $X_i = Z_i - Y_i$, $i = \overline{1,n}$.

В большинстве приложений Y_i и Z_i — характеристики одного и того же объекта, полученные при различных условиях эксперимента. Например, Y_i и Z_i — артериальное давление у i-го пациента до и после принятия лекарства соответственно, а предположение о неэффективности (бесполезности) лекарства равносильно гипотезе H_0: $\theta = 0$. Если Y_i и Z_i — упругость i-го образца стали при традиционном и модифицированном способах закаливания, то гипотеза H_0: $\theta = 0$ равносильна предположению об одинаковых упругих свойствах стали при обоих способах обработки. #

Критерий знаков можно применять и при различных законах распределения независимых случайных величин ε_i, $i = \overline{1,n}$, так как при истинности статистической гипотезы H_0

$$1 - F(-\theta;\theta) \equiv 1 - F_i(-\theta) = 1 - F_i(0) = 1 - \frac{1}{2} = \frac{1}{2},$$

и следствие 9.1 остается справедливым.

Таблицы биномиального распределения* существуют только для небольших значений n. Если же n велико, то квантили s_α статистики $S(\theta_0)$ можно вычислять, основываясь на интегральной теореме Муавра — Лапласа.

Из этой теоремы следует, что при больших n закон распределения случайной величины

$$\frac{S(\theta_0) - \mathrm{M}S(\theta_0)}{\sqrt{\mathrm{D}S(\theta_0)}}$$

*См.: *Большев Л.Н., Смирнов Н.В.*

хорошо аппроксимируется стандартным нормальным распределением. Это позволяет приближенно вычислять квантили s_α при истинности статистической гипотезы H_0, а именно, так как распределение $S(\theta_0)$ биномиально,

$$\mathbf{M}S(\theta_0) = n\big(1 - F(\theta_0 - \theta)\big), \quad \mathbf{D}S(\theta_0) = n\big(1 - F(\theta_0 - \theta)\big)F(\theta_0 - \theta).$$

Поэтому при $\theta = \theta_0$ имеем $\mathbf{M}S(\theta_0) = n/2$, $\mathbf{D}S(\theta_0) = n/4$, и, как следствие,

$$s_\alpha \approx \frac{n}{2} + u_\alpha \frac{\sqrt{n}}{2},$$

где u_α — квантиль стандартного нормального распределения.

Выше (см. 4) отмечалось, что при фиксированном объеме случайной выборки управлять вероятностями α и β *ошибок первого и второго родов* одновременно невозможно — при построении критерия с меньшей ошибкой первого рода растет ошибка второго рода и наоборот. Однако при $n \to \infty$, т.е. когда объем информации о распределении $F(x - \theta)$ растет, естественным требованием к критерию является безошибочное (в пределе) различение основной и альтернативной гипотез. Это приводит нас к следующему понятию.

Определение 9.1. Статистический ***критерий*** проверки гипотез называют ***состоятельным***, если для любой вероятности α ошибки первого рода, вероятность β ошибки второго рода при $n \to \infty$ стремится к нулю.

Таким образом, критерий знаков проверки гипотезы H_0 против альтернатив (9.3) будет состоятельным, если в случае $\theta \neq \theta_0$ для любого α, $0 < \alpha < 1$,

$$\mathbf{P}_\theta\big\{S(\theta_0) < s_{1-\alpha}\big\} \to 1 \text{ при } n \to \infty,$$

где s_α — квантиль распределения статистики $S(\theta_0)$ с уровнем значимости α, которая определяется формулой (9.6) при $\gamma = 1 - \alpha$.

Теорема 9.2*. Критерий знаков для одновыборочной задачи о сдвиге является состоятельным. #

Перейдем к построению точечных оценок параметра θ функции распределения $F(x-\theta)$. Предположим, что H_0 отклонена, т.е. функция распределения случайной величины X имеет вид $F(x-\theta)$, где $\theta \neq \theta_0$ — неизвестный параметр. Как оценить θ по данным случайной выборки $X_1, \ldots, X_n$ из генеральной совокупности X? В 1963 г. Ходжес и Леман** предложили общий способ построения точечных и интервальных оценок θ, основанный на критериях проверки гипотез о параметре θ. Точечная оценка $\widehat{\theta}(\vec{X}_n)$ параметра θ строится аналогично *оценкам максимального правдоподобия.*

Введем обозначение $L(\vec{X}_n;\theta) = S(\theta)$, подчеркивая, что $S(\theta)$ является функцией и параметра θ, и случайной выборки $\vec{X}_n$ из генеральной совокупности X. Функция $L(\vec{X}_n;\theta)$ при построении оценки параметра θ будет играть роль *функции правдоподобия.* Отметим, что для конкретной реализации $\vec{x}_n$ случайной выборки $\vec{X}_n$ функция $L(\vec{x}_n;\theta) = s(\theta)$ есть функция аргумента θ.

В качестве оценки параметра θ возьмем статистику $\widehat{\theta}(\vec{X}_n)$, значение $\widehat{\theta}$ которой для любой выборки $\vec{x}_n$ удовлетворяет условию

$$L(\vec{x}_n;\widehat{\theta}) = \max_{\theta \in \mathbb{R}} L(\vec{x}_n;\theta).$$

Так как случайная величина $S(\theta)$ распределена по биномиальному закону, то для каждого θ у функции $L(\vec{X}_n;\theta)$ существует одно или два наиболее вероятных значения [XVI]. Поэтому в качестве значения оценки параметра θ нужно взять такое число $\widehat{\theta}$, при котором функция $L(\vec{x}_n;\theta)$ принимает наиболее вероятное значение. Из теоремы 9.1 следует, что если θ — истинное значение параметра, то случайная величина $L(\vec{X}_n;\theta)$ имеет биномиальное распределение с параметром $1/2$. Поэтому наиболее вероятное значение $L(\vec{X}_n;\theta)$ при четном n равно $n/2$, а при

*См.: *Хеттмансперггер Т.*

**См.: *Hodges J.L.*, Jr and *Lehmann E.L.*

нечетном n таких значений сразу два: $(n-1)/2$ и $(n+1)/2$. Возьмем в качестве значения оценки $\widehat{\theta}$ параметра θ решение уравнения

$$L(\vec{x}_n;\theta) = \begin{cases} \dfrac{n}{2}, & n \text{ — четное;} \\ \dfrac{n+1}{2}, & n \text{ — нечетное.} \end{cases} \tag{9.9}$$

Заметим, что

$$S(\theta) = \sum_{i=1}^{n} \eta(X_{(i)} - \theta), \tag{9.10}$$

где $X_{(1)}, \ldots, X_{(n)}$ — вариационный ряд случайной выборки $\vec{X}_n$ (представление $S(\theta)$ в виде (9.10) называют ***считающей формой статистики знаков***). Отсюда следует, что для любой реализации $\vec{x}_n$ случайной выборки $\vec{X}_n$ функция $L(\vec{x}_n;\theta)$, рассматриваемая как функция от θ при фиксированных $x_1, \ldots, x_n$, является невозрастающей кусочно-постоянной ступенчатой функцией с „высотой ступеньки", равной единице. На каждом полуинтервале $[x_{(i)}, x_{(i+1)})$ функция $L(\vec{x}_n;\theta)$ постоянна и равна $n-i$, $i = \overline{1, n-1}$, при $\theta < x_{(1)}$ $L(\vec{X}_n;\theta)$ равна n, а при $\theta \geqslant x_{(n)}$ $L(\vec{X}_n;\theta)$ равна нулю (рис. 9.2).

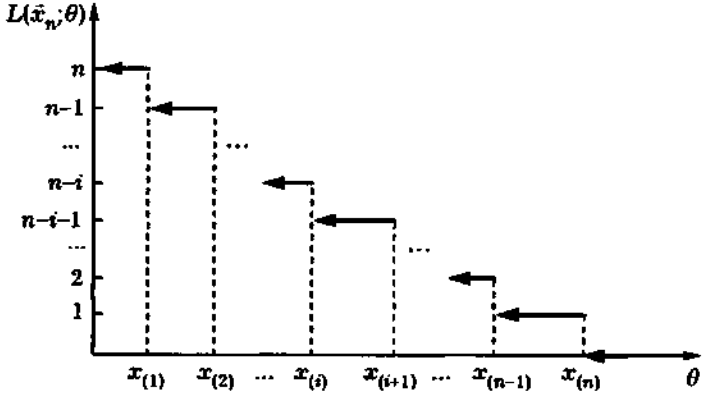


Рис. 9.2

В качестве решения $\widehat{\theta}$ уравнения (9.9) можно брать любое число в полуинтервале $\left[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}\right)$ при четном n и в полуинтервале $\left[x_{\left(\frac{n-1}{2}\right)}, x_{\left(\frac{n+1}{2}\right)}\right)$ — при нечетном. Обычно в качестве $\widehat{\theta}$ берут медиану реализации случайной выборки $\vec{x}_n$, полагая

$$\widehat{\theta} = \begin{cases} \dfrac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & n \text{ — четное;} \\ x_{\left(\frac{n+1}{2}\right)}, & n \text{ — нечетное.} \end{cases} \tag{9.11}$$

Построенную таким образом оценку $\widehat{\theta}(\vec{X}_n)$ называют ***оценкой Ходжеса — Лемана*** параметра θ.

Для построения интервальной оценки параметра θ, входящего в одновыборочную задачу о сдвиге с моделью (9.1), воспользуемся известными результатами (см. **3**) и статистикой $S(\theta)$, определенной равенством (9.10).

Согласно теореме 9.1, при истинности статистической гипотезы H_0: $\theta = \theta_0$ закон распределения статистики $L(\vec{X}_n;\theta) = S(\theta)$ не зависит от параметра θ, и, как следствие (см. **3**), при $\alpha \in (0,1)$

$$\mathbf{P}_\theta\{s_{\alpha/2} \leqslant L(\vec{X}_n;\theta) < s_{1-\alpha/2}\} = \mathbf{P}_{\theta_0}\{s_{\alpha/2} \leqslant L(\vec{X}_n;\theta) < s_{1-\alpha/2}\} = 1-\alpha.$$

Как уже отмечалось, функция $F(\vec{x}_n;\theta)$ при возрастании θ убывает скачками величиной 1 в точках вариационного ряда $x_{(1)}, \ldots, x_{(n)}$. Поэтому для любого $i = \overline{0, n-1}$ неравенство $F(\vec{x}_n;\theta) < n-i$ верно тогда и только тогда, когда $\theta \geqslant x_{(i+1)}$. Аналогично для любого $i = \overline{1, n}$ неравенство $F(\vec{x}_n;\theta) \geqslant i$ верно тогда и только тогда, когда $\theta < x_{(n+1-i)}$. Следовательно, $F(\vec{x}_n;\theta) < s_{1-\alpha/2}$ тогда и только тогда, когда $\theta \geqslant x_{(n+1-s_{1-\alpha/2})}$, а $F(\vec{x}_n;\theta) \geqslant s_{\alpha/2}$ тогда и только тогда, когда $\theta < x_{(n+1-s_{\alpha/2})}$. Так

как (см. следствие 9.1) $n+1-s_{\alpha/2}=s_{1-\alpha/2}$, то нижняя $\underline{\theta}(\vec{X}_n)$ и верхняя $\overline{\theta}(\vec{X}_n)$ границы доверительного интервала уровня значимости α для параметра θ могут быть определены по формулам (рис. 9.3)

$$\underline{\theta}(\vec{X}_n)=X_{(s_{\alpha/2})}, \qquad \overline{\theta}(\vec{X}_n)=X_{(n+1-s_{1-\alpha/2})} \tag{9.12}$$

