

А. Г. Буховец
П. В. Москалев
В. П. Богатова
Т. Я. Бирючинская

Под редакцией
профессора Буховца А. Г.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В СИСТЕМЕ R

Учебное пособие

Элементы линейной алгебры
Сведения из теории вероятностей
Основы математической статистики
Начала регрессионного анализа

ВОРОНЕЖ 2010

Статистический анализ данных в системе R. Учебное пособие / А.Г. Буховец, П.В. Москалев, В.П. Богатова, Т.Я. Бирючинская; Под ред. проф. Буховца А.Г. — Воронеж: ВГАУ, 2010. — 124 с.

Учебное пособие предназначено для студентов, обучающихся по направлениям 080100 — «Экономика» и 110300 — «Агроинженерия», программа которых предусматривает изучение современных средств и методов проведения статистического анализа данных. В учебном пособии кратко излагается соответствующий теоретический материал и приводятся примеры решения практических задач по разделам: линейная алгебра, теория вероятностей и математическая статистика с применением системы статистической обработки данных и программирования **R**. В качестве приложений настоящее пособие содержит описание системы **R** и листинги программ, которые могут быть использованы в учебном процессе.

Рецензенты:

Профессор кафедры высшей математики Воронежского государственного архитектурно-строительного университета, д.ф.-м.н., проф. Семенов М.Е.

Доцент кафедры информационного обеспечения и моделирования Воронежского государственного аграрного университета им. К.Д. Глинки, к.э.н., доц. Кулев С.А.

© А.Г. Буховец, П.В. Москалев, В.П. Богатова, Т.Я. Бирючинская, 2010.

© ФГОУ ВПО «Воронежский государственный аграрный университет им. К.Д. Глинки», 2010.

Оглавление

Введение	4
Глава 1. Элементы линейной алгебры	6
1.1. Векторное пространство	6
1.2. Базис векторного пространства	8
1.3. Скалярное произведение векторов	10
1.4. Матрицы	12
1.5. Транспонирование, произведение и ранг матрицы	14
1.6. Определители и собственные значения	17
Глава 2. Сведения из теории вероятностей	22
2.1. Случайное событие и вероятность	22
2.2. Условная вероятность и независимость событий	24
2.3. Случайные величины и законы распределения	24
2.4. Многомерные случайные величины	26
2.5. Числовые характеристики случайных величин	28
2.6. Наиболее распространённые распределения	31
Глава 3. Основы математической статистики	49
3.1. Генеральная и выборочная совокупности	49
3.2. Выборочные характеристики и точечные оценки	50
3.3. Интервальные оценки параметров распределения	56
3.4. Проверка статистических гипотез	61
Глава 4. Начала регрессионного анализа	84
4.1. Основные понятия регрессионного анализа	84
4.2. Модели множественной линейной регрессии	93
Литература	106
Приложение А. Введение в систему R	107
А.1. Принципы взаимодействия с R	107
Приложение В. Листинги программ	112
В.1. Наиболее распространённые распределения	112
В.2. Основы математической статистики	117
В.3. Начала регрессионного анализа	121

Введение

Предлагаемое вниманию читателей учебное пособие рассчитано для студентов инженерных или экономических специальностей, которые как самостоятельно, так и под руководством преподавателя занимаются изучением методов проведения статистического анализа данных с помощью современных программных средств. В главах 1–4 настоящего пособия в краткой форме излагаются основные сведения из линейной алгебры, теории вероятностей, математической статистики и её приложений.

Сведения, приводимые в первой главе, имеют справочный характер и сопровождаются относительно простыми примерами, иллюстрирующими базовые свойства векторов, матриц и операций над ними, а сведения во второй главе — примерами, иллюстрирующими функции распределения и числовые характеристики случайных величин с некоторыми, наиболее распространёнными законами распределения. Основной теоретический материал излагается в третьей и четвёртой главах и иллюстрируется более развёрнутыми примерами, ориентированными на практические задачи математической статистики и регрессионного анализа. Завершается учебное пособие приложениями с описанием базовых принципов работы системы статистической обработки данных **R**, а также с листингами примеров на языке **R**, оформленными с учётом их самостоятельного применения.

Система статистической обработки данных и программирования **R** возникла в 1993 году как свободная альтернатива системы **S-PLUS**, которая в свою очередь являлась развитием языка **S**, разработанного в конце 1970-х годов в компании Bell Labs специально для решения задач прикладной статистики. Первая реализация **S** была написана на языке FORTRAN и работала под управлением операционной системы GCOS. Однако широкое распространение языка **S** в университетской среде началось только в первой половине 1980-х годов, после его переноса на операционную систему UNIX. В настоящее время язык **S** продолжает своё развитие в составе коммерческого продукта **S-PLUS**, разработанного в 1988 году американской компанией Statistical Sciences, Inc. и на протяжении последних полутора десятилетий прочно входящего в число наиболее развитых систем статистической обработки данных.

Во второй половине 1993 года двое молодых учёных Росс Иейка (Ross Ihaka) и Роберт Джентльмен (Robert Gentleman), специализировавшихся в области вычислительной статистики, анонсировали свою новую разработку, которую назвали **R** [1]. По замыслу создателей, **R** должен был стать свободной реализацией языка **S**, отличающейся от своего прародителя легко расширяемой модульной архитектурой, при сохранении быстрой работы, присущего программам на FORTRAN.

В первые годы проект **R** развивался достаточно медленно, но по мере накопления «критической численности» сообщества пользователей и поддерживаемых ими расширений **R** процесс развития ускорился и в скором времени возникла распределённая система хранения и распространения пакетов к **R**, известная под аббревиатурой «CRAN» [2]. Основная идея организации такой системы состояла в том, что оперативное внедрение все новых и новых функций в монолитную программу требует непрерывных и хорошо скоординированных усилий многих десятков (а быть может и сотен) специалистов из самых разных областей. В то же время, достаточно качественный прикладной пакет, реализующий всего несколько функций, квалифицированный специалист вполне способен написать в одиночку за обозримый промежуток времени, а наличие обратной связи с другими специалистами, заинтересованными в данной разработке, позволяет осуществлять как оперативное тестирование уже написанного кода, так и внедрение новых функций.

В настоящее время реализации **R** существуют для трёх наиболее распространённых семейств операционных систем: GNU/Linux, Apple Mac OS X и Microsoft Windows, а в распределённых хранилищах системы CRAN по состоянию на конец сентября 2010 года были доступны для свободной загрузки 2548 пакетов расширения, ориентированных на специфические задачи обработки данных, возникающие в эконометрике и финансовом анализе, генетике и молекулярной биологии, экологии и геологии, медицине и фармацевтике и многих других прикладных областях. Значительная часть европейских и американских университетов в последние годы активно переходят к использованию **R** в учебной и научно-исследовательской деятельности вместо дорогостоящих коммерческих разработок.

Глава 1

Элементы линейной алгебры

В данной главе приведён краткий обзор основных понятий линейной алгебры и матричного исчисления, используемых в статистических методах обработки экспериментальных данных. Приводимые примеры демонстрируют использование этих понятий для эффективного решения прикладных задач на языке статистической обработки данных и программирования **R** [1]. Излагаемый материал не претендует на полноту и математическую строгость изложения и никоим образом не подменяет основных учебников по освещаемым темам [3, 9].

1.1. Векторное пространство

В традиционных курсах линейной алгебры векторное пространство определяется как некоторое множество объектов (векторов), на котором выполняются некоторые аксиомы. В данном разделе определим *n*-мерный вектор x как столбец, состоящий из n действительных чисел x_i , записанных в определённом порядке $i = 1, 2, \dots, n$ и называемых *координатами* или *компонентами* вектора

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}.$$

Два вектора называются *равными* $x = y$, если равны их соответствующие координаты: $x_i = y_i$, $i = 1, 2, \dots, n$. Для заданных в такой форме векторов определены две линейные операции:

1. *Сложение* векторов x и y

$$x + y = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \dots + \dots \\ x_n + y_n \end{pmatrix};$$

2. *Умножение* вектора x на вещественное число α

$$\alpha x = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \dots \\ \alpha x_n \end{pmatrix}.$$

Для этих операций справедливы следующие *свойства* векторного пространства:

1. $x + y = y + x$, $\alpha x = x\alpha$;
2. $x + (y + z) = (x + y) + z$, $\alpha(\beta x) = (\alpha\beta)x$;
3. $\alpha(x + y) = \alpha x + \alpha y$, $(\alpha + \beta)x = \alpha x + \beta x$;
4. $0x = o$, $x + o = x$, где o — *нулевой вектор*, то есть вектор, все компоненты которого равны нулю.

Множество всех n -мерных векторов с определёнными на нём операциями сложения и умножения на вещественное число называется n -мерным векторным пространством и обозначается R^n .

Пример 1.1. В качестве примера проиллюстрируем вышеуказанные свойства векторов с помощью языка статистической обработки данных и программирования **R**.

```

1 > x <- c(1,2,3,4); y <- c(4,3,2,1)
2 > z <- c(1,3,4,2); o <- c(0,0,0,0)
3 > x+y == y+x; 3*x == x*3
4 [1] TRUE TRUE TRUE TRUE
5 [1] TRUE TRUE TRUE TRUE
6 > x+(y+z) == (x+y)+z; 2*(3*x) == (2*3)*x
7 [1] TRUE TRUE TRUE TRUE
8 [1] TRUE TRUE TRUE TRUE
9 > 2*(x+y) == 2*x + 2*y; (2+3)*x == 2*x + 3*x
10 [1] TRUE TRUE TRUE TRUE
11 [1] TRUE TRUE TRUE TRUE
12 > x+o == x; 0*x == o
13 [1] TRUE TRUE TRUE TRUE
14 [1] TRUE TRUE TRUE TRUE

```

В приведённом листинге все строки, начинающиеся с символа «>», содержат команды, вводимые пользователем в командном окне интерпретатора **R** (смотри номера строк: [1–3], [6], [9], [12]), а все строки, начинающиеся с символов «[1]» — результаты, выводимые **R**: ([4–5], [7–8], [10–11], [13–15]). В общем случае, квадратные скобки в выводе **R** используются для обозначения индекса первого элемента вектора в

текущей строке, что существенно облегчает ориентацию, если выводимый вектор занимает на экране больше одной строки.

В [1–2] строках с помощью функции объединения «с()» поэлементно определяются значения векторов x, y, z, o , присваиваемые затем одноимённым переменным с помощью оператора «<-». Оператор «;» даёт пользователю возможность разместить в одной строке несколько последовательно выполняемых команд.

Далее для переменных « o, x, y, z » иллюстрируется выполнение вышеуказанных свойств [3–15]. Все свойства записываются с использованием логического оператора эквивалентности «==», который производит поэлементное сравнение векторов в левой и правой частях равенства и возвращает результат сравнения в виде логического вектора с константами «TRUE» или «FALSE». Как можно легко убедиться, для приведённых исходных данных все перечисленные свойства векторного пространства выполняются.

Данный пример демонстрирует одну из важнейших особенностей языка **R** — эффективную реализацию векторных операций, позволяющую использовать весьма компактную запись при обработке данных большого объёма.

1.2. Базис векторного пространства

Линейной комбинацией векторов $x_i, i = 1, 2, \dots, k$ в пространстве R^n называется выражение вида

$$\alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k = \sum_{i=1}^k \alpha_i x_i.$$

Система векторов $x_i, i = 1, 2, \dots, k$ называется *линейно независимой*, если равенство

$$\alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k = 0$$

выполняется только в том случае, когда все α_i равны нулю. Если же существует такой набор коэффициентов, в котором хотя бы одно значение α_i отлично от нуля и при этом выполняется указанное равенство, то такая система называется *линейно зависимой*. В линейно зависимой системе любой из векторов может быть представлен как линейная комбинация остальных.

Совокупность линейно независимых векторов $\{e_i\}$, $i = 1, 2, \dots, n$ называется *базисом векторного пространства R^n* , если любой вектор этого пространства x может быть представлен в виде линейной комбинации векторов базиса

$$x = x_1 e_1 + x_2 e_2 + \dots + x_n e_n.$$

Это равенство называется *разложением вектора x по базису $\{e_i\}$* , а числа $\{x_i\}$ — *координатами вектора* в указанном базисе.

Из определения базиса вытекают следующие утверждения:

1. Любой базис n -мерного векторного пространства содержит ровно n векторов, при этом число векторов, образующих базис $\{e_i\}$, $i = 1, 2, \dots, n$, совпадает с *размерностью* векторного пространства, которая обозначается как $\dim R^n = n$.
2. Любой вектор n -мерного векторного пространства *единственным образом* раскладывается по заданному базису $\{e_i\}$, $i = 1, 2, \dots, n$.

Следствием первого утверждения является тот факт, что в R^n любая система, состоящая из s векторов, где $s > n$, является линейно зависимой.

Некоторое подмножество L линейного пространства R^n называется его *линейным подпространством*, если из $x \in L$ и $y \in L$ следует, что $(x + y) \in L$ для любых x и y , а из $x \in L$ следует, что $\alpha x \in L$ при любом вещественном α .

Очевидно, что размерность линейного подпространства не превосходит размерности линейного пространства $\dim L \leq \dim R^n$.

Совокупность всех линейных комбинаций векторов $\{x_i\}$, где $i = 1, 2, \dots, k$ называется *линейной оболочкой* этих векторов.

Пример 1.2. Продолжая предыдущий пример, найдём координаты вектора $a(1, -2, 3, -4)$ в базисе $x(1, 2, 3, 4)$, $y(4, 3, 2, 1)$, $z(1, 3, 4, 2)$, $t(1, 4, 2, 3)$ с помощью языка R . Напомним, что решение этой задачи сводится к решению системы линейных алгебраических уравнений, в которой столбцы векторов базиса (x, y, z, t) формируют матрицу коэффициентов, а разлагаемый по базису вектор a — столбец свободных членов:

$$\begin{pmatrix} x_1 & y_1 & z_1 & t_1 \\ x_2 & y_2 & z_2 & t_2 \\ x_3 & y_3 & z_3 & t_3 \\ x_4 & y_4 & z_4 & t_4 \end{pmatrix} \begin{pmatrix} a'_1 \\ a'_2 \\ a'_3 \\ a'_4 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}.$$

```

15 > x; y; z
16 [1] 1 2 3 4
17 [1] 4 3 2 1
18 [1] 1 3 4 2
19 > a <- c(1,-2,3,-4); t <- c(1,4,2,3)
20 > d <- matrix(c(x,y,z,t), nrow=length(x), byrow=TRUE)
21 > if(det(d) != 0) solve(d,a) else
22 + stop("Векторы линейно зависимы!")
23 [1] -0.3000000 -1.6333333 2.3666667 -0.6333333

```

Так как векторы x , y , z уже были определены ранее [1–2], то для постановки задачи достаточно лишь убедиться в существовании одноимённых переменных [15–18] и определить дополнительные векторы t и a [19].

Для решения системы линейных алгебраических уравнений используется функция «`solve()`» с двумя аргументами [21]: матрицей коэффициентов «`d`» и вектором правых частей «`a`» системы уравнений. Матрица коэффициентов системы линейных алгебраических уравнений «`d`» образуется путём композиции функций «`matrix()`» и «`c()`» из векторов « (x, y, z, t) » с числом строк, определяемым длиной первого вектора «`nrow=length(x)`» [20], а условие «`if(det(d) != 0)`» используется для проверки линейной независимости векторов « x, y, z, t », что является необходимым и достаточным условием для существования одноимённого базиса. Если же указанное условие не будет выполнено: «`det(d) == 0`», то вместо искомых координат в строке [22] будет выдано сообщение об ошибке.

Символ «`+`» в начале строки [22] появляется при переносе слишком длинного выражения с предыдущей строки.

Как видно из приведённого в строке [23] ответа, искомые координаты вектора a' в базисе $\{x, y, z, t\}$ будут равны $(-0.3, -1.6, 2.4, -0.6)$.

1.3. Скалярное произведение векторов

Скалярным произведением векторов x и y называется число (скаляр), обозначаемое как (x, y) или просто xy и определяемое соотношением

$$(x, y) = xy = \sum_{i=1}^n x_i y_i.$$

Основные свойства скалярного произведения:

1. $(x, y) = (y, x)$;

2. $(x_1 + x_2, y) = (x_1, y) + (x_2, y)$;
3. $(\alpha x, y) = \alpha(x, y)$ для любого вещественного α ;
4. $(x, x) = |x|^2 \geq 0$, причём $|x| = 0$ тогда и только тогда, когда $x = 0$, где $|x| = \sqrt{(x, x)}$ — модуль или длина вектора x .

В дополнение к свойствам 1–4 для скалярного произведения двух любых векторов x и y выполняется *неравенство Коши–Буняковского*: $(x, y)^2 \leq (x, x) \cdot (y, y)$.

Векторы x и y называются *коллинеарными*, если $x = \alpha y$. Практически это означает, что координаты векторов x и y пропорциональны друг другу.

Векторы x и y называются *ортogonalными*, если их скалярное произведение равно нулю: $(x, y) = 0$.

Вещественное линейное пространство называется *евклидовым*, если в нём определено скалярное произведение элементов. В евклидовом пространстве удобно использовать базис $\{e_1, e_2, \dots, e_n\}$, все элементы которого взаимно ортогональны и имеют единичную длину:

$$(e_i, e_j) = \delta_{ij}, \quad \delta_{ij} = \begin{cases} 1, & \text{если } i = j; \\ 0, & \text{если } i \neq j, \end{cases}$$

где δ_{ij} — символ *Кронекера*. Такие базисы называются *ортонормированными* и существуют в любом евклидовом пространстве. В ортонормированном базисе координаты вектора x можно представить в виде: $x_i = (x, e_i)$, $i = 1, 2, \dots, n$, а разложение такого вектора по базису

$$x = \sum_{i=1}^n (x, e_i) e_i.$$

Введение в рассмотрение скалярного произведения позволяет в дальнейшем эффективно использовать такие геометрически содержательные понятия, как ортогональность, угол и длина. Эти свойства широко используются при получении системы нормальных уравнений метода наименьших квадратов, а также для объяснения свойств МНК-оценок.

Пример 1.3. В продолжение предыдущего примера выясним ортогональность вектора a с базисом (x, y, z, t) с помощью языка **R**.

24	> a
25	[1] 1 -2 3 -4

```

26 > d
27      [,1] [,2] [,3] [,4]
28 [1,]    1    2    3    4
29 [2,]    4    3    2    1
30 [3,]    1    3    4    2
31 [4,]    1    4    2    3
32 > as.vector(d%*%a)
33 [1]  -10     0    -1   -13

```

Для проверки ортогональности вектора a с векторами базиса требуется вычислить четыре скалярных произведения: (x, a) , (y, a) , (z, a) , (t, a) . Напомним, что в предыдущем примере мы сформировали вспомогательную матрицу «d» из столбцов базисных векторов [21]. Внимательные читатели наверняка обратили внимание, что компоненты матрицы «d» отображаются на экране в обычном порядке [26–31], а компоненты вектора «a» — в транспонированном [24–25]. Это связано с тем, что построчный вывод «длинных» векторов позволяет более эффективно использовать площадь экрана при статистической обработке выборочных данных.

Для вычисления искомых скалярных произведений перемножим матрицу «d» на вектор «a» и представим полученный результат как вектор [32–33]: «as.vector(d%*%a)», где «%*%» означает операцию матричного умножения, определённую далее в разделе 1.5 и позволяющую получить искомые скалярные произведения одной командой.

Как показывают расчёты, ортогональной является вторая пара векторов: $(y, a) = 0 \Rightarrow y(4, 3, 2, 1) \perp a(1, -2, 3, -4)$.

1.4. Матрицы

Прямоугольная таблица чисел, содержащая m строк и n столбцов, называется *числовой матрицей*. Пара чисел m и n называются *размером* матрицы. Обозначаются матрицы следующим образом:

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

Числа a_{ij} , $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$, составляющие матрицу, называются её *элементами*. В случае, если $m = n$, матрица называется *квадратной*, а n — *порядком* матрицы.

Матрицу размера $1 \times n$ называют *матрицей-строкой*, а матрицу размера $m \times 1$ — *матрицей-столбцом*. Очевидно, что последняя может рассматриваться как элемент векторного пространства \mathbf{R}^m .

Главной диагональю квадратной матрицы порядка n называется совокупность элементов: a_{ij} , $i = j = 1, 2, \dots, n$. Квадратная матрица называется *диагональной*, если все её элементы, не лежащие на главной диагонали, равны нулю. Диагональная матрица, все диагональные элементы которой равны единице, называется *единичной* и обозначается I .

Две матрицы A и B называются *равными*, если они имеют одинаковый размер и равные соответствующие элементы.

Основные операции над матрицами:

1. Суммой матриц A и B одинакового размера называется матрица того же размера, определяемая равенством

$$A + B = (a_{ij} + b_{ij}), \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n;$$

2. Произведением матрицы A на число α называется матрица того же размера, определяемая равенством

$$\alpha A = (\alpha a_{ij}), \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n.$$

Основные свойства операций над матрицами:

1. $A + B = B + A$, $\alpha A = A\alpha$;
2. $(A + B) + C = A + (B + C)$, $\alpha(\beta A) = (\alpha\beta)A$;
3. $\alpha(A + B) = \alpha A + \alpha B$, $(\alpha + \beta)A = \alpha A + \beta A$;
4. $A + O = A$, $0A = O$, где O — нулевая матрица, то есть матрица, все элементы которой равны нулю.

Пример 1.4. Проиллюстрируем вышеуказанные свойства для произвольных матриц A, B, C с помощью \mathbf{R} .

```

1 > matrix(round(runif(9, min=-9, max=9)), nrow=3) -> A; A
2      [,1] [,2] [,3]
3 [1,]   -1    9    0
4 [2,]   -6    5    1
5 [3,]    2   -7   -8
6 > matrix(round(runif(9, min=-9, max=9)), nrow=3) -> B; B
7      [,1] [,2] [,3]
8 [1,]    5   -6   -7

```

```

9      [2,]  -5    6    9
10     [3,]   3    4   -5
11 > matrix(round(runif(9, min=-9, max=9)), nrow=3) -> C; C
12      [,1] [,2] [,3]
13     [1,]   0    4   -2
14     [2,]   9    5    8
15     [3,]   0   -4    6
16 > matrix(0, nrow=3, ncol=3) -> O
17 > all(A+B == B+A); all(7*A == A*7)
18      [1] TRUE
19      [1] TRUE
20 > all((A+B)+C == A+(B+C)); all(3*(4*A) == (3*4)*A)
21      [1] TRUE
22      [1] TRUE
23 > all(3*(A+B) == 3*A + 3*B); all((3+4)*A == 3*A + 4*A)
24      [1] TRUE
25      [1] TRUE
26 > all(A+O == A); all(O*A == O)
27      [1] TRUE
28      [1] TRUE

```

Произвольные матрицы A , B , C размером 3×3 формируются с помощью генератора псевдослучайных чисел «runif()»: [1], [6], [11]. Эта функция возвращает вектор из 9 псевдослучайных чисел, равномерно распределённых в диапазоне от «min=-9» до «max=9», которые затем округляются функцией «round()» до целых значений.

Оператор «->» означает операцию присваивания, выполняемую слева — направо: [1], [6], [11], [16].

Нулевая матрица O размером 3×3 формируется с помощью вызова функции «matrix()» [16], повторяющей значение 0 по заданному числу строк «nrow=3» и столбцов «ncol=3».

Функция «all()» используется для сокращённой записи результата проверки свойств матриц: [17], [20], [23], [26]. Эта функция возвращает истинное значение в том случае, если указанное в аргументе условие истинно для всех элементов матрицы.

1.5. Транспонирование, произведение и ранг матрицы

Транспонированием матрицы A называется операция, в результате которой меняются местами строки и столбцы матрицы при сохранении порядка их следования. Полученная в результате этого матрица называется *транспонированной* и обозначается:

$$A^T = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}.$$

Свойства операции транспонирования:

1. $(A^T)^T = A$;
2. $(A + B)^T = A^T + B^T$.

Произведением матриц A размера $m \times n$ и B размера $n \times k$ называется матрица C размера $m \times k$, которая обозначается $C = AB$, и элементы которой определяются по формуле

$$c_{ij} = \sum_{s=1}^n a_{is}b_{sj}, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, k,$$

Если произведение матриц определено, то справедливы его следующие основные свойства:

1. $A(BC) = (AB)C = ABC$;
2. $(A + B)C = AC + BC$, $A(B + C) = AB + AC$;
3. $(AB)^T = B^T A^T$.

Следует особо отметить, что в общем случае произведение матриц не коммутативно: $AB \neq BA$. Более того, существование произведения AB не влечёт за собой существование произведения BA . Тем не менее в частных случаях коммутативность матриц возможна: $AB = BA$, тогда матрицы A и B называются *коммутирующими*.

Также следует отметить, что элементы произведения двух матриц можно рассматривать как скалярные произведения векторов-строк первой матрицы на векторы-столбцы второй. С другой стороны, скалярное произведение двух векторов x и y также может быть записано в виде матричного произведения: $(x, y) = x^T y$.

Рассмотрение столбцов матрицы A размера $m \times n$ в качестве m -мерных векторов позволяет установить их линейную зависимость. Максимальное число линейно-независимых векторов-столбцов матрицы A называется её *рангом по столбцам*. Аналогичным образом можно сформулировать понятие *ранга по строкам* — для этого достаточно перейти к рассмотрению транспонированной матрицы A^T .

Можно доказать, что ранг по столбцам матрицы A равен её рангу по строкам. Обозначается *ранг матрицы* как $\text{rank } A$ или $r(A)$. Из определения очевидно, что $0 \leq \text{rank } A \leq \min(n, m)$. Для нулевой матрицы полагают, что $\text{rank } A = 0$.

Пример 1.5. В продолжение предыдущего примера проиллюстрируем свойства транспонирования и произведения матриц A, B, C , а также вычислим их ранг с помощью **R**.

```

29 > A; t(A)
30      [,1] [,2] [,3]
31 [1,]  -1   9   0
32 [2,]  -6   5   1
33 [3,]   2  -7  -8
34      [,1] [,2] [,3]
35 [1,]  -1  -6   2
36 [2,]   9   5  -7
37 [3,]   0   1  -8

```

Для транспонирования матрицы в приведённом листинге используется функция «`t()`» [29], действие которой можно увидеть из выводимых на экран сообщений [30–37].

```

38 > all(t(t(A)) == A); all(t(A+B) == t(A)+t(B))
39 [1] TRUE
40 [1] TRUE
41 > all(A%*%B != B%*%A); all(A%*%C != C%*%A)
42 [1] TRUE
43 [1] TRUE
44 > all(A%*%(B%*%C) == A%*%B%*%C)
45 [1] TRUE
46 > all((A%*%B)%*%C == A%*%B%*%C)
47 [1] TRUE
48 > all((A+B)%*%C == A%*%C + B%*%C)
49 [1] TRUE
50 > all(A%*%(B+C) == A%*%B + A%*%C)
51 [1] TRUE
52 > all(t(A%*%B) == t(B)%*%t(A))
53 [1] TRUE

```

В строках [41], [44], [46], [49], [50], [52] используется операция матричного умножения, обозначаемая как «`%*%`». Также при проверке коммутативности произведения матриц AB и AC вместо логического равенства «`==`» в строке [41] использовано неравенство «`!=`», причём обе пары матриц AB и AC оказались некоммутирующими.

```

54 > qr(A)$rank; qr(B)$rank; qr(C)[[2]]
55 [1] 3
56 [1] 3
57 [1] 3

```


Для определения ранга матриц A, B, C в строке [54] вызывается функция «`qr()$rank`» или, что равносильно, «`qr()[[2]]`», определяющая ранг передаваемой в качестве аргумента матрицы. Как видно из строк [55–57], ранги матриц A, B, C оказались равными их порядку: $\text{rank } A = \text{rank } B = \text{rank } C = 3$.

1.6. Определители и собственные значения

Каждой квадратной матрице A порядка n по определённому правилу можно поставить в соответствие число, называемое *определителем или детерминантом* матрицы и обозначаемое как $|A|$ или $\det A$. Для вычисления определителя матрицы могут использоваться формулы:

$$\det A = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det A_{ij} = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det A_{ij},$$

где $i, j = 1, 2, \dots, n$; A_{ij} — квадратная матрица порядка $(n-1)$, которая получается из матрицы A вычёркиванием i -ой строки и j -го столбца; $\det A_{ij}$ — *минор* элемента a_{ij} . Эти формулы называются *разложением* определителя матрицы A по j -му столбцу и i -ой строке соответственно.

Основные свойства определителей:

1. Величина определителя *не изменится* при транспонировании матрицы: $\det A^T = \det A$;
2. Определитель произведения двух матриц *равен* произведению их определителей: $\det(AB) = \det A \det B$;
3. При умножении матрицы на вещественное число её определитель *умножается на n -ную степень* этого числа: $\det(\alpha A) = \alpha^n \det A$;
4. Величина определителя *не изменится*, если к элементам одной его строки (столбца) прибавить элементы другой строки (столбца), умноженные на одно и то же вещественное число;
5. При перестановке любых двух строк (столбцов) определитель *меняет знак*;
6. Величина определителя, содержащего две пропорциональные строки (столбца), *равна нулю*;

7. Сумма произведений элементов любой строки (столбца) определителя на алгебраические дополнения к элементам другой его строки (столбца) *равна нулю*:

$$\sum_{i=1}^n (-1)^{i+j} a_{ij} \det A_{tj} = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det A_{it} = 0, \quad i, j \neq t.$$

Матрица A называется *невырожденной*, если её определитель отличен от нуля. Всякая невырожденная матрица имеет единственную *обратную матрицу* A^{-1} , удовлетворяющую равенству: $AA^{-1} = A^{-1}A = I$, где I — единичная матрица.

Основные *свойства* обратных матриц, выполняемые при условии существования всех входящих в соответствующие равенства матриц:

1. $(AB)^{-1} = B^{-1}A^{-1}$, $(A^{-1})^T = (A^T)^{-1}$;
2. $\det A^{-1} = \det^{-1} A$.

Собственным вектором квадратной матрицы A порядка n называется ненулевой вектор x , удовлетворяющий равенству: $Ax = \lambda x$, где λ — некоторое вещественное число, называемое *собственным значением* матрицы A , соответствующим собственному вектору x . Очевидно, что собственный вектор x определён с точностью до коэффициента пропорциональности, и поэтому обычно нормируется условием: $x^T x = 1$.