или по формулам

$$\underline{\theta}(\vec{X}_n)=X_{(n+1-s_{\alpha/2})}, \qquad \overline{\theta}(\vec{X}_n)=X_{(s_{1-\alpha/2})}. \tag{9.13}$$

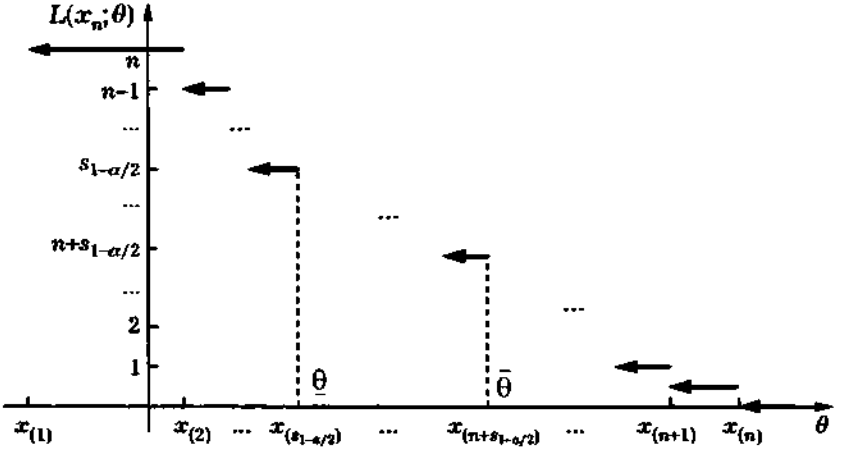


Рис. 9.3

Пример 9.2. По выборке $\vec{x}_n$

5,08; 3,51; 5,78; 4,88; 4,66; 3,94; 4,78; 4,99; 5,33; 5,10;
2,17; 5,32; 4,75; 4,09; 3,98; 3,95; 4,86; 4,89; 5,03; 4,36

объема $n=20$ из генеральной совокупности X проверим на уровне значимости $\alpha=0{,}05$ гипотезу H_0: $\theta=\theta_0=4$ против альтернативной гипотезы H_3: $\theta\neq\theta_0$. Для этого перейдем к

вариационному ряду вида (9.4), построенному для заданной выборки:

−1,83;	−0,49;	−0,06;	−0,05;	−0,02;	0,09;	0,36;
0,66;	0,75;	0,78;	0,86;	0,88;	0,89;	0,99;
1,03;	1,08;	1,10;	1,32;	1,33;	1,78.	

Гипотезу H_0 нужно отклонить на уровне значимости α в пользу H_3, если $s(\theta_0) < s_{\alpha/2}$ или $s(\theta_0) \geqslant s_{1-\alpha/2}$.

Выборочное значение статистики $S(\theta_0) = S(4)$ совпадает с количеством положительных чисел в построенном вариационном ряде вида (9.4) и равно 15. В таблице квантилей биномиального распределения для $n = 20$ и $p_{\theta_0} = p_4 = 0{,}5$ находим

$$\begin{cases} \mathbf{P}_4\{S(4) \geqslant 15\} = 0{,}0207, \\ \mathbf{P}_4\{S(4) \geqslant 14\} = 0{,}0577. \end{cases} \tag{9.14}$$

Отсюда следует, что уравнение

$$\mathbf{P}_4\{S(4) < s_\gamma\} = 0{,}975$$

не имеет решений, т.е. квантили $s_{1-\alpha/2} = s_{0{,}975}$ у статистики $S(4)$ нет. Тем не менее, согласно (9.14), гипотеза H_0 отклоняется в пользу H_3 на уровне значимости

$$\alpha = 2 \cdot 0{,}0207 = 0{,}0414 < 0{,}05.$$

Для оценки неизвестного параметра θ из вариационного ряда находим

$$x_{(10)} = 4 + 0{,}78 = 4{,}78, \qquad x_{(11)} = 4 + 0{,}86 = 4{,}86.$$

Согласно (9.11),

$$\widehat{\theta} = \frac{1}{2}(x_{(10)} + x_{(11)}) = \frac{1}{2}(4{,}78 + 4{,}86) = 4{,}82.$$

Из (9.14) следует, что доверительного интервала для параметра θ уровня доверия $1-0{,}05=0{,}95$ нет. Поэтому построим доверительный интервал уровня доверия $1-\alpha=1-0{,}0414=0{,}9586$. В соответствии с (9.14)

$$s_{1-\alpha/2}=s_{1-0{,}0207}=s_{0{,}9793}=15,$$

$$n+1-s_{1-\alpha/2}=20+1-15=6.$$

Поэтому с вероятностью 0,9586

$$X_{(6)}\leqslant\theta<X_{(15)}.$$

Из вариационного ряда находим

$$x_{(6)}=4+0{,}09=4{,}09,$$

$$x_{(15)}=4+1{,}03=5{,}03.$$

Отсюда следует, что доверительный интервал уровня доверия 0,9586 есть (4,09, 5,03). #

Критерий знаковых рангов Вилкоксона. Определим подмножество $\mathcal{K}_s$ множества $\mathcal{K}_0$, состоящее из всех функций распределения F, соответствующих случайным величинам, плотность которых симметрична относительно нуля, т.е.

$$p(-t)=p(t),\quad t\in\mathbb{R}, \tag{9.15}$$

что равносильно условию

$$F(t)=1-F(-t),\quad t\in\mathbb{R}.$$

Отметим, что функция распределения нормальной случайной величины с нулевым математическим ожиданием принадлежит $\mathcal{K}_s$. Итак,

$$\mathcal{K}_s=\{F\colon F\in\mathcal{K}_0,\ F(t)=1-F(-t)\}.$$

Рассмотрим случайную выборку $\vec{X}_n=(X_1,\ldots,X_n)$. Для произвольного $\tau\in\mathbb{R}$ обозначим через $R_i^{\tau}(\vec{X}_n)$ случайную вели-

чину, представляющую собой *ранг элемента* $|X_i - \tau|$ *случайной выборки*

$$|X_1 - \tau|, \quad \ldots, \quad |X_n - \tau|. \tag{9.16}$$

Определим статистику $T(\tau)$ в соответствии с формулой

$$T(\tau) = \sum_{i=1}^{n} R_i^{\tau}(\vec{X}_n)\eta(X_i - \tau). \tag{9.17}$$

Заметим, что значения $t(\tau)$ статистики $T(\tau)$ — целые числа, наименьшее из которых равно нулю, а наибольшее — $n(n+1)/2$.

Статистику $T(\tau)$ называют ***статистикой знаковых рангов Вилкоксона.***

Можно показать, что если θ — истинное значение параметра θ функции $F(x;\theta) \in \mathcal{K}_s$, то распределение случайной величины $T(\tau)$ зависит от разности $\theta - \tau$, и, следовательно, распределение случайной величины $T(\theta)$ не зависит от θ. В частности, при истинности нулевой гипотезы H_0: $\theta = \theta_0$ распределение $T(\theta_0)$ не зависит от θ_0. Обозначим символом T_γ квантиль уровня γ распределения статистики $T(\theta_0)$ при истинности гипотезы H_0, определяемую из условия

$$\mathbf{P}_{\theta_0}\{T(\theta_0) < T_\gamma\} = \gamma, \quad 0 < \gamma < 1. \tag{9.18}$$

Критерий знаковых рангов Вилкоксона для проверки гипотезы H_0 при альтернативной гипотезе одного из трех возможных видов (9.3) построим следующим образом. Гипотезу H_0 отклоним в пользу гипотезы H_1: $\theta < \theta_0$ на уровне значимости α, $0 < \alpha < 1$, если выборочное значение $t(\theta_0)$ статистики $T(\theta_0)$ удовлетворяет неравенству

$$t(\theta_0) > T_{1-\alpha}.$$

Аналогично гипотезу H_0 отклоним в пользу гипотезы H_2: $\theta > \theta_0$ на уровне значимости α, если

$$t(\theta_0) < T_\alpha.$$

И наконец, H_0 отклоним в пользу двусторонней альтернативной гипотезы H_3: $\theta \neq \theta_0$, если

$$t(\theta_0) < T_{\alpha/2} \quad \text{или} \quad t(\theta_0) > T_{1-\alpha/2}.$$

Это правило мотивируется следующими соображениями. Во-первых, $\mathbf{P}\{X_i - \theta_0 > 0\} > 1/2$ в случае $\theta > \theta_0$, и, во-вторых, чем больше θ, тем больше вероятность $\mathbf{P}\{X_i - \theta_0 > 0\}$, $i = \overline{1,n}$. Поэтому с ростом θ растет вероятность того, что в случайной сумме (9.17) достаточно большим будет и количество ненулевых слагаемых, и каждое слагаемое. Следовательно, большие выборочные значения $t(\theta_0)$ статистики $T(\theta_0)$ должны свидетельствовать об истинности H_1. Аналогичные доводы можно привести для обоснования отклонения критерием гипотезы H_0 в пользу H_2 и H_3.

Заметим, что в отличие от статистики $S(\tau)$ критерия знаков статистика $T(\tau)$ зависит не только от знака каждой разности $X_i - \tau$, но и от ее абсолютной величины, т.е. от расстояния между значениями наблюдения и τ. Эта зависимость как раз и определяется рангом $R_i(\vec{X}_n)$.

Оказывается*, что если θ — истинное значение параметра функции $F(x;\theta)$, то распределение статистики $T(\theta)$ не зависит не только от θ, но и от $F(t)$ и имеет достаточно простой вид. В частности, при истинности H_0, т.е. при $\theta = \theta_0$, распределение $T(\theta_0)$ зависит только от n. При истинности H_0 и небольших n распределение $T(\theta_0)$ при H_0 табулировано. Имеются несложные рекуррентные формулы для вычисления вероятностей** $\mathbf{P}_\theta\{T(\theta) = k\}$, $k = \overline{0, n(n+1)/2}$.

Если n велико, то для T_α существуют приближенные формулы, основанные на аппроксимации распределения статистики $T(\theta)$ нормальным распределением.

*См.: *Хеттмансперггер Т.*

**См. там же.

Известно*, что для любого $t \in \mathbb{R}$ при $n \to \infty$

$$\mathbf{P}\left\{\frac{T(\theta) - \mathbf{M}_\theta T(\theta)}{\sqrt{\mathbf{D}_\theta T(\theta)}} < t\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-s^2/2}\, ds,$$

где $\mathbf{M}_\theta T(\theta)$ и $\mathbf{D}_\theta T(\theta)$ — соответственно математическое ожидание и дисперсия случайной величины $T(\theta)$, вычисленные в предположении, что θ — истинный параметр функции $F(x;\theta)$. Для этих величин верны формулы

$$\mathbf{M}_\theta T(\theta) = \frac{1}{4} n(n+1), \qquad \mathbf{D}_\theta T(\theta) = \frac{1}{24} n(n+1)(2n+1). \quad (9.19)$$

Этот факт позволяет для вычисления квантилей T_α при больших n пользоваться нормальным приближением

$$T_\alpha \approx \frac{1}{4} n(n+1) + u_\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}}, \quad 0 < \alpha < 1, \qquad (9.20)$$

где u_α — квантиль уровня α стандартного нормального распределения.