Для нахождения собственных значений матрицы A исходное уравнение приводят к виду, соответствующему однородной системе линейных алгебраических уравнений

$$(A - \lambda I)x = 0.$$

Для существования ненулевого решения данной системы необходимо и достаточно, чтобы её определитель равнялся нулю

$$\det(A - \lambda I) = 0.$$

Это уравнение называется *характеристическим уравнением* матрицы A . Корнями этого уравнения будут собственные значения матрицы A . При этом, если все корни λ_i характеристического уравнения будут *простыми* (кратность корней равна единице), то соответствующие им собственные векторы x_i будут *линейно независимыми*.

Пример 1.6. Продолжая предыдущий пример, проиллюстрируем свойства определителей и обратных матриц A, B, C , а также найдём их собственные векторы и значения с помощью **R**.

```

58 > det(A) - det(t(A))
59 [1] 3.41061e-13
60 > round(det(A) - det(t(A)), digits=6)
61 [1] 0

```

Важной особенностью функции «`det()`», вычисляющей определитель матрицы, является приближенный характер получаемых результатов, что видно из [58–59]. Запись вида « $3.41061e-13$ » означает весьма близкое, но не равное нулю число, соответствующее заданной предельно допустимой погрешности вычислений: $3.41061 \cdot 10^{-13}$.

В связи с этим, в строке [60] вместо проверки логического равенства, соответствующего первому свойству определителей, мы вычисляем разность между правой и левой частями равенства с последующим округлением до шестого знака с помощью функции «`round()`» с параметром «`digits=6`». В продолжение отметим, что наименование параметра любой функции может быть указано как в сокращённой форме: «`digi=6`» [62], «`d=6`» [64], так и вообще без имени, как в [67]. Вызов функции с именованными параметрами делает исходный код понятнее, а возможность пропускать некоторые имена — компактнее.

```

62 > round(det(A%*%B) - det(A)*det(B), digi=6)
63 [1] 0
64 > round(det(4*A) - 4^3*det(A), d=6)
65 [1] 0
66 > A -> A4; A[,2] - 7*A[,1] -> A4[,2]
67 > round(det(A) - det(A4), 6)
68 [1] 0
69 > A[,c(2,1,3)] -> A5
70 > round(det(A) + det(A5), 6)
71 [1] 0
72 > A -> A6; 7*A[,1] -> A6[,2]
73 > round(det(A6), 6)
74 [1] 0
75 > A[1,1]*det(A[-1,-1]) - A[1,2]*det(A[-1,-2]) +
76 + A[1,3]*det(A[-1,-3]) -> D7a
77 > round(det(A) - D7a, 6)
78 [1] 0
79 > A[1,1]*det(A[-2,-1]) - A[1,2]*det(A[-2,-2]) +
80 + A[1,3]*det(A[-2,-3]) -> D7b
81 > round(D7b, 6)
82 [1] 0

```

В строках [60–82] иллюстрируются основные свойства определителей. Записи вида « $A[,1]$ » и « $A[,2]$ » в [66] означают обращения к первому и второму столбцам матрицы « A », а запись вида « $A[,c(2,1,3)]$ » в [69] — перестановку первого и второго её столбцов.

Символ «+» в начале строк [76] и [80] появляется при переносе слишком длинного выражения с предыдущей строки. Это происходит при нажатии на клавишу [Enter] в том случае, если введённое выражение имеет незакрытую парную скобку («)» или «]») или стоящий в конце строки знак двуместной операции: «+», «-», «*», «/» и т.д.

Выражения вида « $\det(A[-1,-1])$ » в строках [75–76] и [79–80] означают определитель матрицы A без первой строки и первого столбца, то есть минор к элементу a_{11} . Таким образом, в строках [75–76] записано разложение определителя матрицы A по первой строке, а в строках [79–80] записана сумма произведений элементов первой строки матрицы A на алгебраические дополнения к элементам её второй строки.

```

83 > sum(round(A%*%solve(A) - diag(3), 6))
84 [1] 0
85 > sum(round(solve(A%*%B) - solve(B)%*%solve(A), 6))
86 [1] 0
87 > sum(round(t(solve(A)) - solve(t(A)), 6))
88 [1] 0
89 > round(det(solve(A)) - det(A)^-1, 6)
90 [1] 0

```

В строках [83–90] иллюстрируются основные свойства обратных матриц. Функция « $\text{diag}(3)$ » в строке [83] используется для получения единичной матрицы третьего порядка. Для вычисления обратной матрицы используется та же функция, что и для решения системы линейных алгебраических уравнений « $\text{solve}()$ », но только с одним аргументом: [83], [85], [87], [89]. В тех случаях, когда результат предполагал появление нулевой матрицы, использовалась её свёртка с помощью функции суммирования « $\text{sum}()$ »: [83], [85], [87].

```

91 > eigen(A)$values; eigen(B)$values; eigen(C)$values
92 [1] -7.53094+0.00000i 1.76547+6.89017i 1.76547-6.89017i
93 [1] 11.50860 -6.52125 1.01265
94 [1] 6.72278+3.69949i 6.72278-3.69949i -2.44557+0.00000i
95 > round(eigen(B)$vectors, 5)
96 [1,] [2,] [3,]
97 [1,] -0.70254 -0.24472 0.88289
98 [2,] 0.71026 0.49957 0.04099
99 [3,] 0.04443 -0.83099 0.46778

```

Для вычисления собственных значений матриц A, B, C в [91] использованы функции «`eigen()`\$values», а для поиска собственных векторов в [95] — функция «`eigen()`\$vectors». Как видно из результатов расчёта [92–94], матрицы A и C имеют комплексно-сопряжённые собственные значения. Отсюда следует, что вещественные линейно-независимые собственные векторы есть только у B : [95–99].

Контрольные вопросы

1. Сформулируйте определение векторного пространства.
2. Дайте определения операций сложения векторов и умножения вектора на число. Перечислите основные свойства этих операций.
3. Какие векторы называются линейно независимыми и линейно зависимыми?
4. Дайте определение базиса векторного пространства. Сколько различных базисов можно указать в конечномерном векторном пространстве?
5. Дайте определение скалярного произведения векторов. Перечислите основные свойства скалярного произведения.
6. Какие векторы называются ортогональными?
7. Что называется координатами вектора в заданном базисе?
8. Дайте определение матрицы. Что такое размер матрицы? Какие матрицы называются квадратными? Что такое порядок квадратной матрицы?
9. Какие матрицы называются равными?
10. Какие операции определены для матриц. При каких условиях эти операции выполнимы? Укажите основные свойства этих операций.
11. Какие матрицы называются коммутативными?
12. Дайте определение обратной матрицы. Укажите условия, при которых матрица A имеет обратную. Приведите пример квадратной матрицы, не имеющей обратной.

Глава 2

Сведения из теории вероятностей

В данной главе приведён краткий обзор основных понятий теории вероятностей, используемых затем в математической статистике и статистических методах обработки экспериментальных данных. Приводимые примеры демонстрируют использование этих понятий для решения прикладных задач на языке статистической обработки данных и программирования **R** [1]. Излагаемый материал не претендует на полноту и математическую строгость изложения и никоим образом не подменяет основных учебников по освещаемым темам [4–6].

2.1. Случайное событие и вероятность

В теории вероятностей понятие события является первичным и не определяется через другие более простые понятия. Для описания событий как результатов испытаний (также называемых опытами или наблюдениями) с неопределённым исходом используется понятие случайности. Под *испытанием* (или *экспериментом*) понимают любое наблюдение какого-либо явления, выполненное в заданном комплексе условий с фиксацией результата, которое может быть повторено (хотя бы в принципе) достаточное число раз.

Испытание, исход которого не может быть определён однозначно до проведения эксперимента, принято называть *случайным*.

Наряду с самим событием A в рассмотрение вводится *противоположное* к нему событие \bar{A} , которое заключается в том, что событие A не происходит.

Событие, которое при случайном испытании происходит всегда, называется *достоверным* и обозначается как Ω .

Событие, которое никогда не происходит, то есть является противоположным к достоверному, называется *невозможным* и обозначается как \emptyset .

События A и B называются *несовместными*, если появление одного из них исключает появление другого. Иначе говоря, такие собы-

тия никогда не происходят одновременно.

Пусть на рассматриваемом множестве событий определены следующие операции:

1. *Сумма событий* $A + B$ — событие, состоящее в том, что произойдёт хотя бы одно из событий: A и/или B ;
2. *Произведение событий* AB — событие, состоящее в том, что произойдут оба события: и A , и B .

Событие эксперимента (испытания) считается *элементарным* ω , если его нельзя представить через другие события с помощью операций сложения и умножения.

Совокупность всех таких событий $\{\omega_1, \omega_2, \dots, \omega_n\}$ образует *пространство элементарных исходов* Ω :

$$\sum_{i=1}^n \omega_i = \Omega, \quad \omega_i \omega_j = \emptyset, \quad \text{если } i \neq j.$$

Предполагается, что каждому возможному исходу ω_i в данном испытании, может быть сопоставлена неотрицательная числовая функция, такая что $P\{\omega_i\} = p_i$. Значения этой функции, выражающие меру возможности осуществления элементарного события ω_i , называется его *вероятностью*. При этом имеют место следующие *свойства вероятности*: $P\{\omega_i\} \in (0, 1)$, $P\{\emptyset\} = 0$, $P\{\Omega\} = 1$.

В рамках такого подхода любое событие A , связанное с этим экспериментом, определяется как сумма элементарных исходов, а его вероятность — как сумма вероятностей соответствующих элементарных исходов

$$P\{A\} = \sum_{\omega_i \in A} P\{\omega_i\}.$$

Для таких случайных событий справедливы два утверждения, называемых *теоремами сложения вероятностей*:

1. Если события A и B — несовместны: $AB = \emptyset$, то $P\{A + B\} = P\{A\} + P\{B\}$;
2. Если же события A и B — совместны: $AB \neq \emptyset$, то $P\{A + B\} = P\{A\} + P\{B\} - P\{AB\}$.

2.2. Условная вероятность и независимость событий

Если некоторое событие A рассматривается не на всём пространстве элементарных исходов, а лишь на некоторой его части, где кроме A осуществляется и другое событие B , то имеет смысл использовать определение *условной вероятности* события A , откуда следует *теорема умножения вероятностей*:

$$P\{A|B\} = \frac{P\{AB\}}{P\{B\}} \Rightarrow P\{AB\} = P\{B\} P\{A|B\}.$$

Событие A полагают *не зависящим* от B , если $P\{A|B\} = P\{A\}$. Иначе говоря, события A и B считаются *независимыми*, если появление одного из них не изменяет вероятности другого события. Для *независимых событий* теорема умножения вероятностей принимает более простой вид

$$P\{AB\} = P\{A\} P\{B\}.$$

Это равенство часто рассматривают как определение *независимости событий* A и B .

Понятия независимости случайных событий и условной вероятности являются очень важными для математической статистики. Достаточно отметить, что многие свойства статистических оценок получаются именно в предположении независимости входящих в них случайных величин. А понятие условной вероятности используется при определении регрессионной модели.

2.3. Случайные величины и законы распределения

Случайная величина X представляет собой однозначную действительную функцию, заданную на пространстве элементарных событий Ω . Каждая случайная величина задаёт распределение вероятностей на множестве своих возможных значений.

Законом распределения случайной величины X называется всякое соотношение, устанавливающее связь между возможными значениями этой случайной величины и соответствующими им вероятностями. Случайная величина X считается заданной, если известен её закон распределения.

Наиболее общей формой закона распределения является *функция распределения вероятностей* случайной величины, определяемая равенством

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

Основные свойства функции распределения $F(x)$:

1. Значения функции распределения ограничены интервалом:
 $0 \leq F(x) \leq 1$;
2. Функция распределения — *неубывающая* функция:
 $F(x_2) \geq F(x_1)$, если $x_2 > x_1$;
3. Предельные значения аргумента соответствуют предельным значениям функции распределения: $F(-\infty) = 0$, $F(\infty) = 1$;
4. Вероятность события $X \in [\alpha, \beta)$ равна приращению функции распределения на соответствующем интервале:
 $\mathbf{P} \{\alpha \leq X < \beta\} = F(\beta) - F(\alpha)$.

В зависимости от структуры множества возможных значений в практических задачах обычно различают два вида случайных величин: *дискретные* и *непрерывные*.

Дискретной называется случайная величина, множество возможных значений которой конечное или счётное. В качестве закона распределения дискретной случайной величины часто используют ряд распределения, записываемый в виде таблицы $2 \times n$:

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix},$$

где $p_i = \mathbf{P} \{X = x_i\}$ при этом $\sum_{i=1}^n p_i = 1$.

Функция распределения дискретной случайной величины будет иметь разрывы первого рода (скачки), в точках, соответствующих значениям случайной величины x_i (абсциссы скачков). Причем величины этих скачков будут равны вероятностям соответствующих значений p_i (ординаты скачков).

Непрерывной называется случайная величина, имеющая непрерывную и дифференцируемую функцию распределения $F(x)$.

В качестве закона распределения непрерывной случайной величины обычно используется *функция плотности распределения вероятностей*:

$$f(x) = \frac{dF(x)}{dx}.$$

Основные свойства плотности распределения вероятностей $f(x)$:

1. Плотность распределения вероятностей — функция неотрицательная: $f(x) \geq 0$;
2. Плотность распределения удовлетворяет условию нормировки:

$$\int_{-\infty}^{\infty} f(x) dx = 1;$$

3. Вероятность события $X \in [\alpha, \beta]$ равна интегралу на соответствующем отрезке от плотности распределения:

$$\mathbf{P}\{\alpha \leq X \leq \beta\} = \int_{\alpha}^{\beta} f(x) dx;$$

4. Функция распределения равна несобственному интегралу от плотности распределения с переменным верхним пределом:

$$F(x) = \int_{-\infty}^x f(t) dt.$$

2.4. Многомерные случайные величины

Понятие случайной величины может быть обобщено на случай: *системы случайных величин*: $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$, где \mathbf{X} рассматривается как *n -мерный случайный вектор*, а (X_1, X_2, \dots, X_n) — как система случайных величин, определённых на едином пространстве элементарных событий Ω .

Функция распределения n -мерной случайной величины \mathbf{X} задаётся равенством

$$F(x_1, x_2, \dots, x_n) = \mathbf{P}\{X_1 < x_1, X_2 < x_2, \dots, X_n < x_n\}.$$

Случайный вектор \mathbf{X} называется *непрерывным*, если его функция распределения $F(x_1, x_2, \dots, x_n)$ имеет смешанную частную производную n -го порядка, которая называется *плотностью распределения* случайного вектора \mathbf{X} или *совместной плотностью распределения* системы случайных величин (X_1, X_2, \dots, X_n) :

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n F(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

Заметим, что *свойства плотности вероятности n -мерной случайной величины* аналогичны свойствам плотности вероятности *одномерной случайной величины*.

Если рассмотрению подлежит только часть компонент вектора $X = (X_1, X_2, \dots, X_k)^\top$, где $k < n$, то используется *частная (маргинальная) функция распределения*:

$$\begin{aligned} F(x_1, x_2, \dots, x_k) &= \mathbf{P}\{X_1 < x_1, X_2 < x_2, \dots, X_k < x_k\} = \\ &= \mathbf{P}\{X_1 < x_1, X_2 < x_2, \dots, X_k < x_k, X_{k+1} < \infty, \dots, X_n < \infty\} = \\ &= F(x_1, x_2, \dots, x_k, \infty, \dots, \infty), \end{aligned}$$

а также *частная (маргинальная) плотность распределения*:

$$\begin{aligned} f_{1,2,\dots,k}(x_1, x_2, \dots, x_k) &= \\ &= \int \cdots \int f(x_1, x_2, \dots, x_k, x_{k+1}, \dots, x_n) dx_{k+1} \cdots dx_n, \end{aligned}$$

где интегрирование производится по всему множеству возможных значений переменных x_{k+1}, \dots, x_n .

Плотность распределения многомерной случайной величины X , определённая при условии, что значения компонент x_{k+1}, \dots, x_n зафиксированы на соответствующих уровнях x_{k+1}^*, \dots, x_n^* , называется *плотностью условного распределения* случайной величины X :

$$\begin{aligned} f(x_1, x_2, \dots, x_k | x_{k+1} = x_{k+1}^*, \dots, x_n = x_n^*) &= \\ &= \frac{f(x_1, x_2, \dots, x_n)}{f_{k+1,\dots,n}(x_{k+1}, \dots, x_n)}. \end{aligned}$$

Случайные величины X_1, X_2, \dots, X_n называются (*стохастически*) *независимыми*, если функция их совместного распределения $F(x_1, x_2, \dots, x_n)$ представима в виде произведения функций распределения случайных величин:

$$F(x_1, x_2, \dots, x_n) = F(x_1)F(x_2) \cdots F(x_n),$$

или, в случае непрерывных случайных величин, аналогичным образом может быть записана их совместная плотность распределения:

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \cdots f(x_n).$$

2.5. Числовые характеристики случайных величин

Описание случайной величины с помощью функции распределения $F(x)$ является исчерпывающим, но для практических задач иногда оказывается излишне подробным. Бывает, что достаточно охарактеризовать конкретное свойство случайной величины с помощью некоторого числа, то есть перейти к её *числовым характеристикам*.

Для характеристики *центра распределения значений* случайной величины используется математическое ожидание. *Математическим ожиданием (ожидаемым средним значением)* дискретной случайной величины называется величина

$$M(X) = \sum_{i=1}^n p_i x_i.$$

Математическое ожидание непрерывной случайной величины, заданной плотностью распределения, вычисляется как

$$M(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

Основные *свойства* математического ожидания:

1. Если $C = \text{const}$, то $M(C) = C$;
2. Если $C = \text{const}$, то $M(CX) = CM(X)$;
3. $M(X + Y) = M(X) + M(Y)$;
4. Если X, Y — некоррелированы, то $M(XY) = M(X)M(Y)$.

Для характеристики *разброса значений* случайной величины *относительно центра* распределения служит дисперсия, определяемая как математическое ожидание квадрата отклонения случайной величины от своего математического ожидания

$$D(X) = M(X - M(X))^2.$$

Можно показать, что верна *универсальная формула дисперсии*

$$D(X) = M(X^2) - M(X)^2.$$

Для нахождения *дисперсии* дискретной случайной величины используют формулу

$$D(X) = \sum_{i=1}^n (x_i - M(X))^2 p_i = \sum_{i=1}^n x_i^2 p_i - M(X)^2.$$

Дисперсия непрерывной случайной величины, заданной плотностью распределения, вычисляется по формуле

$$D(X) = \int_{-\infty}^{\infty} (x - M(X))^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - M(X)^2.$$

Основные свойства дисперсии:

1. Если $C = \text{const}$, то $D(C) = 0$;
2. Если $C = \text{const}$, то $D(CX) = C^2 D(X)$;
3. Если X, Y — некоррелированы, то $D(X + Y) = D(X) + D(Y)$.

Среднее квадратическое (стандартное) отклонение определяется как квадратный корень из дисперсии $\sigma_X = \sqrt{D(X)}$.

Случайную величину V называют *центрированной*, если её математическое ожидание равно нулю $M(V) = 0$. Для центрирования произвольной случайной величины X служит формула $V = X - M(X)$.

Случайную величину W называют *нормированной*, если её дисперсия равна единице $D(W) = 1$. Для нормирования произвольной случайной величины X служит формула $W = \frac{X}{\sigma_X}$.

Случайную величину Z называют *стандартной*, если её математическое ожидание равно нулю $M(Z) = 0$, а дисперсия равна единице $D(Z) = 1$. Для стандартизации произвольной случайной величины X служит формула $Z = \frac{X - M(X)}{\sigma_X}$.

Медианой $x_{\frac{1}{2}}$ называется такое значение случайной величины X , которое делит область её возможных значений на две равновероятные части. Формально, медиана определяется как решение уравнения $F(x_{\frac{1}{2}}) = \frac{1}{2}$.

Обобщая данное уравнение, приходим к понятию *квантиля* x_p уровня p : $F(x_p) = p$. Квантили, делящие область возможных значений случайной величины X на четыре равновероятные части, называются *первым* $x_{\frac{1}{4}}$, *вторым* $x_{\frac{2}{4}}$ и *третьим* $x_{\frac{3}{4}}$ *квантилями*. Легко увидеть, что второй квантиль совпадает с медианой $x_{\frac{2}{4}} = x_{\frac{1}{2}}$.

С геометрической точки зрения квантиль x_p непрерывной случайной величины есть такая точка на оси абсцисс, что площадь криволинейной трапеции, ограниченная графиком плотности распределения $f(x)$ и лежащая левее вертикальной прямой $x = x_p$, будет равна p . С

другой стороны, квантиль x_p по определению является корнем уравнения $F(x_p) = p$, откуда следует, что квантиль — это абсцисса $x = x_p$ точки пересечения прямой $y = p$ с графиком функции распределения $F(x)$.

Для распределений, чья плотность является четной функцией (к примеру, центрированных равномерного и нормального распределений, распределения Стьюдента и тому подобных), квантили уровней $(1 - p)$ и p будут расположены симметрично относительно начала координат, то есть $x_{1-p} = -x_p$.

Мерой взаимосвязи двух случайных величин X и Y может служить *коэффициент ковариации*, определяемый по формуле

$$\begin{aligned}\operatorname{cov}(X, Y) = \sigma_{XY} &= M((X - M(X))(Y - M(Y))) = \\ &= M(XY) - M(X)M(Y).\end{aligned}$$

Основным *свойством* коэффициента ковариации σ_{XY} является его равенство нулю для независимых случайных величин X и Y . Заметим, что обратное утверждение, вообще говоря, неверно.

Зависимость величины σ_{XY} от масштаба изучаемых величин и делает неудобным её использование в практических приложениях. Поэтому для измерения связи между X и Y обычно используют другую числовую характеристику ρ_{XY} , называемую *коэффициентом корреляции*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Наиболее существенными являются следующие *свойства* коэффициента корреляции:

1. Коэффициент корреляции *симметричен*: $\rho_{XY} = \rho_{YX}$;
2. Модуль коэффициента корреляции *не превосходит единицы*: $|\rho_{XY}| \leq 1$;
3. Модуль коэффициента корреляции *равен единице* $|\rho_{XY}| = 1$ только в том случае, когда случайные величины X и Y связаны *линейной зависимостью*;
4. Если случайные величины X и Y *независимы*, то $\rho_{XY} = 0$, а если $\rho_{XY} = 0$, то говорят о *некоррелированности* случайных величин X и Y ;
5. Величина коэффициента корреляции ρ_{XY} *инвариантна* относительно линейных преобразований.

В случае *многомерных случайных величин* X в рассмотрение вводятся многомерные аналоги числовых характеристик.

Для случайного вектора $X = (X_1, X_2, \dots, X_n)^T$ характеристикой центра группирования будет *вектор средних значений*

$$M(X) = (M(X_1), M(X_2), \dots, M(X_n))^T.$$

В качестве меры рассеяния компонент и их взаимосвязи используется *матрица ковариаций*:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix},$$

где $\sigma_{ij} = \text{cov}(X_i, X_j)$ при $i, j = 1, 2, \dots, n$. Определитель этой матрицы $\det \Sigma$ называется *обобщённой дисперсией*.

По причинам, указанным выше, в практических приложениях чаще используется так называемая *корреляционная матрица*:

$$R = \begin{pmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \dots & \rho_{nn} \end{pmatrix},$$

где $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$ при $i, j = 1, 2, \dots, n$; $\sigma_i = \sqrt{D(X_i)}$; $\sigma_j = \sqrt{D(X_j)}$.

2.6. Наиболее распространённые распределения

2.6.1. Биномиальное распределение

Дискретная случайная величина X имеет *биномиальное распределение* с параметрами $n \in \mathbf{Z}^+$, p : $X \sim \mathcal{B}(n, p)$, если она принимает целочисленные значения $k = 0, 1, \dots, n$ с вероятностями, определяемыми формулой Бернулли

$$p_k = P\{X = k\} = C_n^k p^k q^{n-k},$$

где $C_n^k = \frac{n!}{k!(n-k)!}$; $p \in (0, 1)$; $q = 1 - p$.

Биномиальное распределение возникает в последовательности из n независимых испытаний с постоянной вероятностью успеха в каждом испытании $p = \text{const}$ и полностью определяется значениями параметров n и p :

$$X \sim \begin{pmatrix} 0 & 1 & \dots & k & \dots & n \\ q^n & C_n^1 p^1 q^{n-1} & \dots & C_n^k p^k q^{n-k} & \dots & p^n \end{pmatrix}.$$

Функция распределения случайной величины, подчиняющейся биномиальному закону $X \sim \mathcal{B}(n, p)$, имеет вид:

$$F(x) = \begin{cases} 0, & \text{если } x \leq 0; \\ \sum_{k \leq x} C_n^k p^k q^{n-k}, & \text{если } 0 < x \leq n; \\ 1, & \text{если } x > n. \end{cases}$$

Математическое ожидание и дисперсия случайной величины, подчиняющейся биномиальному закону $X \sim \mathcal{B}(n, p)$, вычисляются по формулам:

$$M(X) = np, \quad D(X) = npq.$$

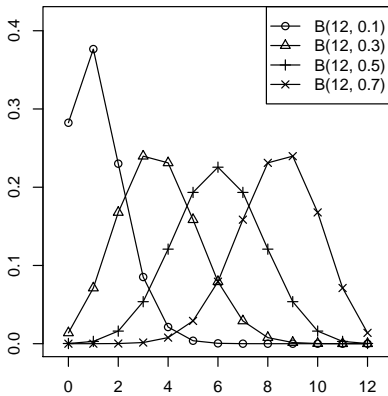


Рис. 2.1. Биномиальное распределение вероятностей p_k

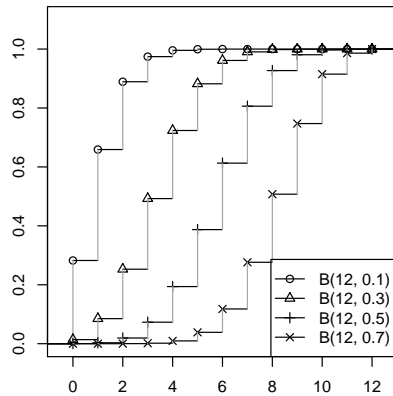


Рис. 2.2. Функция биномиального распределения $F(x)$

На рис. 2.1 и 2.2 показаны примеры построения графиков распределения вероятностей p_k и функции распределения $F(x)$ биномиально распределённой случайной величины $X \sim \mathcal{B}(n, p)$ при $n = 12$ и p , принимающей последовательные значения от $\frac{1}{10}$ до $\frac{7}{10}$ через $\frac{2}{10}$, то есть $p \in \{\frac{1}{10}, \frac{3}{10}, \frac{5}{10}, \frac{7}{10}\}$.

Пример 2.1. В качестве примера построим вышеприведённые графики вероятностей p_k и функции распределения $F(x)$ биномиально распределённой случайной величины $X \sim \mathcal{B}(n, p)$ с помощью R.


```

1 > source("probGraph.r")
2 > p <- seq(1, 7, 2)/10
3 > n <- 12; x <- seq(0, n)
4 > P <- sapply(p, function(pp) dbinom(x, n, pp))
5 > F <- sapply(p, function(pp) pbinom(x, n, pp))
6 > l <- sapply(p, function(pp) sprintf("B(%.0f, %.3g)", n, pp))
7 > dgraph(x, P, l)
8 > pgraph(x, F, l)

```

Команда «source("probGraph.r")» в строке [1] производит загрузку исходного кода библиотеки, содержащей функции для построения графиков по теории вероятностей.

Функция «seq()» в строках [2–3] генерирует вектор последовательных значений от первого до второго аргумента; третий аргумент функции позволяет указать приращение в последовательности значений, равное по-умолчанию ± 1 .

Функция «sapply(p, ...)» производит подстановку каждой компоненты вектора «p» в указанную далее функцию. Таким образом, в строках [4], [5] с помощью функций «dbinom()» и «pbinom()» по вектору абсцисс «x» вычисляются ординаты вероятности «P» и функции биномиального распределения «F» для каждой пары параметров «n, p», а в строке [6] значения этих параметров формируют поясняющие надписи на графиках.

Функции «dgraph()» и «pgraph()» определены в пользовательской библиотеке «probGraph.r» и производят построение графиков вероятностей и функций распределения дискретной случайной величины по переданным векторам абсцисс «x» и ординат «P» или «F». Полный текст исходного кода библиотеки «probGraph.r» приведён в Приложении B.1.

2.6.2. Распределение Пуассона

Дискретная случайная величина X имеет *распределение Пуассона* с параметром $\lambda > 0$: $X \sim \mathcal{P}(\lambda)$, если она принимает целочисленные значения $k = 0, 1, \dots, \infty$ с вероятностями, определяемыми формулой Пуассона

$$p_k = \mathbf{P}\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda},$$

где $\lambda > 0$.

Распределение Пуассона является предельным случаем биномиального распределения при $n \rightarrow \infty$, а $p \rightarrow 0$ так, что $\lambda = np = \text{const}$.

Оно возникает при рассмотрении единичных независимых случайных событий с постоянной интенсивностью λ и полностью определяется её значением

$$X \sim \begin{pmatrix} 0 & 1 & 2 & \dots \\ \frac{1}{e^\lambda} & \frac{\lambda^1}{e^\lambda 1!} & \frac{\lambda^2}{e^\lambda 2!} & \dots \end{pmatrix}.$$

Функция распределения случайной величины, подчиняющейся закону Пуассона $X \sim \mathcal{P}(\lambda)$, имеет вид:

$$F(x) = \begin{cases} 0, & \text{если } x \leq 0; \\ \sum_{k < x} \frac{\lambda^k}{k!} e^{-\lambda}, & \text{если } x > 0. \end{cases}$$

Математическое ожидание и дисперсия случайной величины, подчиняющейся закону Пуассона $X \sim \mathcal{P}(\lambda)$, вычисляются по формулам:

$$M(X) = D(X) = \lambda = np.$$

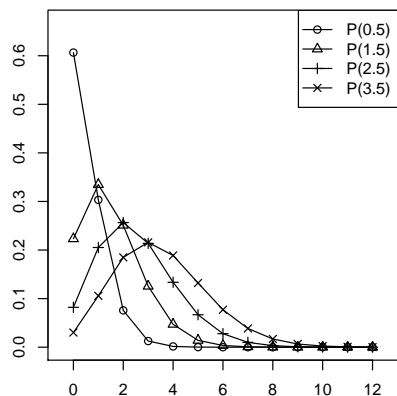


Рис. 2.3. Пуассоновское распределение вероятностей p_k

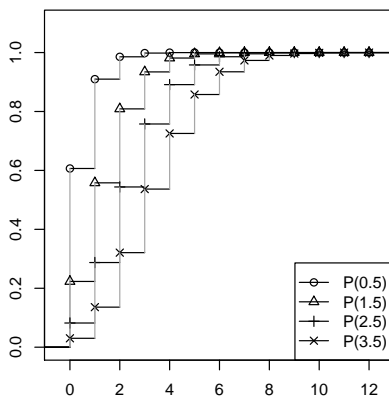


Рис. 2.4. Функция пуассоновского распределения $F(x)$

На рис. 2.3 и 2.4 показаны примеры построения графиков вероятностей p_k и функции распределения $F(x)$ пуассоновской случайной величины $X \sim \mathcal{P}(\lambda)$ при $\lambda \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}\}$.