В основе построения точечных и интервальных оценок параметра θ при $F \in \mathcal{K}_s$ лежат уже рассмотренные идеи Ходжеса и Лемана. Сначала определим $N = n(n+1)/2$ случайных величин $V_1, \ldots, V_N$ вида

$$\frac{1}{2}(X_i + X_j), \quad i, j = \overline{1, n}, \quad i \leqslant j, \qquad (9.21)$$

называемых ***средними Уолша***. Оказывается, что статистике $T(\theta)$ можно придать форму, схожую со статистикой знаков $S(\theta)$, а именно**:

$$T(\theta) = \sum_{i=k}^{N} \eta(V_k - \theta). \qquad (9.22)$$

*См.: *Хеттмансперггер Т.*

**См. там же.

Фактически статистика $T(\theta)$ — это статистика критерия знаков $S(\theta)$ вида (9.5), построенная по случайной „выборке“ $V_1, \ldots, V_N$. Точечную и доверительную оценки параметра θ, основанные на статистике $T(\theta)$, получают по формулам (9.9)–(9.13) с заменой в них n на N, X_k на V_k, $S(\theta)$ на $T(\theta)$. В качестве оценки $\widehat{\theta}(\vec{X}_n)$ параметра θ используют медиану последовательности $V_1, \ldots, V_N$:

$$\widehat{\theta}(\vec{X}_n) = \begin{cases} \frac{1}{2}\left(V_{\left(\frac{N}{2}\right)} + V_{\left(\frac{N}{2}+1\right)}\right), & N \text{ — четное;} \\ V_{\left(\frac{N+1}{2}\right)}, & N \text{ — нечетное.} \end{cases} \tag{9.23}$$

Доверительный интервал уровня доверия $1-2\alpha$, основанный на статистике $T(\theta)$, определяется либо неравенствами

$$V_{(T_{\alpha/2})} \leqslant \theta < V_{(N+1-T_{\alpha/2})}, \tag{9.24}$$

либо неравенствами

$$V_{(N+1-T_{1-\alpha/2})} \leqslant \theta < V_{(T_{1-\alpha/2})}, \tag{9.25}$$

где $V_{(T_{\alpha/2})}$, $V_{(N+1-T_{\alpha/2})}$, $V_{(N+1-T_{1-\alpha/2})}$, $V_{(T_{1-\alpha/2})}$ — элементы вариационного ряда $V_{(1)}, \ldots, V_{(N)}$ с соответствующими номерами, а $T_{\alpha/2}$ и $T_{1-\alpha/2}$ — квантили уровней $\alpha/2$ и $1-\alpha/2$ статистики $T(\theta)$ при истинном значении θ, которые находятся в соответствии с формулой (9.18).

Пример 9.3. Имеется выборка наблюдений

$$-1{,}90; \quad -2{,}53; \quad -0{,}53; \quad -1{,}04; \quad 1{,}98; \quad 1{,}22 \tag{9.26}$$

объема $n=6$. На уровне значимости $\alpha = 0{,}05$ проверим гипотезу H_0: $\theta = \theta_0$, где $\theta_0 = -2$, против альтернативной гипотезы H_1: $\theta < \theta_0$. Для этого перейдем от исходной выборки (9.26) к выборке вида (9.16) с $\tau = -2$:

$$0{,}10; \quad -0{,}53; \quad 1{,}47; \quad 0{,}96; \quad 3{,}98; \quad 3{,}22,$$

имеющей вариационный ряд

$$-0{,}53;\quad 0{,}10;\quad 0{,}96;\quad 1{,}47;\quad 3{,}22;\quad 3{,}98. \tag{9.27}$$

В данном случае последовательность рангов имеет вид

$$2;\quad 1;\quad 4;\quad 3;\quad 6;\quad 5. \tag{9.28}$$

Поэтому, согласно (9.17), $t(-2)=19$. По таблицам находим

$$\begin{cases} \mathbf{P}_{-2}\{T(-2)\geqslant 18\}=0{,}078;\\ \mathbf{P}_{-2}\{T(-2)\geqslant 19\}=0{,}047;\\ \mathbf{P}_{-2}\{T(-2)\geqslant 20\}=0{,}031. \end{cases} \tag{9.29}$$

Видно, что квантили $T_{0,95}$ у распределения случайной величины $T(-2)$ не существует. Тем не менее, как следует из (9.29), H_0 отклоняется в пользу H_1 даже на более низком уровне значимости $\alpha=0{,}047$.

Для нахождения значения $\widehat{\theta}$ оценки $\widehat{\theta}(\vec{X}_n)$ параметра θ вычислим $N=n(n+1)/2=21$ значений $v_1,\ldots,v_N$ средних Уолша (9.21) наблюдений (9.26). Возрастающая последовательность значений средних Уолша будет иметь вид

$$\begin{array}{rrrrrrr} -2{,}53; & -2{,}26; & -2{,}00; & -1{,}78; & -1{,}53; & -1{,}52; & -1{,}26;\\ -1{,}04; & -0{,}78; & -0{,}65; & -0{,}53; & -0{,}39; & -0{,}27; & -0{,}01;\\ 0{,}09; & 0{,}35; & 0{,}47; & 0{,}73; & 1{,}22; & 1{,}60; & 1{,}98. \end{array}$$

Так как число $N=21$ нечетное, медиана последовательности средних Уолша равна $v_{\left(\frac{N+1}{2}\right)}=v_{(11)}=-0{,}53$. Значит,

$$\widehat{\theta}=v_{(11)}=-0{,}53.$$

Построим доверительный интервал уровня доверия $1-\alpha=$ $=1-0{,}062=0{,}938$, где $\alpha/2=0{,}031$. Из (9.29) находим, что

$$T_{1-\alpha/2}=T_{0{,}969}=20,\qquad N+1-T_{1-\alpha/2}=21+1-20=2.$$

Поэтому из (9.25) следует, что с вероятностью 0,938

$$V_{(2)} \leqslant \theta < V_{(20)}.$$

Так как $v_{(2)} = -2{,}26$, $v_{(20)} = 1{,}60$, то доверительный интервал для параметра θ уровня доверия 0,938 есть $(-2{,}26,\ 1{,}60)$.

9.2. Двухвыборочная задача о сдвиге

Пусть $\varepsilon_1, \ldots, \varepsilon_{m+n}$ — независимые одинаково распределенные ненаблюдаемые случайные величины с функцией распределения $F \in \mathcal{K}_0$. Определим наблюдаемые *случайные выборки* $\vec{X}_m = (X_1, \ldots, X_m)$ и $\vec{Y}_n = (Y_1, \ldots, Y_n)$ следующим образом:

$$X_i = \theta_x + \varepsilon_i, \quad i = \overline{1, m}, \quad \theta_x \in \mathbb{R},$$
$$Y_j = \theta_y + \varepsilon_{m+j}, \quad j = \overline{1, n}, \quad \theta_y \in \mathbb{R},$$

где θ_x и θ_y — неизвестные параметры сдвига. Функция распределения случайной величины X_i равна $F(x;\theta_x)$, $i = \overline{1, m}$, а функция распределения случайной величины Y_j равна $F(x;\theta_y)$, $j = \overline{1, n}$. Обычно случайную выборку $X_1, \ldots, X_m$ называют ***контрольной выборкой*** (или ***выборкой из контрольной совокупности***), а $Y_1, \ldots, Y_n$ — ***рабочей*** или ***экспериментальной выборкой.***

Например, $X_1, \ldots, X_m$ могут быть измерениями некоторой характеристики изделия, изготовляемого по традиционной технологии, а $Y_1, \ldots, Y_n$ — по новой экспериментальной. На практике исследователей обычно интересует неизвестный параметр

$$\theta = \theta_y - \theta_x, \tag{9.30}$$

представляющий собой сдвиг в положении, обусловленный переходом на новую технологию.

Задачу проверки *статистической гипотезы* H_0: $\theta = \theta_0$ против одной из альтернативных гипотез H_1: $\theta < \theta_0$, H_2: $\theta > \theta_0$

или H_3: $\theta \neq \theta_0$ называют ***двухвыборочной задачей о сдвиге***. Таким образом, задачи, рассмотренные в примерах 4.25, 4.26, 9.1 а также задачи 4.32, 4.33 являются частными случаями двухвыборочной задачи о сдвиге.

Заметим, что если случайные величины ε_i имеют нормальное распределение, то нормально распределены и случайные величины X_i, $i = \overline{1,m}$, Y_j, $j = \overline{1,n}$. Поэтому решение двухвыборочной задачи о сдвиге может быть получено при помощи критерия Стьюдента (см. пример 4.14).

При решении задач проверки гипотезы H_0 против одной из альтернативных гипотез H_1, H_2, H_3, а также при построении *точечной и интервальной оценок* для θ применяется та же схема, что и в случае *одновыборочной задачи о сдвиге* (см. **9.1**).

Для произвольного $\tau \in \mathbb{R}$ обозначим через $R_j^{\tau}(\vec{X}_m, \vec{Y}_n)$ ранг элемента $Y_j - \tau$, $j = \overline{1,n}$, в объединенной случайной выборке

$$X_1, \quad \ldots, \quad X_m, \quad Y_1 - \tau, \quad \ldots, \quad Y_n - \tau$$

и рассмотрим *статистику*

$$W(\tau) = \sum_{j=1}^{n} R_j^{\tau}(\vec{X}_m, \vec{Y}_n), \tag{9.31}$$

называемую ***статистикой рангов Вилкоксона*** или ***ранговой статистикой Вилкоксона***. Значения $w(\tau)$ случайной величины $W(\tau)$ — целые числа в диапазоне от $n(n+1)/2$ до $m + n(n+1)/2$. Рассуждая так же, как и выше (см. **9.1**), убеждаемся в том, что если $\theta_y - \theta_x = \theta$, то функция распределения случайной величины $W(\tau)$ зависит лишь от разности $\theta - \tau$, и, в частности, распределение случайной величины $W(\theta)$ не зависит от θ. Обозначим через W_γ — квантиль уровня γ распределения $W(\theta)$ при $\theta_y - \theta_x = \theta$, т.е.

$$\mathbf{P}_\theta\{W(\theta) < W_\gamma\} = \gamma, \quad 0 < \gamma < 1.$$

Эмпирическое обоснование ***двухвыборочного критерия Вилкоксона*** для проверки *основной гипотезы* H_0 против одной из альтернативных гипотез H_1, H_2, H_3 состоит в следующем. Чем больше θ в (9.30), тем более вероятно, что значения $y_1, \ldots, y_n$ случайных величин $Y_1, \ldots, Y_n$ превысят значения $x_1, \ldots, x_m$ случайных величин $X_1, \ldots, X_m$. Следовательно, при больших θ ранги $R_j^{\tau}(\vec{X}_m, \vec{Y}_n)$, $j = \overline{1,n}$, а вместе с ними и $W(\theta_0)$ при фиксированном θ_0, имеют тенденцию принимать большие значения. Напротив, при $\theta < \theta_0$ значения случайных величин $Y_1, \ldots, Y_n$ в основном меньше, чем значения случайных величин $X_1, \ldots, X_m$, что приводит к небольшим значениям случайных величин $R_j^{\tau}(\vec{X}_m, \vec{Y}_n)$, $j = \overline{1,n}$, а следовательно, и к небольшим значениям $w(\theta_0)$ статистики $W(\theta_0)$.

После этих наводящих соображений определим двухвыборочный критерий Вилкоксона. При проверке гипотезы H_0 против H_1 на уровне значимости α при помощи двухвыборочного критерия Вилкоксона основную гипотезу H_0 нужно принять, если

$$w(\theta_0) > W_{1-\alpha},$$

и отклонить, если

$$w(\theta_0) < W_{1-\alpha},$$

где $W_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения $W(\theta_0)$ при истинности основной гипотезы H_0.

При проверке H_0 против альтернативной гипотезы H_2 гипотезу H_0 следует принять, если

$$w(\theta_0) > W_{\alpha},$$

и отклонить при

$$w(\theta_0) < W_{\alpha}.$$

При проверке H_0 против альтернативной гипотезы H_3 гипотезу H_0 принимают, если

$$W_{\alpha/2} < w(\theta_0) < W_{1-\alpha/2},$$

и отклоняют в противном случае.

В некоторых справочниках приведены квантили не статистики рангов Вилкоксона (9.31), а квантили ***статистики Манна — Уитни*** $U(\tau)$, которая определяется следующим образом:

$$U(\tau)=\sum_{i=1}^{m}\sum_{j=1}^{n}\eta(Y_j-X_i-\tau)= \\ =\sum_{k=1}^{mn}\eta(V_k-\tau)=\sum_{k=1}^{mn}\eta(V_{(k)}-\tau), \quad (9.32)$$

где η — *функция Хевисайда*, $V_1, V_2, \ldots, V_{mn}$ — последовательность всевозможных разностей вида Y_j-X_i, $i=\overline{1,m}$, $j=\overline{1,n}$, а $V_{(1)}, V_{(2)}, \ldots, V_{(mn)}$ — вариационный ряд „случайной выборки“ $V_1, V_2, \ldots, V_{mn}$.

Можно показать, что статистики $W(\tau)$ и $U(\tau)$ отличаются на неслучайную величину

$$W(\tau)=U(\tau)+\frac{n(n+1)}{2}. \quad (9.33)$$

Поэтому, во-первых, квантили W_γ и U_γ статистик $W(\theta)$ и $U(\theta)$ при $\theta_y-\theta_x=\theta$ связаны равенством

$$W_\gamma=U_\gamma+\frac{n(n+1)}{2}, \quad (9.34)$$

а во-вторых, у статистики $W(\tau)$, так же как у статистик $S(\tau)$ и $T(\tau)$, есть *считающая форма* (9.32).

Если m и n велики, то можно вычислять квантили W_γ по приближенным формулам. Известно*, что если m и n стремятся к бесконечности так, что $m/(m+n)\to\lambda$, $0<\lambda<1$, то для любого $t\in\mathbb{R}$

$$\mathbf{P}\left\{\frac{W(\theta)-\mathbf{M}_\theta W(\theta)}{\sqrt{\mathbf{D}_\theta W(\theta)}}<t\right\}\to\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t}e^{-s^2/2}\,ds, \quad (9.35)$$

*См.: *Хеттмансперггер Т.*

где математическое ожидание $\mathbf{M}_\theta W(\theta)$ и дисперсия $\mathbf{D}_\theta W(\theta)$ статистики $W(\theta)$ определяются по формулам

$$\mathbf{M}_\theta W(\theta) = \frac{n(m+n+1)}{2}, \quad \mathbf{D}_\theta W(\theta) = \frac{mn(m+n+1)}{12}. \tag{9.36}$$

Поэтому

$$W_\gamma \approx \frac{n(m+n+1)}{2} + u_\gamma \sqrt{\frac{mn(m+n+1)}{12}}, \tag{9.37}$$

где u_γ — квантиль уровня γ стандартного нормального распределения.

Так же как и при построении точечных оценок в одновыборочной задаче (см. **9.1**), значение $\widehat{\theta}$ *оценки Ходжеса — Лемана* $\widehat{\theta}(\vec{X}_m, \vec{Y}_n)$ параметра $\theta = \theta_y - \theta_x$ в двухвыборочной задаче определяется как такое число θ, при котором для выборок $x_1, \ldots, x_m, y_1, \ldots, y_n$ достигается максимум значений $w(\theta)$ статистики $W(\theta)$ или, что то же самое, значений $u(\theta)$ статистики $U(\theta)$. Рассуждения, аналогичные рассуждению при построении оценок Ходжеса — Лемана в одновыборочной задаче, приводят к тому, что $\widehat{\theta}(\vec{X}_m, \vec{Y}_n)$ — медиана вариационного ряда $V_{(1)}, V_{(2)}, \ldots, V_{(mn)}$:

$$\widehat{\theta}(\vec{X}_m, \vec{Y}_n) = \begin{cases} \frac{1}{2}\left(V_{\left(\frac{mn}{2}\right)} + V_{\left(\frac{mn}{2}+1\right)}\right), & mn \text{ — четное;} \\ V_{\left(\frac{mn+1}{2}\right)}, & mn \text{ — нечетное.} \end{cases} \tag{9.38}$$

При построении интервальной оценки для параметра θ в двухвыборочной задаче также сохраняется схема, использовавшаяся в одновыборочной задаче (см. **9.1**). Для статистики Манна — Уитни

$$\mathbf{P}_{\theta_0}\{U_{\alpha/2} \leqslant U(\theta_0) < U_{1-\alpha/2}\} = 1 - \alpha.$$

Из определения $U(\tau)$ в (9.32) следует, что $w(\tau)$ является невозрастающей кусочно-постоянной функцией от τ, убывающей

скачками в точках $v_{(k)}$, $k=\overline{1, mn}$, и равной $mn-i$ на полуинтервале $[v_{(i)}, v_{(i+1)})$, $i=\overline{1, mn-1}$, где v_i — значение V_i, $i=\overline{1, mn}$. Поэтому доверительный интервал уровня доверия $1-\alpha$ определяется либо неравенствами

$$V_{(U_{\alpha/2})} \leqslant \theta < V_{(mn+1-U_{\alpha/2})}, \tag{9.39}$$

либо неравенствами

$$V_{(mn+1-U_{1-\alpha/2})} \leqslant \theta < V_{(U_{1-\alpha/2})}. \tag{9.40}$$

Используя в неравенствах (9.39) и (9.40) квантили статистики Манна — Уитни, которые выражаются через квантили статистики Вилкоксона по формуле (9.34), получим еще два представления доверительного интервала:

$$V_{(W_{\alpha/2}-\frac{n(n+1)}{2})} \leqslant \theta < V_{(mn+\frac{n(n+1)}{2}-W_{\alpha/2})}, \tag{9.41}$$

$$V_{(mn+\frac{n(n+1)}{2}-W_{1-\alpha/2})} \leqslant \theta < V_{(W_{1-\alpha/2}-\frac{n(n+1)}{2})}. \tag{9.42}$$

Пример 9.4. Рассмотрим выборку объема $m=6$

$$3{,}9; \quad 4{,}3; \quad 4{,}4; \quad 4{,}6; \quad 4{,}9; \quad 5{,}8$$

из генеральной совокупности X и выборку объема $n=5$

$$7{,}7; \quad 8{,}1; \quad 8{,}3; \quad 8{,}6; \quad 8{,}9$$

из генеральной совокупности Y. Предположим, что функции распределения генеральных совокупностей X и Y отличаются лишь сдвигом на неизвестную величину $\theta \in \mathbb{R}$. Проверим на уровне значимости $\alpha=0{,}05$ гипотезу H_0: $\theta=\theta_0$ при $\theta_0=3$ против альтернативной гипотезы H_2: $\theta<\theta_0$.

Объединим обе выборки и построим вариационный ряд объединенной выборки, предварительно вычтя из всех элементов

второй выборки $\theta_0 = 3$:

3,9; 4,3; 4,4; 4,6; 4,7; 4,9; 5,1; 5,3; 5,6; 5,8; 5,9.

Из этого вариационного ряда находим последовательность значений рангов $R_1^3(\vec{X}_m, \vec{Y}_n), \ldots, R_5^3(\vec{X}_m, \vec{Y}_n)$ элементов второй выборки в объединенной выборке:

5; 7; 8; 9; 11.

Затем по формуле (9.31) получаем $w(\theta_0) = w(3) = 40$, а по таблицам распределения статистики $W(3)$ при $m = 6$, $n = 5$ находим

$$\begin{cases} \mathrm{P}_3\{W(3) \geqslant 39\} = 0{,}063, \\ \mathrm{P}_3\{W(3) \geqslant 40\} = 0{,}041, \\ \mathrm{P}_3\{W(3) \geqslant 41\} = 0{,}026. \end{cases} \tag{9.43}$$

Таким образом, квантили $W_{0,95}$ при $m = 6$, $n = 5$ не существует. Из (9.43) видно, что H_0 отклоняется на уровне значимости $\alpha = 0{,}041$ в пользу H_1.

Чтобы найти значение $\widehat{\theta}$ оценки $\widehat{\theta}(\vec{X}_m, \vec{Y}_n)$ Ходжеса — Лемана для параметра сдвига θ, рассмотрим вариационный ряд $V_{(1)}, V_{(2)}, \ldots, V_{(mn)}$ для последовательности разностей $Y_j - X_i$, который в данном случае имеет вид

−2,00; −1,70; −1,60; −1,50; −1,40; −1,30; −1,30; −1,20;
−1,20; −1,00; −1,00; −1,00; −0,90; −0,80; −0,80; −0,70;
−0,70; −0,70; −0,50; −0,40; −0,40; −0,30; −0,20; −0,10;
−0,10; 0,20; 0,20; 0,50; 0,70; 1,10.

Так как $mn = 6 \cdot 5 = 30$, то выборочная медиана вариационного ряда $V_{(1)}, V_{(2)}, \ldots, V_{(mn)}$ есть

$$\widehat{\theta} = \frac{1}{2}\left(v_{(15)} + v_{(16)}\right) = \frac{-0{,}7 - 0{,}8}{2} = -0{,}75.$$

При построении доверительного интервала для θ уровня доверия $1-\alpha=0{,}95$ нужно найти квантиль $W_{\alpha/2}$ или $W_{1-\alpha/2}$, где $\alpha=0{,}025$. Из (9.43) видно, что нельзя построить доверительный интервал при $\alpha/2=0{,}025$, но можно при $\alpha/2=0{,}026$. Так как $W_{1-0{,}026}=W_{0{,}974}=41$, то из (9.43) получаем

$$mn+1+\frac{n(n+1)}{2}-W_{1-\alpha/2}=30+1+15-41=5,$$

$$W_{1-\alpha/2}-\frac{n(n+1)}{2}=41-15=26.$$

Поэтому используя вариационный ряд $V_{(1)}$, $V_{(2)}$, ..., $V_{(mn)}$, находим

$$v_{(5)}=-1{,}4, \quad v_{(26)}=0{,}2.$$

Отсюда и из (9.42) вытекает, что с вероятностью 0,949

$$V_{(5)}\leqslant\theta\leqslant V_{(26)},$$

т.е. доверительный интервал параметра θ с уровнем доверия 0,949 есть $(-1{,}4,\ 0{,}2)$.