Пример 2.2. Продолжая предыдущий пример, построим вышеприведённые графики вероятностей p_k и функции распределения $F(x)$ для пуассоновской случайной величины $X \sim \mathcal{P}(\lambda)$ с помощью **R**.

```

9 > a <- seq(0.5, 3.5, 1)
10 > P <- sapply(a, function(aa) dpois(x, aa))
11 > F <- sapply(a, function(aa) ppois(x, aa))
12 > l <- sapply(a, function(aa) sprintf("P(%.3g)", aa))
13 > dgraph(x, P, l)
14 > pgraph(x, F, l)

```

2.6.3. Геометрическое распределение

Дискретная случайная величина X имеет *геометрическое распределение* с параметром p : $X \sim \mathcal{G}(p)$, если она принимает целочисленные значения $k = 0, 1, \dots, \infty$ с вероятностями, определяемыми формулой

$$p_k = \mathbf{P}\{X = k\} = q^k p,$$

где $p \in (0, 1)$; $q = 1 - p$.

Геометрическое распределение имеет случайная величина X , равная числу испытаний в последовательности Бернулли, проходящих до появления первого успеха. Геометрическое распределение полностью определяется значениями параметра p :

$$X \sim \begin{pmatrix} 0 & 1 & 2 & \dots \\ p & qp & q^2p & \dots \end{pmatrix}.$$

Функция распределения случайной величины X , подчиняющейся геометрическому закону $X \sim \mathcal{G}(p)$, имеет вид

$$F(x) = \begin{cases} 0, & \text{если } x \leq 0; \\ \sum_{k < x} q^k p, & \text{если } x > 0. \end{cases}$$

Математическое ожидание и дисперсия случайной величины X , подчиняющейся геометрическому закону $X \sim \mathcal{G}(p)$, вычисляются по формулам:

$$\mathbf{M}(X) = \frac{1}{p}, \quad \mathbf{D}(X) = \frac{q}{p^2}.$$

На рис. 2.5 и 2.6 показаны примеры построения графиков распределения вероятностей (k, p_k) и функции распределения $F(x)$ геометрически распределённой случайной величины $X \sim \mathcal{G}(p)$ при $p \in \{\frac{3}{10}, \frac{5}{10}, \frac{7}{10}, \frac{9}{10}\}$.

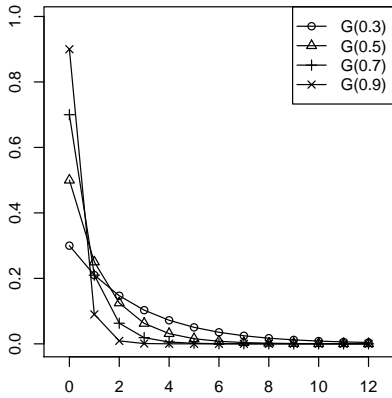


Рис. 2.5. Геометрическое распределение вероятностей p_k

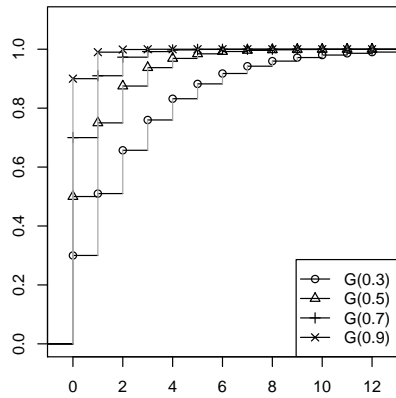


Рис. 2.6. Функция геометрического распределения $F(x)$

Пример 2.3. Продолжая предыдущий пример, построим вышеприведённые графики вероятностей p_k и функции распределения $F(x)$ для геометрически распределённой случайной величины $X \sim \mathcal{G}(p)$ с помощью **R**.

```

15 > p <- seq(3, 9, 2)/10
16 > P <- sapply(p, function(pp) dgeom(x, pp))
17 > F <- sapply(p, function(pp) pgeom(x, pp))
18 > l <- sapply(p, function(pp) sprintf("G(%.3g)", pp))
19 > dgraph(x, P, l)
20 > pgraph(x, F, l)

```

2.6.4. Равномерное распределение

Простейшим из непрерывных распределений является равномерное распределение, возникающее при обобщении понятия n равновероятных случайных событий на случай $n \rightarrow \infty$. Непрерывная случайная величина X имеет *равномерное распределение* на отрезке $[a, b]$: $X \sim \mathcal{U}(a, b)$, если её плотность вероятности постоянна и отлична от нуля только на этом отрезке:

$$f(x) = \begin{cases} 0, & \text{если } x \notin [a, b]; \\ \frac{1}{b-a}, & \text{если } x \in [a, b]. \end{cases}$$

Равномерное распределение полностью определяется координатами концов отрезка $[a, b]$. Функция распределения случайной величины, подчиняющейся равномерному закону $X \sim \mathcal{U}(a, b)$, имеет вид:

$$F(x) = \begin{cases} 0, & \text{если } x \leq a; \\ \frac{x-a}{b-a}, & \text{если } a < x \leq b; \\ 1, & \text{если } x > b. \end{cases}$$

Математическое ожидание и дисперсия равномерно распределённой случайной величины $X \sim \mathcal{U}(a, b)$ вычисляются по формулам:

$$M(X) = \frac{a+b}{2}, \quad D(X) = \frac{(b-a)^2}{12}.$$

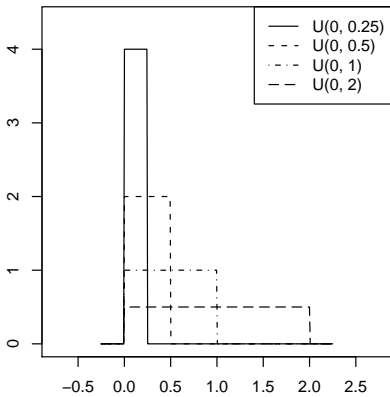


Рис. 2.7. Плотность равномерного распределения $f(x)$

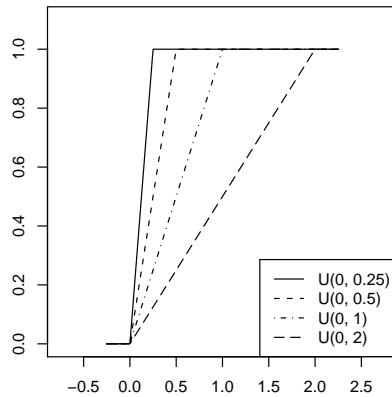


Рис. 2.8. Функция равномерного распределения $F(x)$

На рис. 2.7 и 2.8 показаны примеры построения графиков плотности вероятности $f(x)$ и функции распределения $F(x)$ равномерно распределённой случайной величины $X \sim \mathcal{U}(a, b)$ при значениях параметров: $a = 0$, $b \in \{\frac{1}{4}, \frac{1}{2}, 1, 2\}$.

Пример 2.4. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для равномерно распределённой случайной величины $X \sim \mathcal{U}(a, b)$ с помощью **R**.

```

21 > a <- 0; b <- c(1/4, 1/2, 1, 2)
22 > x <- seq(a-1/4, max(b)+1/4, len=300)
23 > f <- sapply(b, function(bb) dunif(x, a, bb))
24 > F <- sapply(b, function(bb) punif(x, a, bb))
25 > l <- sapply(b, function(bb) sprintf("U(%.3g, %.3g)", a, bb))
26 > cgraph(x, f, l)
27 > fgraph(x, F, l)

```

2.6.5. Показательное распределение

Показательное распределение возникает при моделировании *времени между* последовательными реализациями одного и того же случайного события. Непрерывная случайная величина X имеет показательное распределение: $X \sim \mathcal{E}(\lambda)$, если её плотность вероятности имеет вид:

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

где $\lambda > 0$ — параметр, интерпретируемый как среднее число случайных событий в единицу времени.

Функция распределения показательно распределённой случайной величины: $X \sim \mathcal{E}(\lambda)$ имеет вид:

$$F(x) = \begin{cases} 0, & \text{если } x < 0, \\ 1 - e^{-\lambda x}, & \text{если } x \geq 0. \end{cases}$$

Математическое ожидание и дисперсия показательно распределённой случайной величины $X \sim \mathcal{E}(\lambda)$ вычисляются по формулам:

$$M(X) = \frac{1}{\lambda}, \quad D(X) = \frac{1}{\lambda^2}.$$

На рис. 2.9 и 2.10 показаны примеры построения графиков плотности вероятности $f(x)$ и функции распределения $F(x)$ показательно распределённой случайной величины $X \sim \mathcal{E}(\lambda)$ при значениях параметра: $\lambda \in \{\frac{1}{4}, \frac{1}{2}, 1, 2\}$.

Пример 2.5. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для показательно распределённой случайной величины $X \sim \mathcal{E}(\lambda)$ с помощью **R**.

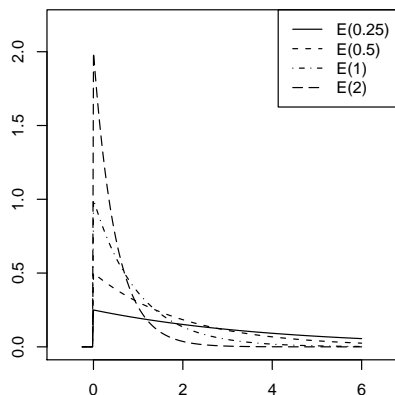


Рис. 2.9. Плотность показательного распределения $f(x)$

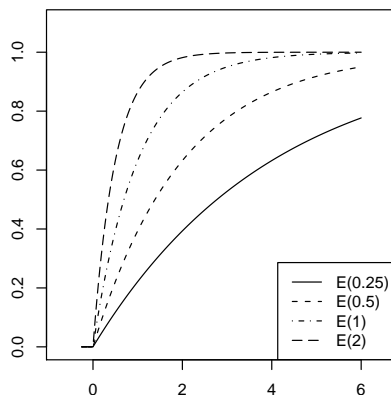


Рис. 2.10. Функция показательного распределения $F(x)$

```

28 > a <- c(1/4, 1/2, 1, 2)
29 > x <- seq(0, 1/min(a), len=300)
30 > f <- sapply(a, function(aa) dexp(x, aa))
31 > F <- sapply(a, function(aa) pexp(x, aa))
32 > l <- sapply(a, function(aa) sprintf("E(%.3g)", aa))
33 > cgraph(x, f, l)
34 > fgraph(x, F, l)

```

2.6.6. Нормальное распределение

Нормальное распределение обычно возникает при рассмотрении суммы большого количества независимо распределённых случайных величин с конечной дисперсией. Непрерывная случайная величина X имеет *нормальное распределение*: $X \sim \mathcal{N}(a, \sigma)$, если её плотность вероятности имеет вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} = \varphi\left(\frac{x-a}{\sigma}\right),$$

где $x, a \in \mathbf{R}$; $\sigma > 0$; $\varphi(z)$ — функция Гаусса, определяемая равенством

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Нормальное распределение полностью определяется параметрами a и σ . Функция распределения случайной величины, подчиняющейся нормальному закону $X \sim \mathcal{N}(a, \sigma)$, имеет вид:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-a)^2}{2\sigma^2}} dt = \frac{1}{2} + \Phi\left(\frac{x-a}{\sigma}\right),$$

где $\Phi(z)$ — функция Лапласа, определяемая равенством

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{y^2}{2}} dy.$$

Математическое ожидание и дисперсия нормально распределённой случайной величины $X \sim \mathcal{N}(a, \sigma)$ вычисляются по формулам:

$$M(X) = a, \quad D(X) = \sigma^2.$$

Свойства нормального распределения:

1. Если $Y = \alpha X + \beta$, где $\alpha, \beta \in \mathbf{R}$, а случайная величина $X \sim \mathcal{N}(a, \sigma)$, то случайная величина $Y \sim \mathcal{N}(\alpha a + \beta, \alpha \sigma)$;
2. Если $X_i \sim \mathcal{N}(a_i, \sigma_i)$, при $i = 1, 2, \dots, n$ — независимые случайные величины, то $Y = \sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n a_i, \sqrt{\sum_{i=1}^n \sigma_i^2}\right)$;
3. Если $X_i \sim \mathcal{N}(a_i, \sigma_i)$, при $i = 1, 2, \dots, n$ — зависимые случайные величины, то $Y = \sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n a_i, \sqrt{\sum_{i=1}^n \sigma_i^2 + \sum_{i < j} \rho_{ij} \sigma_i \sigma_j}\right)$.

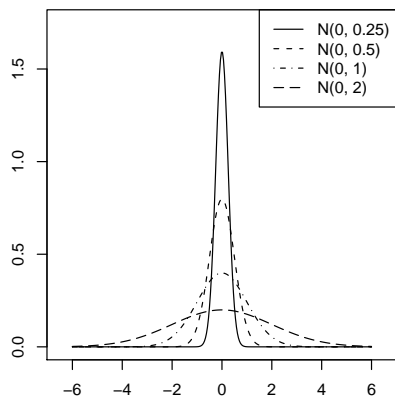
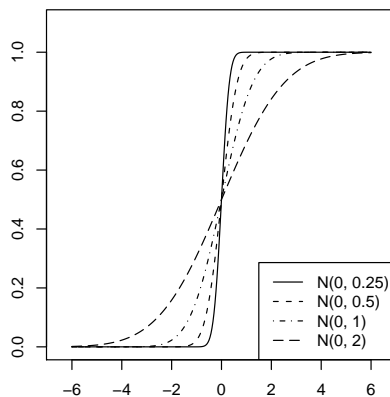
На рис. 2.11 и 2.12 показаны примеры построения графиков плотности вероятности $f(x)$ и функции распределения $F(x)$ нормально распределённой случайной величины $X \sim \mathcal{N}(a, \sigma)$ при значениях параметров: $a = 0$, $\sigma \in \{\frac{1}{4}, \frac{1}{2}, 1, 2\}$.

Пример 2.6. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для нормально распределённой случайной величины $X \sim \mathcal{N}(a, \sigma)$ с помощью **R**.

```

35 > a <- 0; s <- c(1/4, 1/2, 1, 2)
36 > x <- seq(a-3*max(s), a+3*max(s), len=300)

```


Рис. 2.11. Плотность нормального распределения $f(x)$ Рис. 2.12. Функция нормального распределения $F(x)$

```

37 > f <- sapply(s, function(ss) dnorm(x, a, ss))
38 > F <- sapply(s, function(ss) pnorm(x, a, ss))
39 > l <- sapply(s, function(ss) sprintf("N(%.3g, %.3g)", a, ss))
40 > cgraph(x, f, l)
41 > fgraph(x, F, l)

```

2.6.7. Логнормальное распределение

Непрерывная случайная величина X имеет *логарифмически нормальное* или *логнормальное распределение*, если её логарифм нормально распределён. Подобно нормальному распределению логнормальное возникает при рассмотрении *произведения* большого числа независимых случайных величин с конечной дисперсией. Плотность вероятности логарифмически нормального распределения имеет вид:

$$f(x) = \begin{cases} 0, & \text{если } x \leq 0; \\ \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - a)^2}{2\sigma^2}}, & \text{если } x > 0, \end{cases}$$

где $x \in \mathbf{R}$; $x, \sigma > 0$.