9.3. Решение типовых примеров

Пример 9.5. Для определения предела текучести некоторой марки стали по просьбе заказчика, которому была необходима сталь с пределом текучести в $30\frac{\text{кгс}}{\text{мм}^2}$, были проведены стандартные испытания $n=25$ образцов. Результаты испытаний (в $\frac{\text{кгс}}{\text{мм}^2}$) следующие:

32,00;	30,69;	35,68;	34,41;	41,95;	40,05;	32,63;
32,77;	30,41;	28,84;	29,70;	28,61;	34,39;	35,48;
29,97;	34,80;	30,45;	30,36;	34,66;	30,71;	33,19;
29,49;	29,60;	28,43;	29,29.			

Выясним, удовлетворяет ли данная марка стали требованиям заказчика, и оценим по результатам экспериментов истинный предел ее текучести. Другими словами, проверим по заданной выборке *основную гипотезу* H_0 вида (9.2) о том, что $\theta_0 = 30$, против *альтернативной гипотезы* H_3 о том, что $\theta_0 \neq 30$. Гипотезу H_0 будем проверять при помощи *критерия знаков*.

Вариационный ряд рассматриваемой выборки имеет вид

28,43;	28,61;	28,84;	29,29;	29,49;	29,60;	29,70;
29,97;	30,36;	30,41;	30,45;	30,69;	30,71;	32,00;
32,63;	32,77;	33,19;	34,39;	34,41;	34,66;	34,80;
35,48;	35,68;	40,05;	41,95.			

Выборочное значение статистики $S(30)$, вычисленное в соответствии с формулой (9.5) по заданной выборке, равно 17. По таблице распределения $S(30)$ для $n = 25$ находим

$$\mathbf{P}_{30}\{S(30) \geqslant 17\} = 0{,}0539, \qquad \mathbf{P}_{30}\{S(30) \geqslant 18\} = 0{,}0216.$$

Поэтому гипотеза H_0 отклоняется на уровне значимости $\alpha < 0{,}0539$ и принимается на уровне значимости $\alpha \geqslant 0{,}0539$.

Неизвестный параметр θ оценим медианой вариационного ряда рассматриваемой выборки. Эта медиана равна 30,71.

Построим *доверительный интервал уровня доверия* $1 - \alpha$, где $\alpha/2 = 0{,}0216$. Так как $s_{1-\alpha/2} = 18$, то с вероятностью $1 - \alpha = 0{,}9568$

$$X_{(8)} \leqslant \theta < X_{(18)}.$$

Поскольку $X_{(8)} = 29{,}97$, $X_{(18)} = 34{,}39$, то с вероятностью 0,9568

$$29{,}97 \leqslant \theta < 34{,}39.$$

Пример 9.6. Решим предыдущую задачу (см. пример 9.5) при помощи *критерия знаковых рангов Вилкоксона*.

Упорядоченный массив *средних Уолша* выборки из примера 9.5 состоит из $N = 325$ чисел и имеет вид

28,43; 28,52; 28,61; 28,64; 28,72; 28,84; 28,86; 28,95; 28,96; 29,02;
29,05; 29,06; 29,06; 29,11; 29,15; 29,17; 29,20; 29,22; 29,27; 29,29;
29,29; 29,39; 29,39; 29,40; 29,42; 29,44; 29,45; 29,48; 29,49; 29,49;
29,51; 29,53; 29,55; 29,56; 29,57; 29,60; 29,60; 29,60; 29,62; 29,63;
29,64; 29,65; 29,65; 29,66; 29,70; 29,73; 29,76; 29,77; 29,79; 29,82;
29,83; 29,85; 29,87; 29,93; 29,95; 29,97; 29,97; 29,98; 29,99; 30,00;
30,01; 30,03; 30,03; 30,05; 30,07; 30,09; 30,10; 30,15; 30,16; 30,16;
30,19; 30,19; 30,20; 30,21; 30,22; 30,30; 30,33; 30,34; 30,36; 30,38;
30,40; 30,41; 30,42; 30,43; 30,45; 30,52; 30,53; 30,53; 30,55; 30,56;
30,57; 30,58; 30,60; 30,62; 30,64; 30,69; 30,69; 30,70; 30,71; 30,73;
30,75; 30,80; 30,81; 30,81; 30,85; 30,90; 30,96; 30,98; 31,02; 31,03;
31,06; 31,12; 31,13; 31,16; 31,18; 31,19; 31,20; 31,22; 31,23; 31,24;
31,30; 31,34; 31,34; 31,35; 31,37; 31,40; 31,41; 31,42; 31,44; 31,49;
31,50; 31,51; 31,52; 31,54; 31,55; 31,56; 31,58; 31,59; 31,61; 31,61;
31,62; 31,63; 31,64; 31,66; 31,67; 31,70; 31,73; 31,74; 31,75; 31,77;
31,80; 31,82; 31,82; 31,84; 31,85; 31,94; 31,94; 31,95; 31,95; 31,96;
31,97; 32,00; 32,00; 32,01; 32,04; 32,04; 32,05; 32,05; 32,06; 32,08;
32,13; 32,14; 32,15; 32,16; 32,18; 32,18; 32,19; 32,20; 32,25; 32,26;
32,31; 32,31; 32,38; 32,38; 32,38; 32,38; 32,39; 32,40; 32,41; 32,42;
32,43; 32,48; 32,49; 32,51; 32,53; 32,54; 32,54; 32,55; 32,55; 32,56;
32,56; 32,58; 32,59; 32,59; 32,60; 32,60; 32,62; 32,63; 32,64; 32,67;
32,68; 32,69; 32,70; 32,72; 32,74; 32,75; 32,77; 32,82; 32,91; 32,92;
32,94; 32,96; 32,98; 33,02; 33,04; 33,06; 33,08; 33,09; 33,18; 33,19;
33,19; 33,20; 33,20; 33,33; 33,40; 33,51; 33,52; 33,58; 33,59; 33,64;
33,71; 33,72; 33,74; 33,78; 33,79; 33,80; 33,84; 33,93; 33,99; 34,05;
34,13; 34,15; 34,23; 34,24; 34,33; 34,33; 34,39; 34,40; 34,41; 34,44;
34,45; 34,53; 34,54; 34,60; 34,60; 34,66; 34,67; 34,73; 34,77; 34,80;
34,83; 34,87; 34,94; 34,94; 35,01; 35,04; 35,05; 35,07; 35,14; 35,17;
35,19; 35,21; 35,23; 35,24; 35,25; 35,28; 35,37; 35,38; 35,39; 35,48;
35,58; 35,62; 35,68; 35,72; 35,78; 35,82; 35,96; 36,03; 36,15; 36,18;
36,20; 36,32; 36,33; 36,34; 36,41; 36,62; 36,97; 37,22; 37,23; 37,29;
37,36; 37,36; 37,43; 37,57; 37,77; 37,87; 38,17; 38,18; 38,30; 38,37;
38,71; 38,81; 40,05; 41,00; 41,95.

Проверим гипотезу H_0 против альтернативной гипотезы H_3 на *уровне значимости* $\alpha = 0{,}05$. Значение статистики $T(30)$,

совпадающее с числом членов вариационного ряда выборки, превышающих 30, будет равно 265. По формулам (9.19) находим

$$\mathbf{M}_{30}T(30) = \frac{25\cdot 26}{4} = 162{,}5, \qquad \mathbf{D}_{30}T(30) = \frac{25\cdot 26\cdot 51}{24} = 1381{,}25,$$

$$\sqrt{\mathbf{D}_{30}T(30)} \approx 37{,}165.$$

Поэтому, согласно (9.20),

$$T_\alpha \approx 162{,}5 + u_\alpha \cdot 37{,}165.$$

Для $\alpha = 0{,}05$ по таблице квантилей стандартного нормального распределения (см. табл. П.2) находим $u_{\alpha/2} = 1{,}96$. Поэтому $T_{0{,}975} \approx 235{,}34$. Так как $265 > 235{,}34$, то гипотеза H_0 отклоняется. Параметр θ оценивается медианой упорядоченного массива средних Уолша, которая есть 163-й элемент массива и равна 32,00.

Построим доверительный интервал уровня доверия $\alpha = 0{,}95$. Так как $T_{1-\alpha/2} = T_{0{,}975} = 235{,}34$, то, согласно формуле (9.25), нижняя и верхняя границы этого интервала есть 90-й и 236-й элементы упорядоченного массива средних Уолша. Поэтому с вероятностью 0,95

$$30{,}56 \leqslant \theta < 33{,}51.$$

Пример 9.7. Даны выборка

0,00;	−0,53;	1,47;	0,96;	3,98;	3,22;	0,25;
0,31;	−0,64;	−1,26;	−0,92;	−1,36;	0,96;	1,39;
−0,81;	1,12;	−0,62;	−0,66;	1,07;	−0,52;	0,48;
−1,00;	−0,96;	−1,43;	−1,09			

объема $m = 25$ из распределения Коши с плотностью

$$p_X(x) = \frac{1}{\pi(1+x^2)},$$

и выборка

−0,88;	0,41;	−0,64;	−0,81;	−0,09;	−0,71;	−0,00;
0,49;	−0,65;	0,59;	0,17;	−0,46;	0,99;	−0,24;
−0,98;	−0,85;	−0,09;	−0,63;	0,68;	0,02;	−0,59;
−0,02;	−0,45;	−0,50;	0,40;	0,29;	−0,17;	−0,43

объема $n = 28$ из равномерного распределения на отрезке $[-1, 1]$ с плотностью $p_Y(x)$. Проверим при помощи *критерия Смирнова* гипотезу о равенстве функций p_X и p_Y.

Вариационный ряд объединенной выборки имеет вид

−1,43;	−1,36;	−1,26;	−1,09;	−1,00;	−0,98;	−0,96;	−0,92;
−0,88;	−0,85;	−0,81;	−0,81;	−0,71;	−0,66;	−0,65;	−0,64;
−0,64;	−0,63;	−0,62;	−0,59;	−0,53;	−0,52;	−0,50;	−0,46;
−0,45;	−0,43;	−0,24;	−0,17;	−0,09;	−0,09;	−0,02;	−0,00;
0,00;	0,02;	0,17;	0,25;	0,29;	0,31;	0,40;	0,41;
0,48;	0,49;	0,59;	0,68;	0,96;	0,96;	0,99;	1,07;
1,12;	1,39;	1,47;	3,22;	3,98.			

Соответствующие им величины δ_i, $i = \overline{1, N}$, вычисленные по формуле (5.15), таковы:

0; 0; 0; 0; 0; 1; 0; 0; 1; 1; 1; 0; 1; 0; 1; 1; 0; 1; 0; 1; 0; 0; 1; 1;
1; 1; 1; 1; 1; 1; 1; 0; 1; 1; 0; 1; 0; 1; 1; 0; 1; 1; 1; 0; 0; 1; 0;
0; 0; 0; 0; 0.

Поэтому

$$D = 0{,}473, \qquad \sqrt{\frac{mn}{m+n}} = 1{,}718.$$

Так как m и n велики, то для проверки гипотезы H_0 воспользуемся асимптотической формулой (5.13), в соответствии с которой

$$\mathbf{P}\left\{\sqrt{\frac{25 \cdot 28}{25+28}} > 1{,}718\right\} \approx 0{,}004.$$

Поэтому гипотезу H_0 следует отклонить на уровне значимости $\alpha \geqslant 0{,}004$.

Проверим эту же гипотезу с помощью *двухвыборочного критерия Вилкоксона*. Вычисляя по формулам (9.32)–(9.33) при $\theta_0 = 0$ реализацию $w(0)$ случайной величины $W(0)$, получим $w(0) = 754$. Так как m и n велики, то для нахождения квантилей распределения *статистики рангов Вилкоксона* воспользуемся приближенной формулой (9.37), выражающей их через квантили стандартного нормального распределения. Имеем

$$\frac{w(0) - \mathrm{M}_0 W(0)}{\sqrt{\mathrm{D}_0 W(0)}} = \frac{754 - 756}{\sqrt{3150}} \approx -0{,}036.$$

По таблицам квантилей стандартного нормального распределения находим

$$\mathrm{P}\left\{\left|\frac{W(0) - \mathrm{M}_0 W(0)}{\sqrt{\mathrm{D}_0 W(0)}}\right| > 0{,}036\right\} = 0{,}48.$$

Поэтому двухвыборочный критерий Вилкоксона гипотезу об однородности не отклоняет. Это произошло из-за того, что медианы обоих распределений совпадают (равны нулю), а не сдвинуты относительно друг друга.

Вопросы и задачи

9.1. Какие методы математической статистики называются непараметрическими?

9.2. В чем преимущества и недостатки непараметрических методов по сравнению с классическими?

9.3. Дайте определение одновыборочной задачи о сдвиге.

9.4. В какой ситуации лучше всего применять методы, основанные на статистике критерия знаков?

9.5. В какой ситуации лучше всего применять методы, основанные на статистике критерия знаковых рангов Вилкоксона?

9.6. Таблицы какого распределения достаточно иметь для решения одновыборочной задачи при помощи критерия знаков?

9.7. Какой критерий называется состоятельным?

9.8. Дайте определение ранга элемента числовой последовательности.

9.9. Какая задача называется двухвыборочной задачей о сдвиге?

9.10. Можно ли применять критерии знаков и знаковых рангов Вилкоксона для проверки гипотез о математическом ожидании нормального распределения?

9.11. Что называется средними Уолша?

9.12. Какие критерии называются критериями согласия?

9.13. Можно ли гипотезу о параметре сдвига в двухвыборочной задаче проверять не ранговым критерием Вилкоксона, а критерием Смирнова?

9.14. Являются ли состоятельными критерии Колмогорова, ω^2 и Смирнова?

9.15. В каких случаях в двухвыборочной задаче лучше применять критерий Колмогорова, критерий ω^2 и двухвыборочный критерий Вилкоксона?

9.16. Докажите теорему 9.2 о состоятельности критерия знаков.

Указание: при помощи центральной предельной теоремы аппроксимировать квантили статистики критерия знаков квантилями нормального распределения.

9.17. Докажите, что определения статистики $T(\tau)$ по формулам (9.17) и (9.22) равносильны.

9.18. Найдите математическое ожидание и дисперсию статистики $T(\tau)$ знаковых рангов Вилкоксона (см. формулу (9.19)).

9.19. Докажите формулы (9.36).

9.20. Докажите асимптотическую нормальность статистики $W(\tau)$.

9.21. Для проверки влияния нейтронного облучения на деформируемость меди были проведены эксперименты на растяжение двух партий образцов. В первой необлученной (контрольной) партии из 13 образцов результаты экспериментов при деформации 0,5 оказались следующими:

6,01; 6,23; 5,75; 6,17; 5,97; 6,22; 6,19;
5,94; 6,01; 5,87; 6,23; 5,78; 5,99.

Вторая партия из 13 образцов после облучения потоком нейтронов интенсивностью $2 \cdot 10^{18}$ нейтрон/см^2 при той же деформации 0,5 привела к следующим результатам:

5,75; 5,86; 6,13; 6,18; 5,63; 5,74; 5,97;
5,49; 6,22; 5,79; 6,32; 5,45; 6,03.

Изменяется ли прочность меди после облучения?

9.22. Для упрочнения алюминиевых изделий используется операция нагартовки (наклепа), заключающаяся в пластической деформации. Семь образцов алюминия были подвержены 2%-ной нагартовке, а десять образцов — 5%-ной. Прочность (в $\frac{\text{кг}}{\text{мм}^2}$) образцов первой партии составила

17; 18; 16; 19; 15; 20; 14,

а второй

21; 22; 20; 23; 19; 24; 18; 21,5; 20,6.

Можно ли на основании этих данных сделать вывод об увеличении прочности алюминия при увеличении пластической деформации?

ПРИЛОЖЕНИЕ

Таблица П.1

Функция распределения $\Phi(x)$

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Таблица **П.2**

Квантили нормального распределения

P	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
q	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Таблица **П.3**

Квантили распределения χ^2

	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999
1	3,9E-05	0,0002	0,001	0,004	0,02	0,1	0,1	1,1	1,6	2,7	3,8	5,0	6,6	7,9	10,8
2	0,010	0,02	0,051	0,10	0,21	0,45	0,71	2,41	3,22	4,61	5,99	7,38	9,21	10,60	13,82
3	0,072	0,11	0,22	0,35	0,58	1,01	1,42	3,66	4,64	6,25	7,81	9,35	11,34	12,84	16,27
4	0,21	0,30	0,48	0,71	1,06	1,65	2,19	4,88	5,99	7,78	9,49	11,14	13,28	14,86	18,47
5	0,41	0,55	0,83	1,15	1,61	2,34	3,00	6,06	7,29	9,24	11,07	12,83	15,09	16,75	20,52
6	0,68	0,87	1,24	1,64	2,20	3,07	3,83	7,23	8,56	10,64	12,59	14,45	16,81	18,55	22,46
7	0,99	1,24	1,69	2,17	2,83	3,82	4,67	8,38	9,80	12,02	14,07	16,01	18,48	20,28	24,32
8	1,34	1,65	2,18	2,73	3,49	4,59	5,53	9,52	11,03	13,36	15,51	17,53	20,09	21,95	26,12
9	1,73	2,09	2,70	3,33	4,17	5,38	6,39	10,66	12,24	14,68	16,92	19,02	21,67	23,59	27,88
10	2,16	2,56	3,25	3,94	4,87	6,18	7,27	11,78	13,44	15,99	18,31	20,48	23,21	25,19	29,59
11	2,60	3,05	3,82	4,57	5,58	6,99	8,15	12,90	14,63	17,28	19,68	21,92	24,72	26,76	31,26
12	3,07	3,57	4,40	5,23	6,30	7,81	9,03	14,01	15,81	18,55	21,03	23,34	26,22	28,30	32,91
13	3,57	4,11	5,01	5,89	7,04	8,63	9,93	15,12	16,98	19,81	22,36	24,74	27,69	29,82	34,53
14	4,07	4,66	5,63	6,57	7,79	9,47	10,82	16,22	18,15	21,06	23,68	26,12	29,14	31,32	36,12
15	4,60	5,23	6,26	7,26	8,55	10,31	11,72	17,32	19,31	22,31	25,00	27,49	30,58	32,80	37,70
16	5,14	5,81	6,91	7,96	9,31	11,15	12,62	18,42	20,47	23,54	26,30	28,85	32,00	34,27	39,25

Окончание табл. **П.3**

	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999
17	5,70	6,41	7,56	8,67	10,09	12,00	13,53	19,51	21,61	24,77	27,59	30,19	33,41	35,72	40,79
18	6,26	7,01	8,23	9,39	10,86	12,86	14,44	20,60	22,76	25,99	28,87	31,53	34,81	37,16	42,31
19	6,84	7,63	8,91	10,12	11,65	13,72	15,35	21,69	23,90	27,20	30,14	32,85	36,19	38,58	43,82
20	7,43	8,26	9,59	10,85	12,44	14,58	16,27	22,77	25,04	28,41	31,41	34,17	37,57	40,00	45,31
21	8,03	8,90	10,28	11,59	13,24	15,44	17,18	23,86	26,17	29,62	32,67	35,48	38,93	41,40	46,80
22	8,64	9,54	10,98	12,34	14,04	16,31	18,10	24,94	27,30	30,81	33,92	36,78	40,29	42,80	48,27
23	9,26	10,20	11,69	13,09	14,85	17,19	19,02	26,02	28,43	32,01	35,17	38,08	41,64	44,18	49,73
24	9,89	10,86	12,40	13,85	15,66	18,06	19,94	27,10	29,55	33,20	36,42	39,36	42,98	45,56	51,18
25	10,52	11,52	13,12	14,61	16,47	18,94	20,87	28,17	30,68	34,38	37,65	40,65	44,31	46,93	52,62
26	11,16	12,20	13,84	15,38	17,29	19,82	21,79	29,25	31,79	35,56	38,89	41,92	45,64	48,29	54,05
27	11,81	12,88	14,57	16,15	18,11	20,70	22,72	30,32	32,91	36,74	40,11	43,19	46,96	49,64	55,48
28	12,46	13,56	15,31	16,93	18,94	21,59	23,65	31,39	34,03	37,92	41,34	44,46	48,28	50,99	56,89
29	13,12	14,26	16,05	17,71	19,77	22,48	24,58	32,46	35,14	39,09	42,56	45,72	49,59	52,34	58,30
30	13,79	14,95	16,79	18,49	20,60	23,36	25,51	33,53	36,25	40,26	43,77	46,98	50,89	53,67	59,70
35	17,19	18,51	20,57	22,47	24,80	27,84	30,18	38,86	41,78	46,06	49,80	53,20	57,34	60,27	66,62
40	20,71	22,16	24,43	26,51	29,05	32,34	34,87	44,16	47,27	51,81	55,76	59,34	63,69	66,77	73,40
45	24,31	25,90	28,37	30,61	33,35	36,88	39,58	49,45	52,73	57,51	61,66	65,41	69,96	73,17	80,08
50	27,99	29,71	32,36	34,76	37,69	41,45	44,31	54,72	58,16	63,17	67,50	71,42	76,15	79,49	86,66
75	47,21	49,48	52,94	56,05	59,79	64,55	68,13	80,91	85,07	91,06	96,22	100,84	106,39	110,29	118,60
100	67,33	70,06	74,22	77,93	82,36	87,95	92,13	106,91	111,67	118,50	124,34	129,56	135,81	140,17	149,45

Таблица **П.4**

Квантили распределения Стьюдента

	0,75	0,9	0,95	0,975	0,99	0,995	0,999
1	1,000	3,078	6,314	12,706	31,821	63,657	318,309
2	0,816	1,886	2,920	4,303	6,965	9,925	22,327
3	0,765	1,638	2,353	3,182	4,541	5,841	10,215
4	0,741	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,476	2,015	2,571	3,365	4,032	5,893
6	0,718	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,383	1,833	2,262	2,821	3,250	4,297
10	0,700	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,363	1,796	2,201	2,718	3,106	4,025
12	0,695	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,350	1,771	2,160	2,650	3,012	3,852
14	0,692	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,341	1,753	2,131	2,602	2,947	3,733
16	0,690	1,337	1,746	2,120	2,583	2,921	3,686
17	0,689	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,330	1,734	2,101	2,552	2,878	3,610
19	0,688	1,328	1,729	2,093	2,539	2,861	3,579
20	0,687	1,325	1,725	2,086	2,528	2,845	3,552
21	0,686	1,323	1,721	2,080	2,518	2,831	3,527
22	0,686	1,321	1,717	2,074	2,508	2,819	3,505
23	0,685	1,319	1,714	2,069	2,500	2,807	3,485
24	0,685	1,318	1,711	2,064	2,492	2,797	3,467
25	0,684	1,316	1,708	2,060	2,485	2,787	3,450
26	0,684	1,315	1,706	2,056	2,479	2,779	3,435
27	0,684	1,314	1,703	2,052	2,473	2,771	3,421
28	0,683	1,313	1,701	2,048	2,467	2,763	3,408
29	0,683	1,311	1,699	2,045	2,462	2,756	3,396
30	0,683	1,310	1,697	2,042	2,457	2,750	3,385
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307
60	0,679	1,296	1,671	2,000	2,390	2,660	3,232
120	0,677	1,289	1,658	1,980	2,358	2,617	3,160
200	0,676	1,286	1,653	1,972	2,345	2,601	3,131