Логарифмически нормальное распределение полностью определяется параметрами a и σ . Функция распределения логарифмически нормальной случайной величины $X \sim \ln \mathcal{N}(a, \sigma)$ имеет вид:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x t^{-1} e^{-\frac{(\ln t - a)^2}{2\sigma^2}} dt = \Phi\left(\frac{\ln x - a}{\sigma}\right) + \frac{1}{2},$$

где $\Phi(x)$ — функция Лапласа.

Математическое ожидание и дисперсия логарифмически нормальной случайной величины $X \sim \ln \mathcal{N}(a, \sigma)$ зависят:

$$M(X) = e^{\frac{2a+\sigma^2}{2}}, \quad D(X) = (e^{\sigma^2} - 1)e^{2a+\sigma^2} = (e^{\sigma^2} - 1)M(X)^2.$$

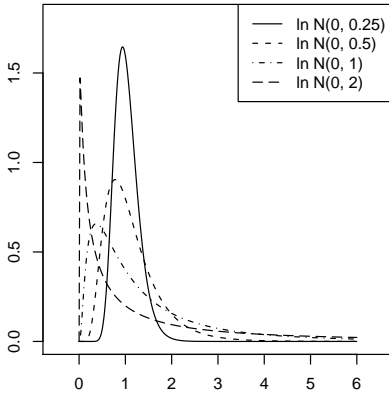


Рис. 2.13. Плотность логарифмически нормального распределения $f(x)$

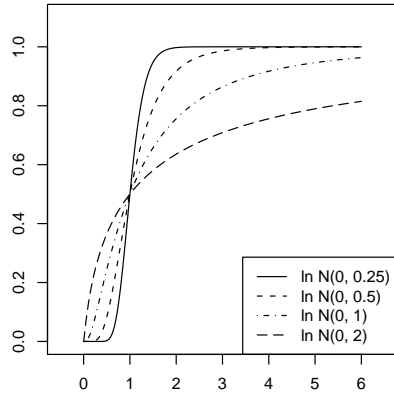


Рис. 2.14. Функция логарифмически нормального распределения $F(x)$

На рис. 2.13 и 2.14 показаны примеры построения графиков плотности вероятности $f(x)$ и функции распределения $F(x)$ логарифмически нормально распределённой случайной величины $X \sim \ln \mathcal{N}(a, \sigma)$ при значениях параметров: $a = 0$, $\sigma \in \{\frac{1}{4}, \frac{1}{2}, 1, 2\}$.

Пример 2.7. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для логарифмически нормально распределённой случайной величины $X \sim \ln \mathcal{N}(a, \sigma)$ с помощью **R**.

```
42 > a <- 0; s <- c(1/4, 1/2, 1, 2)
43 > x <- seq(a, a+3*max(s), len=300)
44 > f <- sapply(s, function(ss) dlnorm(x, a, ss))
```

```

45 > F <- sapply(s, function(ss) plnorm(x, a, ss))
46 > l <- sapply(s, function(ss) sprintf("ln N(%.3g, %.3g)", a, ss))
47 > cgraph(x, f, l)
48 > fgraph(x, F, l)

```

2.6.8. Пирсона χ^2 -распределение

Если $X_i \sim \mathcal{N}(0, 1)$, где $i = 1, 2, \dots, n$ — независимые стандартные нормальные случайные величины, то сумма n квадратов этих величин имеет χ^2 -распределение (Пирсона) с n степенями свободы:

$$\chi_n^2 = \sum_{i=1}^n X_i^2.$$

Плотность распределения χ^2 выражается формулой:

$$f(x) = \begin{cases} 0, & \text{если } x \leq 0; \\ \frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n-2}{2}} e^{-\frac{x}{2}}, & \text{если } 0 < x; \end{cases} \quad \Gamma(p) = \int_0^{\infty} t^{p-1} e^{-t} dt,$$

где $\Gamma(p)$ — гамма-функция Эйлера. При возрастании числа степеней свободы $n \rightarrow \infty$ распределение χ_n^2 асимптотически нормально.

Математическое ожидание и дисперсия распределения χ_n^2 имеют вид:

$$M(\chi_n^2) = n, \quad D(\chi_n^2) = 2n.$$

На рис. 2.15 и 2.16 показаны примеры построения графиков плотности вероятности $f(x)$ и функции распределения $F(x)$ случайной величины $X \sim \chi_n^2$ при числе степеней свободы: $n \in \{2, 3, 4, 5\}$.

Пример 2.8. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для случайной величины $X \sim \chi_n^2$ с помощью **R**.

```

49 > k <- c(2, 3, 4, 5)
50 > x <- seq(0, n, len=300)
51 > f <- sapply(k, function(kk) dchisq(x, kk))
52 > F <- sapply(k, function(kk) pchisq(x, kk))
53 > l <- sapply(k, function(kk) sprintf("chi^2(%.0f)", kk))
54 > cgraph(x, f, l)
55 > fgraph(x, F, l)

```

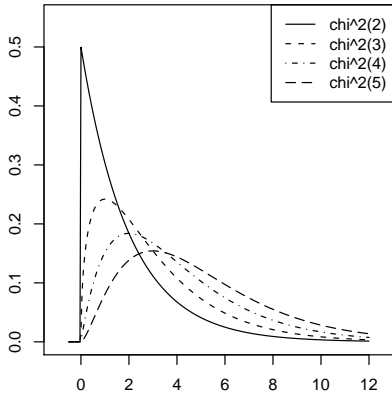


Рис. 2.15. Плотность χ^2 -распределения $f(x)$

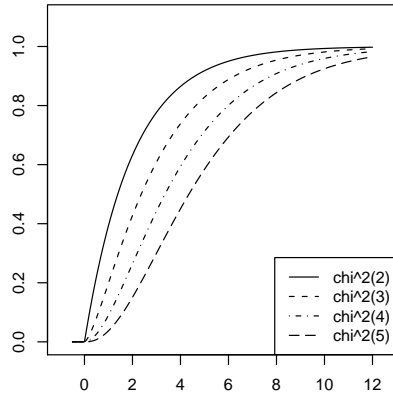


Рис. 2.16. Функция χ^2 -распределения $F(x)$

2.6.9. Стьюдента t -распределение

Если случайные величины $Z \sim \mathcal{N}(0, 1)$ и $U \sim \chi_n^2$ — независимы, то случайная величина

$$t_n = \frac{Z}{\sqrt{\frac{U}{n}}}$$

имеет распределение Стьюдента или t -распределение с n степенями свободы.

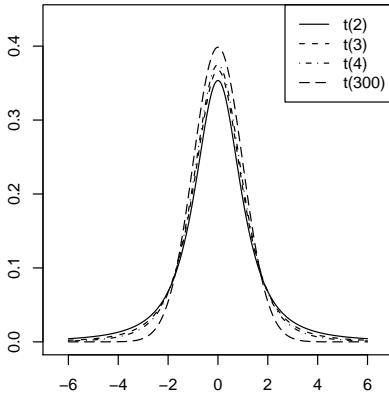
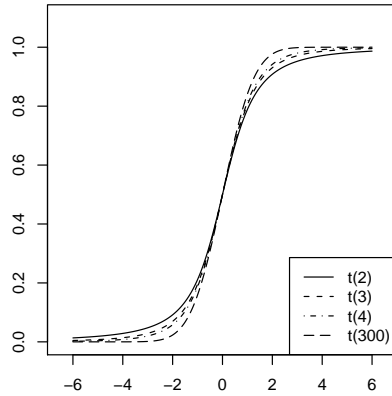
Плотность t -распределения имеет вид:

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

где $x \in \mathbf{R}$; $\Gamma(p)$ — гамма-функция Эйлера. При возрастании числа степеней свободы $n \rightarrow \infty$ распределение Стьюдента асимптотически нормально.

Математическое ожидание и дисперсия t -распределения выражаются формулами:

$$\mathbf{M}(t_n) = 0, \quad \mathbf{D}(t_n) = \begin{cases} \infty, & \text{если } 1 < n \leq 2; \\ \frac{n}{n-2}, & \text{если } n > 2. \end{cases}$$

Рис. 2.17. Плотность t -распределения $f(x)$ Рис. 2.18. Функция t -распределения $F(x)$

На рис. 2.17 и 2.18 показаны примеры построения графиков плотности вероятности $f(x)$ и функции распределения $F(x)$ случайной величины $X \sim t_n$ при числе степеней свободы: $n \in \{2, 3, 4, 300\}$.

Пример 2.9. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ случайной величины $X \sim t_n$ с помощью **R**.

```

56 > k <- c(2, 3, 4, 300)
57 > x <- seq(-6, 6, len=300)
58 > f <- sapply(k, function(kk) dt(x, kk))
59 > F <- sapply(k, function(kk) pt(x, kk))
60 > l <- sapply(k, function(kk) sprintf("t(%.0f)", kk))
61 > cgraph(x, f, l)
62 > fgraph(x, F, l)

```

2.6.10. Фишера F -распределение

Если случайные величины $U \sim \chi_m^2$ и $V \sim \chi_n^2$ — независимы, то случайная величина

$$F_{\frac{m}{n}} = \frac{\frac{1}{m}U}{\frac{1}{n}V}$$

имеет распределение Фишера или F -распределение со степенями свободы числителя m и знаменателя n ¹. Плотность F -распределения:

$$f(x) = \begin{cases} 0, & \text{если } x \leq 0; \\ \frac{\sqrt{\frac{(mx)^{m-1}}{(mx+n)^{m+n}}} \cdot \frac{1}{x B(\frac{m}{2}, \frac{n}{2})}}, & \text{если } x > 0; \end{cases} \quad B(u, v) = \int_0^1 t^{u-1} (1-t)^{v-1} dt,$$

где $m, n > 0$; $B(u, v)$ — бета-функция Эйлера. При возрастании числа степеней свободы $n \rightarrow \infty$ распределение Фишера асимптотически нормально.

Математическое ожидание и дисперсия F -распределения выражаются формулами:

$$M(F_{\frac{m}{n>2}}) = \frac{n}{n-2}, \quad D(F_{\frac{m}{n>4}}) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}.$$

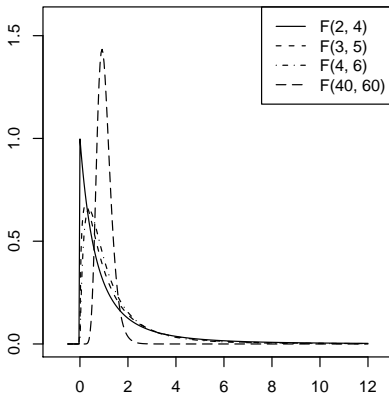


Рис. 2.19. Плотность F -распределения $f(x)$

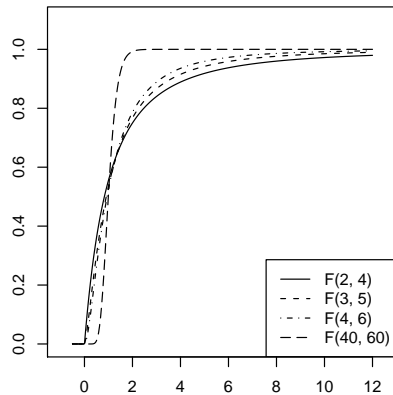


Рис. 2.20. Функция F -распределения $F(x)$

На рис. 2.19 и 2.20 показаны примеры построения графиков плотности вероятности $f(x)$ и функции распределения $F(x)$ случайной величины $X \sim F_{\frac{m}{n}}$ при значении чисел степеней свободы: $m \in \{2, 3, 4, 40\}$, $n \in \{4, 5, 6, 60\}$.

¹ Используемое в настоящем пособии обозначение $F_{\frac{m}{n}}$ для распределения Фишера со степенями свободы числителя m и знаменателя n не является общепринятым, но по мнению авторов оно порождает меньше двусмысленностей, по сравнению с обычно применяемым $F_{m,n}$.

Пример 2.10. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ случайной величины $X \sim F_m^n$ с помощью **R**.

```
63 > k1 <- c(2, 3, 4, 40); k2 <- c(4, 5, 6, 60); k <- seq(4)
64 > x <- seq(0, 12, len=300)
65 > f <- sapply(k, function(kk) df(x, k1[kk], k2[kk]))
66 > F <- sapply(k, function(kk) pf(x, k1[kk], k2[kk]))
67 > l <- sapply(k, function(kk) sprintf("F(%.0f, %.0f)",
68 +   k1[kk], k2[kk]))
69 > cgraph(x, f, l)
70 > fgraph(x, F, l)
```

Контрольные вопросы

1. Что называется случайным событием? Дайте определения достоверного и невозможного событий.
2. Какие события называются: несовместными, равновозможными и противоположными?
3. Что называют пространством элементарных исходов?
4. Дайте определение суммы событий. Приведите примеры сумм событий.
5. Дайте определение произведения событий. Приведите примеры произведения двух событий.
6. Сформулируйте теоремы сложения вероятностей для совместных и несовместных событий.
7. Сформулируйте определение зависимых и независимых событий. Приведите формулы умножения вероятностей для зависимых и независимых событий.
8. Дайте определение условной вероятности. Сформулируйте теорему о полной вероятности и запишите формулу Байеса.
9. Дайте определение случайной величины и закона её распределения. Перечислите типы случайных величин. Что называют рядом распределения дискретной случайной величины?
10. Дайте определение математического ожидания для дискретной и непрерывной случайных величин. Перечислите свойства математического ожидания.
11. Дайте определение дисперсии и среднего квадратического отклонения случайной величины. Перечислите свойства дисперсии и среднего квадратического отклонения.

12. Дайте определение плотности распределения случайной величины. Укажите основные свойства функции плотности распределения.
13. Как определяется система двух случайных величин (двумерная случайная величина). Как определяется закон распределения двумерной случайной величины.
14. Приведите определения условного математического ожидания и дисперсии случайной величины. Перечислите их свойства.
15. Дайте определение числовых характеристик системы случайных величин: ковариации и коэффициента корреляции. Сформулируйте их свойства.
16. Какие случайные величины называют независимыми и некоррелированными?
17. Сформулируйте законы распределения дискретных случайных величин: биномиальный, геометрический и распределения Пуассона? Как найти числовые характеристики этих распределений?
18. Сформулируйте законы распределения непрерывных случайных величин: равномерный, показательный? Как найти числовые характеристики этих распределений?
19. Какие случайные величины называют нормально и логнормально распределёнными? Как найти числовые характеристики этих распределений?
20. Дайте определение функции случайных величин. Приведите примеры законов распределения функций случайных величин, зависящих от нормального: χ^2 -распределения Пирсона, t -распределения Стьюдента и F -распределения Фишера.

Глава 3

Основы математической статистики

Математическая статистика изучает *методы оценивания и сравнения распределений* случайных величин и их характеристик по наблюдаемым значениям. *Первая задача* математической статистики состоит в упорядочении и представлении наблюдаемых значений в виде, удобном для анализа. *Вторая задача* заключается в оценке, хотя бы приближительной, характеристик и параметров распределений наблюдаемой случайной величины. *Третьей задачей* математической статистики является решение вопроса о согласовании результатов оценивания с наблюдаемыми значениями, то есть проверка статистических гипотез.