Квантили распределения Фишера

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120
									$p = 0{,}9$									
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,05	62,26	62,53	62,79	63,06
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,17	5,17	5,16	5,15	5,14
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,81	2,80	2,78	2,76	2,74
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,57	2,56	2,54	2,51	2,49
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,27	2,25	2,23	2,21	2,18
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,17	2,16	2,13	2,11	2,08
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,03	2,01	1,99	1,96	1,93
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,93	1,91	1,89	1,86	1,83
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,89	1,87	1,85	1,82	1,79
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,86	1,84	1,81	1,78	1,75
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,83	1,81	1,78	1,75	1,72
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,80	1,78	1,75	1,72	1,69
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,78	1,76	1,73	1,70	1,67
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,76	1,74	1,71	1,68	1,64
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,74	1,72	1,69	1,66	1,62
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,71	1,69	1,66	1,62	1,59
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,68	1,66	1,63	1,59	1,56
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,67	1,65	1,61	1,58	1,54

Продолжение табл. **П.5**

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,66	1,64	1,60	1,57	1,53
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,65	1,63	1,59	1,56	1,52
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,64	1,62	1,58	1,55	1,51
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,63	1,61	1,57	1,54	1,50
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,50	1,48	1,44	1,40	1,35
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,44	1,41	1,37	1,32	1,26
									$p = 0{,}95$									
1	161,5	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,3	250,1	251,1	252,2	253,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,46	19,46	19,47	19,48	19,49
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,63	8,62	8,59	8,57	8,55
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,52	4,50	4,46	4,43	4,40
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,83	3,81	3,77	3,74	3,70
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,40	3,38	3,34	3,30	3,27
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,11	3,08	3,04	3,01	2,97
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,89	2,86	2,83	2,79	2,75
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,73	2,70	2,66	2,62	2,58
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,60	2,57	2,53	2,49	2,45
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,50	2,47	2,43	2,38	2,34
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,41	2,38	2,34	2,30	2,25
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,34	2,31	2,27	2,22	2,18
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,28	2,25	2,20	2,16	2,11
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,23	2,19	2,15	2,11	2,06
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,18	2,15	2,10	2,06	2,01
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,14	2,11	2,06	2,02	1,97
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,07	2,04	1,99	1,95	1,90
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,02	1,98	1,94	1,89	1,84

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,00	1,96	1,91	1,86	1,81
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,97	1,94	1,89	1,84	1,79
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,94	1,90	1,85	1,80	1,75
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,92	1,88	1,84	1,79	1,73
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,89	1,85	1,81	1,75	1,70
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,88	1,84	1,79	1,74	1,68
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,78	1,74	1,69	1,64	1,58
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,69	1,65	1,59	1,53	1,47
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,60	1,55	1,50	1,43	1,35
									$p = 0{,}975$									
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6	976,7	984,9	993,1	998,1	1001,4	1005,6	1009,8	1014,0
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,50	8,46	8,41	8,36	8,31
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,27	6,23	6,18	6,12	6,07
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,11	5,07	5,01	4,96	4,90
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,40	4,36	4,31	4,25	4,20
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,94	3,89	3,84	3,78	3,73
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,60	3,56	3,51	3,45	3,39
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,35	3,31	3,26	3,20	3,14
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,16	3,12	3,06	3,00	2,94
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,01	2,96	2,91	2,85	2,79
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,88	2,84	2,78	2,72	2,66
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,78	2,73	2,67	2,61	2,55
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,69	2,64	2,59	2,52	2,46
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,61	2,57	2,51	2,45	2,38
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,55	2,50	2,44	2,38	2,32
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,49	2,44	2,38	2,32	2,26

Продолжение табл. **П.5**

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,44	2,39	2,33	2,27	2,20
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,40	2,35	2,29	2,22	2,16
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,36	2,31	2,25	2,18	2,11
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,32	2,27	2,21	2,14	2,08
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,29	2,24	2,18	2,11	2,04
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,26	2,21	2,15	2,08	2,01
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,23	2,18	2,12	2,05	1,98
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,21	2,16	2,09	2,03	1,95
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,18	2,13	2,07	2,00	1,93
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,16	2,11	2,05	1,98	1,91
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,14	2,09	2,03	1,96	1,89
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,12	2,07	2,01	1,94	1,87
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	1,99	1,94	1,88	1,80	1,72
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,87	1,82	1,74	1,67	1,58
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,75	1,69	1,61	1,53	1,43
									$p = 0{,}99$									
1	4052	4999	5403	5624	5763	5858	5928	5981	6022	6055	6106	6157	6208	6239	6260	6286	6313	6339
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,58	26,50	26,41	26,32	26,22
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,91	13,84	13,75	13,65	13,56
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,45	9,38	9,29	9,20	9,11
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,30	7,23	7,14	7,06	6,97
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,06	5,99	5,91	5,82	5,74
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,26	5,20	5,12	5,03	4,95
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,71	4,65	4,57	4,48	4,40
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,31	4,25	4,17	4,08	4,00
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,01	3,94	3,86	3,78	3,69
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,76	3,70	3,62	3,54	3,45
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,57	3,51	3,43	3,34	3,25
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,41	3,35	3,27	3,18	3,09

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,28	3,21	3,13	3,05	2,96
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,16	3,10	3,02	2,93	2,84
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,07	3,00	2,92	2,83	2,75
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	2,98	2,92	2,84	2,75	2,66
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,91	2,84	2,76	2,67	2,58
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,84	2,78	2,69	2,61	2,52
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,79	2,72	2,64	2,55	2,46
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,73	2,67	2,58	2,50	2,40
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,69	2,62	2,54	2,45	2,35
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,64	2,58	2,49	2,40	2,31
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,60	2,54	2,45	2,36	2,27
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,57	2,50	2,42	2,33	2,23
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,54	2,47	2,38	2,29	2,20
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,51	2,44	2,35	2,26	2,17
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,48	2,41	2,33	2,23	2,14
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,45	2,39	2,30	2,21	2,11
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,27	2,20	2,11	2,02	1,92
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,10	2,03	1,94	1,84	1,73
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,93	1,86	1,76	1,66	1,53
									$p = 0{,}995$									
1	162†	199†	216†	224†	230†	234†	237†	239†	241†	242†	244†	246†	248†	249†	250†	251†	252†	253†
2	198,5	199,0	199,2	199,3	199,3	199,3	199,4	199,4	199,4	199,4	199,4	199,4	199,5	199,5	199,5	199,5	199,5	199,5
3	55,55	49,80	47,47	46,19	45,39	44,84	44,43	44,13	43,88	43,69	43,39	43,08	42,78	42,59	42,47	42,31	42,15	41,99
4	31,33	26,28	24,26	23,15	22,46	21,97	21,62	21,35	21,14	20,97	20,70	20,44	20,17	20,00	19,89	19,75	19,61	19,47
5	22,78	18,31	16,53	15,56	14,94	14,51	14,20	13,96	13,77	13,62	13,38	13,15	12,90	12,76	12,66	12,53	12,40	12,27
6	18,63	14,54	12,92	12,03	11,46	11,07	10,79	10,57	10,39	10,25	10,03	9,81	9,59	9,45	9,36	9,24	9,12	9,00
7	16,24	12,40	10,88	10,05	9,52	9,16	8,89	8,68	8,51	8,38	8,18	7,97	7,75	7,62	7,53	7,42	7,31	7,19
8	14,69	11,04	9,60	8,81	8,30	7,95	7,69	7,50	7,34	7,21	7,01	6,81	6,61	6,48	6,40	6,29	6,18	6,06
9	13,61	10,11	8,72	7,96	7,47	7,13	6,88	6,69	6,54	6,42	6,23	6,03	5,83	5,71	5,62	5,52	5,41	5,30
10	12,83	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,85	5,66	5,47	5,27	5,15	5,07	4,97	4,86	4,75

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120
11	12,23	8,91	7,60	6,88	6,42	6,10	5,86	5,68	5,54	5,42	5,24	5,05	4,86	4,74	4,65	4,55	4,45	4,34
12	11,75	8,51	7,23	6,52	6,07	5,76	5,52	5,35	5,20	5,09	4,91	4,72	4,53	4,41	4,33	4,23	4,12	4,01
13	11,37	8,19	6,93	6,23	5,79	5,48	5,25	5,08	4,94	4,82	4,64	4,46	4,27	4,15	4,07	3,97	3,87	3,76
14	11,06	7,92	6,68	6,00	5,56	5,26	5,03	4,86	4,72	4,60	4,43	4,25	4,06	3,94	3,86	3,76	3,66	3,55
15	10,80	7,70	6,48	5,80	5,37	5,07	4,85	4,67	4,54	4,42	4,25	4,07	3,88	3,77	3,69	3,58	3,48	3,37
16	10,58	7,51	6,30	5,64	5,21	4,91	4,69	4,52	4,38	4,27	4,10	3,92	3,73	3,62	3,54	3,44	3,33	3,22
17	10,38	7,35	6,16	5,50	5,07	4,78	4,56	4,39	4,25	4,14	3,97	3,79	3,61	3,49	3,41	3,31	3,21	3,10
18	10,22	7,21	6,03	5,37	4,96	4,66	4,44	4,28	4,14	4,03	3,86	3,68	3,50	3,38	3,30	3,20	3,10	2,99
19	10,07	7,09	5,92	5,27	4,85	4,56	4,34	4,18	4,04	3,93	3,76	3,59	3,40	3,29	3,21	3,11	3,00	2,89
20	9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,96	3,85	3,68	3,50	3,32	3,20	3,12	3,02	2,92	2,81
21	9,83	6,89	5,73	5,09	4,68	4,39	4,18	4,01	3,88	3,77	3,60	3,43	3,24	3,13	3,05	2,95	2,84	2,73
22	9,73	6,81	5,65	5,02	4,61	4,32	4,11	3,94	3,81	3,70	3,54	3,36	3,18	3,06	2,98	2,88	2,77	2,66
23	9,63	6,73	5,58	4,95	4,54	4,26	4,05	3,88	3,75	3,64	3,47	3,30	3,12	3,00	2,92	2,82	2,71	2,60
24	9,55	6,66	5,52	4,89	4,49	4,20	3,99	3,83	3,69	3,59	3,42	3,25	3,06	2,95	2,87	2,77	2,66	2,55
25	9,48	6,60	5,46	4,84	4,43	4,15	3,94	3,78	3,64	3,54	3,37	3,20	3,01	2,90	2,82	2,72	2,61	2,50
26	9,41	6,54	5,41	4,79	4,38	4,10	3,89	3,73	3,60	3,49	3,33	3,15	2,97	2,85	2,77	2,67	2,56	2,45
27	9,34	6,49	5,36	4,74	4,34	4,06	3,85	3,69	3,56	3,45	3,28	3,11	2,93	2,81	2,73	2,63	2,52	2,41
28	9,28	6,44	5,32	4,70	4,30	4,02	3,81	3,65	3,52	3,41	3,25	3,07	2,89	2,77	2,69	2,59	2,48	2,37
29	9,23	6,40	5,28	4,66	4,26	3,98	3,77	3,61	3,48	3,38	3,21	3,04	2,86	2,74	2,66	2,56	2,45	2,33
30	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34	3,18	3,01	2,82	2,71	2,63	2,52	2,42	2,30
40	8,83	6,07	4,98	4,37	3,99	3,71	3,51	3,35	3,22	3,12	2,95	2,78	2,60	2,48	2,40	2,30	2,18	2,06
60	8,49	5,79	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90	2,74	2,57	2,39	2,27	2,19	2,08	1,96	1,83
120	8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,93	2,81	2,71	2,54	2,37	2,19	2,07	1,98	1,87	1,75	1,61
									$p = 0{,}999$									
1	405‡	500‡	540‡	562‡	576‡	585‡	592‡	598‡	602‡	605‡	610‡	615‡	620‡	624‡	626‡	628‡	631‡	634‡
2	998,5	999,0	999,2	999,3	999,3	999,3	999,4	999,4	999,4	999,4	999,4	999,4	999,5	999,5	999,5	999,5	999,5	999,5
3	167,0	148,5	141,1	137,1	134,6	132,8	131,6	130,6	129,9	129,2	128,3	127,4	126,4	125,8	125,5	125,0	124,5	124,0
4	74,14	61,25	56,18	53,44	51,71	50,53	49,66	49,00	48,47	48,05	47,41	46,76	46,10	45,70	45,43	45,09	44,75	44,40
5	47,18	37,12	33,20	31,09	29,75	28,83	28,16	27,65	27,24	26,92	26,42	25,91	25,39	25,08	24,87	24,60	24,33	24,06
6	35,51	27,00	23,70	21,92	20,80	20,03	19,46	19,03	18,69	18,41	17,99	17,56	17,12	16,85	16,67	16,44	16,21	15,98