3.1. Генеральная и выборочная совокупности

В математической статистике исследуемую случайную величину X , в общем случае — многомерную, принято называть *генеральной совокупностью*, а её реализации в последовательности независимых испытаний $\{x_1, x_2, \dots, x_n\}$ — *выборочной совокупностью* или *случайной выборкой*. Составляющие выборку случайные величины x_i называют *элементами выборки*, а их количество n — *объёмом выборки*.

Основной задачей статистического исследования является описание генеральной совокупности X по имеющейся случайной выборке $\{x_i\}$, $i = 1, 2, \dots, n$. Как правило, эта задача сводится к нахождению закона распределения случайной величины $X \sim F(x)$ и определению её числовых характеристик.

Статистикой называется любая функция элементов случайной выборки $g(x_1, x_2, \dots, x_n)$. Очевидно, что если рассматривать элементы выборки как независимые одинаково распределённые случайные величины $x_i \sim X \sim F(x)$, то и статистика будет случайной величи-

ной, имеющей свой закон распределения $g(x_1, x_2, \dots, x_n) \sim G(x)$.

Элементы выборки, упорядоченные по неубыванию, называются *вариационным рядом* $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$:

$$\min(x_i) = x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} = \max(x_i).$$

3.2. Выборочные характеристики и точечные оценки

Любые характеристики случайной величины X , полученные по её выборке $\{x_1, x_2, \dots, x_n\}$, называются *выборочными* или *эмпирическими*. *Статистической оценкой* называется выборочная характеристика, используемая в качестве приближённого значения неизвестной характеристики генеральной совокупности.

Статистическая оценка, представленная в виде числа (точки на числовой прямой), называется *точечной*. Тогда практическая применимость точечной оценки определяется такими её свойствами как *несмещённость*, *состоятельность* и *эффективность*.

Пусть $\{x_1, x_2, \dots, x_n\}$ — случайная выборка, тогда $\theta_n(x_1, x_2, \dots, x_n)$ — выборочная оценка некоторого параметра θ . Оценка θ_n называется *несмещённой*, если для любого фиксированного n верно, что $M(\theta_n) = \theta$. Это свойство гарантирует, что использование несмещённой оценки не порождает систематических ошибок.

Оценка θ_n называется *состоятельной*, если она *сходится по вероятности* к истинному значению параметра θ , то есть для любого $\varepsilon > 0$ выполняется условие

$$\lim_{n \rightarrow \infty} P\{|\theta_n - \theta| < \varepsilon\} = 1 \quad \text{или кратко} \quad \theta_n \xrightarrow[n \rightarrow \infty]{P} \theta.$$

Выполнение этого условия означает, что с увеличением объёма выборки n возрастает наша уверенность в малом по абсолютной величине отклонении оценки θ_n от истинного значения параметра θ .

Оценка θ_n называется *эффективной*, если она обладает наименьшей дисперсией, а значит и средним квадратическим отклонением от истинного значения параметра θ , по сравнению с любыми другими оценками данного класса.

Так, несмещённой и состоятельной оценкой вероятности появления значения x_k является его *относительная частота* w_k , а несмещёнными и состоятельными оценками для математического ожидания $M(X)$ и дисперсии $D(X)$ являются *выборочное среднее* \bar{x} и *исправленная выборочная дисперсия* s_x^2 :

$$w_k \xrightarrow[n \rightarrow \infty]{\mathbf{P}} p_k, \quad \bar{x} \xrightarrow[n \rightarrow \infty]{\mathbf{P}} \mathbf{M}(X), \quad s_x^2 \xrightarrow[n \rightarrow \infty]{\mathbf{P}} \mathbf{D}(X);$$

$$w_k = \frac{n_k}{n}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где $p_k = \mathbf{P}\{X = x_k\}$, n_k — вероятность и частота появления значения x_k дискретной случайной величины X .

Несмещённой и состоятельной оценкой коэффициента ковариации σ_{XY} случайных величин X и Y является *выборочная ковариация* s_{xy} , определяемая по формуле

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

где \bar{x} и \bar{y} — выборочные средние случайных величин x и y , соответственно.

Несмещённой и состоятельной оценкой коэффициента корреляции ρ_{XY} случайных величин X и Y является *выборочный коэффициент корреляции* r_{xy} , определяемый по формуле

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Оценками функций распределения $F(x)$ и плотности вероятности $f(x)$ непрерывной случайной величины X будут построенные по её выборке *эмпирическая функция распределения* $F_n(x)$ и *гистограмма* $f_{n,h}(x)$:

$$F_n(x) \xrightarrow[n \rightarrow \infty]{\mathbf{P}} F(x), \quad f_{n,h}(x) \xrightarrow[nh \rightarrow \infty, h \rightarrow 0]{\mathbf{P}} f(x);$$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(-\infty, x)}(x_i), \quad f_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}_{[kh, (k+1)h)}(x_i),$$

где точки вокруг обозначения вероятности \mathbf{P} указывают на поточечную сходимость по вероятности гистограммы $f_{n,h}(x)$ к функции плотности вероятности при выполнении условий $nh \rightarrow \infty, h \rightarrow 0$; $h = \text{const}$ — длина интервала группировки; $k = [\frac{x}{h}] \in \mathbf{Z}$ — номер интервала группировки; $[a]$ — целая часть числа a ; $\mathbf{1}_A(x_i)$ — *индикаторная функция* заданного подмножества A , позволяющая подсчитать количество элементов выборки x_i , принадлежащих A :

$$1_A(x_i) = \begin{cases} 0, & \text{если } x_i \notin A; \\ 1, & \text{если } x_i \in A. \end{cases}$$

В качестве подмножеств A при построении эмпирической функции распределения $F_n(x)$ выбираются полубесконечные интервалы с переменной границей $(-\infty, x)$, $x \in \mathbf{R}$, а при построении гистограммы $f_{n,h}(x)$ — разбиение области определения на интервалы равной длины $[kh, (k+1)h)$, $k \in \mathbf{Z}$.

Замечание 3.1. Для выбора длины интервала группировки h существует множество эмпирических формул, но для обеспечения поточечной сходимости $f_{n,h}(x)$ к $f(x)$ должно выполняться условие, чтобы при больших объёмах выборок n и малых длинах интервалов h их произведение nh оставалось бы достаточно большим, например, $h = \frac{1}{\sqrt{n}}$ и тому подобное.

Замечание 3.2. С учётом заведомо дискретного характера реализаций случайной выборки $\{x_i\}$ статистические оценки функций распределения $F_n(x)$ и плотности вероятности $f_{n,h}(x)$ представляют собой кусочно-постоянные функции, примеры которых будут приведены ниже.

Выборочная медиана $x_{\frac{1}{2},n}$ эмпирического распределения определяется с помощью вариационного ряда $\{x_{(i)}\}$ по формуле:

$$x_{\frac{1}{2},n} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{если } \frac{n}{2} \notin \mathbf{Z}; \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}), & \text{если } \frac{n}{2} \in \mathbf{Z}. \end{cases}$$

Обобщая предыдущую формулу, найдём *выборочный квантиль* p -*рядка* p :

$$x_{p,n} = \begin{cases} x_{([np]+1)}, & \text{если } np \notin \mathbf{Z}; \\ \frac{1}{2}(x_{([np])} + x_{([np]+1)}), & \text{если } np \in \mathbf{Z}, \end{cases}$$

где $[a]$ — целая часть числа a . При анализе распределений с большими выбросами для характеристики центра распределения вместо *выборочного среднего* \bar{x} часто используется *выборочная медиана* $x_{\frac{1}{2},n}$. Аналогично, для характеристики разброса значений вместо *исправленной выборочной дисперсии* s_x^2 в таких случаях используется *выборочный интерквартильный размах*, то есть разность между третьей и первой выборочными квантилями: $x_{\frac{3}{4},n} - x_{\frac{1}{4},n}$.

Пример 3.1. В качестве примера вычислим основные выборочные характеристики и построим графики эмпирической функции распределения $F_n(x)$ и гистограммы $f_{n,h}(x)$ для выборки 100 значений случайной величины X с помощью **R**.

```

1 > source("samples.r")
2 > n <- 100; x <- samples(n, seed=20100625); x
3   [1] 10.415234  6.226460  5.335232  7.626587 10.634318
4   [6]  8.719651  5.328345  6.046289  9.491963 11.339810
5  [11]  7.040767 10.098266 10.743983  7.857163  7.221059
6  [16] 10.729515  5.757895  8.625306  6.882450  7.310465
7  [21]  6.407871 10.063467  8.974054 10.797271  5.870843
8  [26]  9.474356  6.239238  8.361255  8.920304  7.167161
9  [31]  8.245096  6.396023 10.001735  8.193686  6.238837
10 [36]  8.394063  7.985744  6.862698  7.728719  7.198774
11 [41]  9.277423  8.440071  9.040913  8.768474  6.939044
12 [46]  9.519461  7.029109  7.589811  5.405237  7.825954
13 [51]  9.671514  7.853999  8.546850 10.353427 10.423974
14 [56]  8.558719  6.707131  8.413566  7.706516  9.676796
15 [61]  5.166097  7.741313  8.501668  4.130656  9.746505
16 [66] 11.362920  9.501895 14.872348 13.412845  9.878529
17 [71]  8.478567 10.123378  9.040385  8.062636  9.040725
18 [76]  9.094855  9.590011 12.443799  7.228846  4.561412
19 [81]  9.654944  8.533316  8.083763  9.139730  7.054913
20 [86] 11.271246  8.594706 10.764662  6.686746  8.989150
21 [91]  6.190772  6.475368  8.065002  6.734576  9.072982
22 [96]  6.610282  9.775622  7.343985  4.356711  8.230622

```

Команда «source("samples.r")» в первой строке листинга загружает вспомогательную функцию, которая вызывается во второй строке и обеспечивает генерирование 100 выборочных значений случайной величины X . Параметр «seed=20100625» устанавливает начальное состояние генератора псевдослучайных чисел, формирующего выборку. Если вы пожелаете получить другую последовательность выборочных значений, то укажите другое значение этого параметра. При организации индивидуальной работы студентов значение параметра «seed» может быть указано ведущим занятием преподавателем.

```

23 > a1 <- mean(x); s1 <- sd(x); a1; s1
24   [1] 8.343084
25   [1] 1.891777
26 > quantile(x, c(0, .25, .5, .75, 1))
27   0%      25%      50%      75%     100%
28 4.130656 7.037852 8.403814 9.537099 14.872348

```

В строках [23–28] вычисляются основные выборочные характеристики: среднее значение $\bar{x} \approx 8.34$, исправленное среднее квадратиче-

ское отклонение $s_x \approx 1.89$, а также квантили: $x_{0,n} \approx 4.13$, $x_{\frac{1}{4},n} \approx 7.04$, $x_{\frac{1}{2},n} \approx 8.40$, $x_{\frac{3}{4},n} \approx 9.54$, $x_{1,n} \approx 14.87$. Заметим, что квантили уровней 0 и 1 соответствуют минимальному и максимальному элементам выборки: $x_{0,n} = \min(x_i)$, $x_{1,n} = \max(x_i)$.

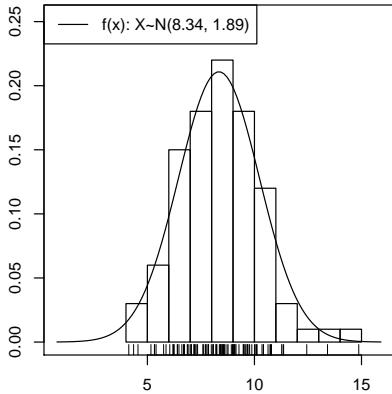


Рис. 3.1. Гистограмма $f_{n,h}(x)$ выборки значений и график функции плотности вероятности $f(x)$ с.в. $X \sim \mathcal{N}(8.34, 1.89)$

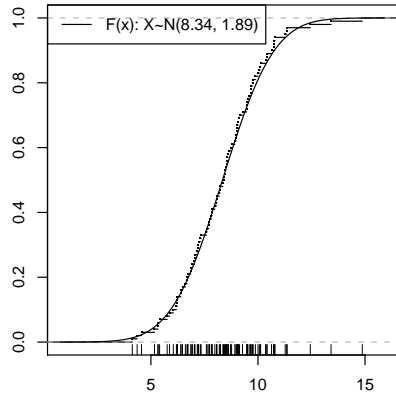


Рис. 3.2. Эмпирическая функция распределения $F_n(x)$ выборки и график функции распределения $F(x)$ с.в. $X \sim \mathcal{N}(8.34, 1.89)$

```

29 > x2 <- seq(a1-4*s1, a1+4*s1, len=n); range(x2)
30 [1] 0.7759757 15.9101930
31 > f1 <- dnorm(x2, a1, s1); F1 <- pnorm(x2, a1, s1)
32 > ltext <- sprintf("X~N(%.2f, %.2f)", a1, s1); ltext
33 [1] "X~N(8.34, 1.89)"
34 > hist(x, breaks="Scott", xlim=range(x2), ylim=c(0, 1.2*max(f1)),
35 +     freq=FALSE, main="", xlab="", ylab="")
36 > rug(x); lines(x2, f1); box()
37 > legend("topleft", lty=1, legend=paste("f(x):", ltext))
38 > windows()
39 > plot(ecdf(x), pch=".", xlim=range(x2),
40 +     main="", xlab="", ylab="")
41 > rug(x); lines(x2, F1)
42 > legend("topleft", lty=1, legend=paste("F(x):", ltext))

```

В строках [29–42] выполняются построения гистограммы $f_{n,h}(x)$ и эмпирической функции распределения $F_n(x)$ для выборки 100 значений случайной величины X , см. рис. 3.1 и 3.2. В дополнение к эмпирическим оценкам $f_{n,h}(x)$ и $F_n(x)$ на тех же рисунках для сравнения

приводятся графики теоретических функций плотности вероятности $f(x)$ и функции распределения $F(x)$, построенных по несмещённым оценкам параметров: $a \approx 8.34$, $\sigma \approx 1.89$.

В строке [29] с помощью оценок параметров a и σ формируется вектор абсцисс «x2». Наименьшее и наибольшее значения вектора «x2» отображаются с помощью функции «range()». В строке [31] с помощью функций «dnorm()» и «pnorm()» формируются соответствующие абсциссам «x2» векторы ординат теоретических функций плотности вероятности «f1» и функции распределения «F1» случайной величины $X \sim \mathcal{N}(8.34, 1.89)$.

В строке [32] с помощью функции «sprintf()» и оценок параметров «a1» и «s1» формируется строка «ltext», содержащая описание предполагаемого закона распределения случайной величины X , [33].

В строках [34–35] с помощью функции «hist()» выполняется построение гистограммы для вектора выборочных значений «x». Параметр «breaks="Scott"» позволяет указать алгоритм вычисления числа интервалов группировки; в данном случае видно, что для построения гистограммы было использовано 11 интервалов. Для уточнения пределов изменения по осям абсцисс и ординат используются параметры «xlim=range(x2)» — размах по оси абсцисс составляет $\bar{x} \pm 4s_x$, и «ylim=c(0, 1.2*max(f1))» — верхнее значение по оси ординат на 20% превышает максимум плотности вероятности, равный $\frac{1}{s_x \sqrt{2\pi}}$. Логический параметр «freq=FALSE» указывает, что при построении гистограммы по оси ординат откладывается не абсолютная частота n_k , а плотность относительной частоты $\frac{n_k}{nh_k}$, где $h_k = h = \text{const}$, что обеспечивает нормировку гистограммы по площади: $\frac{1}{nh} \sum n_k = 1$, где $k = 1, 2, \dots, 11$. Параметры «main», «xlab», «ylab» позволяют установить подписи как для графика в целом, так и индивидуально для осей абсцисс и ординат.