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120
7	29,25	21,69	18,77	17,20	16,21	15,52	15,02	14,63	14,33	14,08	13,71	13,32	12,93	12,69	12,53	12,33	12,12	11,91
8	25,41	18,49	15,83	14,39	13,48	12,86	12,40	12,05	11,77	11,54	11,19	10,84	10,48	10,26	10,11	9,92	9,73	9,53
9	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89	9,57	9,24	8,90	8,69	8,55	8,37	8,19	8,00
10	21,04	14,91	12,55	11,28	10,48	9,93	9,52	9,20	8,96	8,75	8,45	8,13	7,80	7,60	7,47	7,30	7,12	6,94
11	19,69	13,81	11,56	10,35	9,58	9,05	8,66	8,35	8,12	7,92	7,63	7,32	7,01	6,81	6,68	6,52	6,35	6,18
12	18,64	12,97	10,80	9,63	8,89	8,38	8,00	7,71	7,48	7,29	7,00	6,71	6,40	6,22	6,09	5,93	5,76	5,59
13	17,82	12,31	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80	6,52	6,23	5,93	5,75	5,63	5,47	5,30	5,14
14	17,14	11,78	9,73	8,62	7,92	7,44	7,08	6,80	6,58	6,40	6,13	5,85	5,56	5,38	5,25	5,10	4,94	4,77
15	16,59	11,34	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08	5,81	5,54	5,25	5,07	4,95	4,80	4,64	4,47
16	16,12	10,97	9,01	7,94	7,27	6,80	6,46	6,19	5,98	5,81	5,55	5,27	4,99	4,82	4,70	4,54	4,39	4,23
17	15,72	10,66	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58	5,32	5,05	4,78	4,60	4,48	4,33	4,18	4,02
18	15,38	10,39	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39	5,13	4,87	4,59	4,42	4,30	4,15	4,00	3,84
19	15,08	10,16	8,28	7,27	6,62	6,18	5,85	5,59	5,39	5,22	4,97	4,70	4,43	4,26	4,14	3,99	3,84	3,68
20	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08	4,82	4,56	4,29	4,12	4,00	3,86	3,70	3,54
21	14,59	9,77	7,94	6,95	6,32	5,88	5,56	5,31	5,11	4,95	4,70	4,44	4,17	4,00	3,88	3,74	3,58	3,42
22	14,38	9,61	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83	4,58	4,33	4,06	3,89	3,78	3,63	3,48	3,32
23	14,20	9,47	7,67	6,70	6,08	5,65	5,33	5,09	4,89	4,73	4,48	4,23	3,96	3,79	3,68	3,53	3,38	3,22
24	14,03	9,34	7,55	6,59	5,98	5,55	5,23	4,99	4,80	4,64	4,39	4,14	3,87	3,71	3,59	3,45	3,29	3,14
25	13,88	9,22	7,45	6,49	5,89	5,46	5,15	4,91	4,71	4,56	4,31	4,06	3,79	3,63	3,52	3,37	3,22	3,06
26	13,74	9,12	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48	4,24	3,99	3,72	3,56	3,44	3,30	3,15	2,99
27	13,61	9,02	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41	4,17	3,92	3,66	3,49	3,38	3,23	3,08	2,92
28	13,50	8,93	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35	4,11	3,86	3,60	3,43	3,32	3,18	3,02	2,86
29	13,39	8,85	7,12	6,19	5,59	5,18	4,87	4,64	4,45	4,29	4,05	3,80	3,54	3,38	3,27	3,12	2,97	2,81
30	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24	4,00	3,75	3,49	3,33	3,22	3,07	2,92	2,76
40	12,61	8,25	6,59	5,70	5,13	4,73	4,44	4,21	4,02	3,87	3,64	3,40	3,14	2,98	2,87	2,73	2,57	2,41
60	11,97	7,77	6,17	5,31	4,76	4,37	4,09	3,86	3,69	3,54	3,32	3,08	2,83	2,67	2,55	2,41	2,25	2,08
120	11,38	7,32	5,78	4,95	4,42	4,04	3,77	3,55	3,38	3,24	3,02	2,78	2,53	2,37	2,26	2,11	1,95	1,77

Примечание. Знак „†“ означает умножение на 10^2, а знак „‡“ — умножение на 10^3 (например, 162† означает $162 \cdot 10^2$, а 405‡ означает $405 \cdot 10^3$).

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

Учебники и учебные пособия

Беляев Ю.К., Чепурин Е.В. Основы математической статистики. М.: Наука, 1979.

Боровков А.А. Математическая статистика. Оценка параметров, проверка гипотез. М.: Наука, 1984.

Бочаров П.П., Печинкин А.В. Теория вероятностей. Математическая статистика. М.: Гардарика, 1998. 328 с.

Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высш. шк., 1972. 477 с.

Ивченко Г.И., Медведев Ю.И. Математическая статистика. М.: Высш. шк., 1992. 304 с.

Пугачев В.С. Теория вероятностей и математическая статистика. М.: Наука, 1979. 495 с.

Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. М.: Наука, 1965. 511 с.

Элементы математической статистики / *О.И. Тескин, Н.Е. Козлов, Г.М. Цветкова, Е.М. Пашовкин.* М.: Изд-во МГТУ, 1995. 107 с.

Задачники

Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М.: Высш. шк., 1975. 334 с.

Емельянов Г.В., Скитович В.П. Задачник по теории вероятностей и математической статистике. Л.: Изд-во ЛГУ, 1967. 330 с.

Мешалкин Л.Д. Сборник задач по теории вероятностей. М.: Изд-во МГУ, 1963. 143 с.

Сборник задач по математике для втузов. Ч. 3. Теория вероятностей и математическая статистика / Под ред. *А.В. Ефимова.* М.: Наука, 1990. 428 с.

Справочная литература и монографии

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика, 1985. 487 с.

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983. 471 с.

Андерсон Т. Введение в многомерный статистический анализ / Пер. с англ. М.: Физматгиз, 1963. 500 с.

Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1983. 416 с.

Вальд А. Последовательный анализ. М.: Физматгиз. 1960. 328 с.

Гнеденко В.Б., Беляев Ю.К., Соловьев А.Д. Математические методы в теории надежности. М., 1965. 524 с.

Кашьяп Р.Л., Рао А.Р. Построение динамических стохастических моделей по экспериментальным данным / Пер. с англ. под ред. *В.С. Пугачева*. М.: Наука, 1983. 384 с.

Кендалл М., Стюарт А. Статистические выводы и связи / Пер. с англ. под. ред. *А.Н. Колмогорова*. М.: Наука, 1973. 900 с.

Крамер Г. Математические методы статистики / Пер. с англ. под. ред. *А.Н. Колмогорова*. М.: Мир, 1975. 648 с.

Леман Э. Проверка статистических гипотез / Пер. с англ. *Ю.В. Прохорова*. М.: Наука, 1979. 408 с.

Мартынов Г.В. Критерии омега-квадрат. М.: Наука, 1978. 80 с.

Прикладная статистика. Классификация и снижение размерности / *С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин*. М.: Финансы и статистика, 1989. 607 с.

Рао С.Р. Линейные статистические методы и их применения / Пер. с англ. под. ред. *Ю.В. Линника*. М.: Наука, 1968. 548 с.

Робастность в статистике. Подход на основе функций влияния: Пер. с англ. / *Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль*. М.: Мир, 1989. 512 с.

Смирнов Н.В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках. Бюл. Моск. ун-та. Серия А, 1939. Т.2. С. 973–994.

Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. М.: ИНФРА-М, 1998. 528 с.

Хеттманспергер Т. Статистические выводы, основанные на рангах / Пер. с англ. М.: Финансы и статистика, 1987. 334 с.

Холлендер М., Вулф Д. Непараметрические методы статистики / Пер. с англ. М.: Финансы и статистика, 1983. 518 с.

Хьюбер Дж. П. Робастность в статистике / Пер. с англ. М.: Мир, 1984. 304 с.

Шеффе Г. Дисперсионный анализ / Пер. с англ. М.: Физматгиз, 1963. 626 с.

Ширяев А.Н. Статистический последовательный анализ. М.: Наука, 1976. 272 с.

Hodges J.L., Jr and *Lehmann E.L.* Estimates of location based on rank tests // Ann. Math. Stat. 1963. 34. 598–611 p..

Gnedenko B.V., Pavlov I.V., Ushakov I.A. Statistical reliability engineering. N.Y.: John Wiley, 1999. 514 p.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Учебное издание

Математика в техническом университете
Выпуск XVII

Горяинов Владимир Борисович
Павлов Игорь Валерианович
Цветкова Галина Михайловна
Тескин Олег Иванович

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Редактор *Е.В. Авалова*
Художник *С.С. Водчиц*
Корректор *О.В. Калашникова*

Оригинал-макет подготовлен
в Издательстве МГТУ им. Н.Э. Баумана
под руководством *А.Н. Канатникова*

Изд. лиц. № 020523 от 25.04.97

Подписано в печать 20.12.2000. Формат 60×88 1/16.
Печать офсетная. Бумага офсетная № 1.
Усл. печ. л. 26,5. Уч.-изд. л. 24,32.
Тираж 3000 экз. Заказ № 7482.

Издательство МГТУ им. Н.Э. Баумана.
107005, Москва, 2-я Бауманская, 5.

Отпечатано в Производственно-издательском комбинате ВИНИТИ.
140010, г. Люберцы Московской обл., Октябрьский пр-т, 403.
Тел. 554-21-86

ISBN 5-7038-1730-7

9 785703 817308