Функция «windows()» в строке [38] открывает новое графическое окно, а функция «plot(ecdf(x)...)» в строке [39] строит в этом окне график эмпирической функции распределения $F_n(x)$, соответствующей выборочным данным «x». Параметр «pch="."» указывает на символ, которым отмечается начало каждого постоянного участка после скачка функции $F_n(x)$. Все остальные параметры имеют тот же смысл, что и для функции «hist()».

Функция «rug(x)» в строках [36], [41] отображает над осью абсцисс метки, соответствующие координатам выборочных значений, а функция «lines()» строит кривые, соответствующие теоретическим функциям плотности вероятности $f(x)$ и распределения $F(x)$.

Функции «legend("topleft"...)» в строках [37], [42] отображают

в левом верхнем углу графика пояснительную надпись, часть которой была получена ранее в строке [32]: «`legend=paste(...ltext)`». Параметр «`lty=1`» указывает, что теоретические функции $f(x)$ и $F(x)$, соответствующие распределению случайной величины $X \sim \mathcal{N}(8.34, 1.89)$, отображаются на графиках сплошной линией.

3.3. Интервальные оценки параметров распределения

При оценивании неизвестных параметров распределения наряду с рассмотренными выше точечными оценками получили распространение *интервальные оценки*. В отличие от точечной интервальная оценка позволяет получить *вероятностную характеристику* точности оценивания неизвестного параметра θ .

Пусть имеется случайная выборка объёма n из *непрерывного распределения* случайной величины с неизвестным параметром θ , для оценки которого строится интервал: (θ_n^-, θ_n^+) , где θ_n^\pm — функции случайной выборки, такие, что верно равенство

$$P\{\theta \in (\theta_n^-, \theta_n^+)\} = \gamma.$$

Тогда интервал $I_\gamma(\theta) = (\theta_n^-, \theta_n^+)$ называют *доверительным интервалом*, накрывающим неизвестный параметр θ с заданной *доверительной вероятностью* γ или γ -*доверительным интервалом*.

Заметим, что при построении доверительных интервалов для *дискретных случайных величин* вместо равенства удаётся обеспечить лишь неравенство

$$P\{\theta \in (\theta_n^-, \theta_n^+)\} \geq \gamma.$$

Доверительная вероятность γ , как правило, считается заданной, близкой к единице и при отсутствии других соображений выбирается среди значений: 0.9, 0.95, 0.975, 0.99, 0.995, ...

Один из типичных методов построения доверительного интервала основан на использовании статистики $T(\theta)$, функция распределения которой $F(t)$ не зависит от оцениваемого параметра θ . При этом используются следующие предположения:

1. Функция распределения статистики $F(t)$ является непрерывной и возрастающей;
2. Для любой реализации выборки статистика $T(\theta)$ является непрерывной и монотонной функцией параметра θ ;

3. Задана доверительная вероятность γ .

Согласно первому предположению для любого числа $p \in [0, 1]$ существует единственный квантиль t_p уровня p функции распределения $F(t)$. Отсюда с учётом третьего предположения получим равенства:

$$\mathbf{P} \left\{ T(\theta) \in \left(t_{\frac{1-\gamma}{2}}, t_{\frac{1+\gamma}{2}} \right) \right\} = F \left(t_{\frac{1+\gamma}{2}} \right) - F \left(t_{\frac{1-\gamma}{2}} \right) = \gamma,$$

справедливые для любых допустимых значений параметра θ , так как функция распределения статистики $F(t)$ от θ не зависит. Согласно второму предположению для любой реализации выборочной совокупности уравнения $T(\theta) = t_{\frac{1\pm\gamma}{2}}$ имеют единственные решения $\theta = \theta_n^\pm$, определяющие искомый доверительный интервал $\mathbf{I}_\gamma(\theta) = (\theta_n^-, \theta_n^+)$.

Доверительный интервал для $M(X)$ при $X \sim \mathcal{N}(a, \sigma)$

При построении *доверительного интервала для математического ожидания* нормально распределённой случайной величины $X \sim \mathcal{N}(a, \sigma)$ по случайной выборке объёмом n используется статистика вида

$$T(a) = \frac{\bar{x} - a}{s_x} \sqrt{n}.$$

Действительно, если $\bar{x} \sim \mathcal{N}(a, \frac{\sigma}{\sqrt{n}})$, то $Z = \frac{\bar{x} - a}{\sigma} \sqrt{n} \sim \mathcal{N}(0, 1)$. В то же время $V = s_x^2 \frac{n-1}{\sigma^2} \sim \chi_{n-1}^2$, причём случайные величины Z и V независимы и статистика $T(a)$ может быть представлена в виде

$$T(a) = \frac{\bar{x} - a}{s_x} \sqrt{n} = Z \sqrt{\frac{n-1}{V}}.$$

Отсюда следует, что статистика $T(a)$ имеет распределение Стьюдента с $n-1$ числом степеней свободы. Для убывающей по параметру a функции $T(a)$ определяющие доверительный интервал уравнения $T(\theta) = t_{\frac{1\pm\gamma}{2}}$ принимают вид

$$T(a_n^\pm) = \frac{\bar{x} - a_n^\pm}{s_x} \sqrt{n} = t_{\frac{1\pm\gamma}{2}, n-1}.$$

Решая эти уравнения, находим нижнюю и верхнюю границы γ -доверительного интервала для математического ожидания $M(X) = a$ нормально распределённой случайной величины $X \sim \mathcal{N}(a, \sigma)$:

$$I_{\gamma}(a) = \left(\bar{x} - \frac{s_x}{\sqrt{n}} \cdot t_{\frac{1+\gamma}{2}, n-1}, \bar{x} - \frac{s_x}{\sqrt{n}} \cdot t_{\frac{1-\gamma}{2}, n-1} \right),$$

где $t_{\frac{1\pm\gamma}{2}, n-1}$ — квантили уровней $\frac{1\pm\gamma}{2}$ распределения Стьюдента с числом степеней свободы $n - 1$.

Доверительный интервал для $D(X)$ при $X \sim \mathcal{N}(a, \sigma)$

При построении *доверительного интервала для дисперсии* нормально распределённой случайной величины $X \sim \mathcal{N}(a, \sigma)$ по случайной выборке объёмом n используется статистика, имеющая распределение χ^2 с числом степеней свободы $n - 1$

$$V(\sigma) = \frac{n-1}{\sigma^2} \cdot s_x^2 \sim \chi_{n-1}^2.$$

Для убывающей по параметру σ функции $V(\sigma)$ нижняя и верхняя границы γ -доверительного интервала определяются уравнениями

$$V(\sigma_n^{2\pm}) = \frac{n-1}{\sigma_n^{2\pm}} \cdot s_x^2 = \chi_{\frac{1\pm\gamma}{2}, n-1}^2.$$

Решая эти уравнения, находим γ -доверительный интервал для дисперсии $D(X) = \sigma^2$ нормально распределённой случайной величины $X \sim \mathcal{N}(a, \sigma)$:

$$I_{\gamma}(\sigma^2) = \left(s_x^2 \cdot \frac{(n-1)}{\chi_{\frac{1+\gamma}{2}, n-1}^2}, s_x^2 \cdot \frac{(n-1)}{\chi_{\frac{1-\gamma}{2}, n-1}^2} \right),$$

где $\chi_{\frac{1\pm\gamma}{2}, n-1}^2$ — квантили уровней $\frac{1\pm\gamma}{2}$ распределения Пирсона с числом степеней свободы $n - 1$.

Пример 3.2. Построим реализации доверительных интервалов для математического ожидания $M(X)$ и дисперсии $D(X)$ стандартной нормально распределённой случайной величины $X \sim \mathcal{N}(0, 1)$ при различных значениях доверительной вероятности $\gamma \in [0.95, 0.999]$ и объёма выборок $n \in [100, 1000]$ с помощью **R**.

```

1 > set.seed(20100625)
2 > n <- seq(100, 1000, 20)
3 > g <- seq(0.95, 0.995, , length(n))
4 > ciM <- function(x,n,g) mean(x)-sd(x)/sqrt(n)*qt((1+c(g,0,-g))/2,n-1)
5 > ciD <- function(x,n,g) sd(x)^2*(n-1)/qchisq((1+c(g,0,-g))/2,n-1)
6 > ciMn <- sapply(n, function(nn) ciM(rnorm(nn), nn, g[1]))

```

```

7 > ciDn <- sapply(n, function(nn) ciD(rnorm(nn), nn, g[1]))
8 > ciMg <- sapply(g, function(gg) ciM(rnorm(n[1]), n[1], gg))
9 > ciDg <- sapply(g, function(gg) ciD(rnorm(n[1]), n[1], gg))
10 > txtMn <- sprintf("MX(n,g=%.3g)", g[1])
11 > txtMg <- sprintf("MX(g,n=%.0f)", n[1])
12 > txtDg <- sprintf("DX(g,n=%.0f)", n[1])
13 > ci_graph <- function(x, y, point, text) { windows()
14 +   plot(range(x), range(y), type="n", xlab="", ylab="")
15 +   for(j in seq(length(y))) {
16 +     lines(x[,j], rep(y[j],3), lwd=2)
17 +     points(x[2,j], y[j], pch=16, lwd=2) }
18 +   legend("topright", legend=text, bg="white")
19 +   abline(v=point, lty=2, lwd=2) }
20 > ci_graph(ciMn, n, 0, txtMn)
21 > ci_graph(ciMg, g, 0, txtMg)
22 > ci_graph(ciDn, n, 1, txtDn)
23 > ci_graph(ciDg, g, 1, txtDg)

```

В строке [1] устанавливается состояние генератора псевдослучайных чисел, а в строках [2–3] формируются векторы значений объема выборки «n» и доверительной вероятности «g».

В строках [4–5] определяются функции «ciM» и «ciD», которые с заданной вероятностью «g» вычисляют границы доверительных интервалов для $M(X)$ и $D(X)$ по выборке «x» заданного объема «n».

Далее, в [6–7] с помощью функции «sapply()» для каждого значения объема выборки из интервала $n \in [100, 1000]$ с фиксированной вероятностью $\gamma_1 = 0.95$ строятся реализации доверительных интервалов для $M(X)$ и $D(X)$ случайной величины $X \sim \mathcal{N}(0, 1)$, показанные на рис. 3.3 и 3.4 в верхнем ряду.

Аналогично, в [8–9] с помощью функции «sapply()» для каждого значения доверительной вероятности из интервала $\gamma \in [0.95, 0.995]$ при фиксированном объеме выборки $n_1 = 100$ строятся реализации доверительных интервалов для $M(X)$ и $D(X)$ случайной величины $X \sim \mathcal{N}(0, 1)$, показанные на рис. 3.3 и 3.4 в нижнем ряду.

В строках [10–13] формируются поясняющие надписи для каждого графика, а затем в [14–20] описывается функция «ci_graph», выполняющая построение самих графиков.

Функция «windows()» в строке [14] открывает новое графическое окно, в котором функция «plot()» рисует оси координат, используя размахи абсцисс «range(x)» и ординат «range(y)». Никаких построений кроме осей координат функция «plot()» не выполняет: «type="n"», [15].

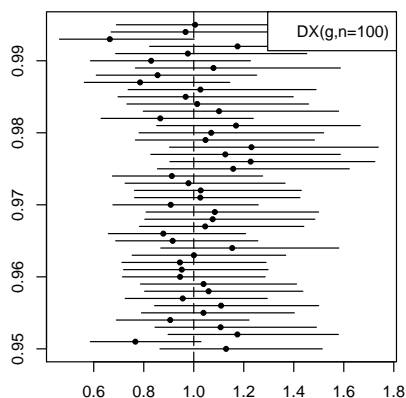
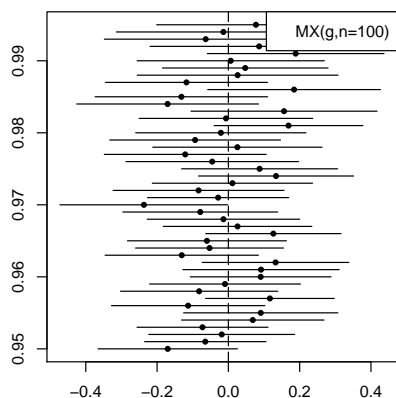
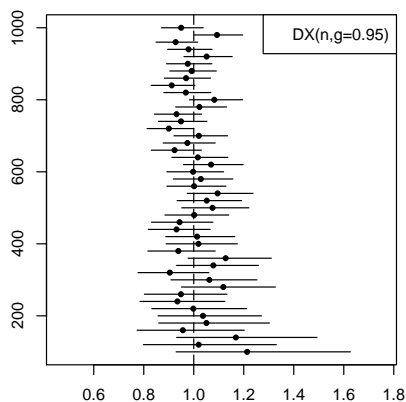
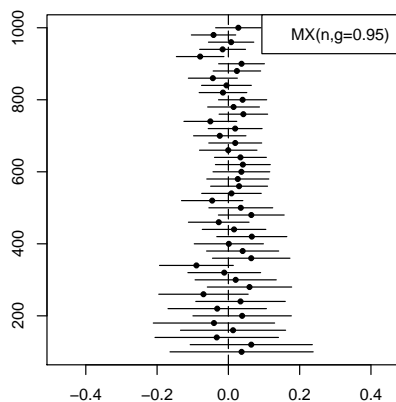


Рис. 3.3. Реализации доверительных интервалов $M(X)$ для различных $n \in [100, 1000]$ и $\gamma \in [0.95, 0.995]$ при $X \sim \mathcal{N}(0, 1)$

Рис. 3.4. Реализации доверительных интервалов $D(X)$ для различных $n \in [100, 1000]$ и $\gamma \in [0.95, 0.995]$ при $X \sim \mathcal{N}(0, 1)$

Фактическими построениями занимается цикл «for()» функций «lines()» и «points()» в строках [16–18], который для каждой тройки абсцисс «x[,j]» и ординат «rep(y[j],3)» рисует пару горизонтальных линий двойной толщины с точкой между ними. Абсцисса точки соответствует j -ой точечной оценке, а отрезки горизонтальных линий — j -му доверительному интервалу для $M(X)$ или $D(X)$.

Функция «legend("topright...")» в строке [19] рисует в правом верхнем углу графика заданный текст «legend=text» на белом фоне

«bg="white"», а функция «abline()» в строке [20] — штриховую линию двойной толщины «lty=2, lwd=2», пересекающую ось абсцисс по нормали в точке «v=point», соответствующей истинным значениям оцениваемых величин: $M(X) = 0$, $D(X) = 1$.

В строках [21–24] описанная функция используется для непосредственного построения реализаций доверительных интервалов $M(X)$ и $D(X)$ при различных значениях $n \in [100, 1000]$ и $\gamma \in [0.95, 0.999]$ для $X \sim \mathcal{N}(0, 1)$.

Из графиков, показанных на рис. 3.3 и 3.4, хорошо видно, что доверительные интервалы и для $M(X)$ и для $D(X)$ представляют собой коллинеарные оси абсцисс случайные векторы, длина которых уменьшается по мере увеличения объема выборки n и уменьшения доверительной вероятности γ .

3.4. Проверка статистических гипотез

Статистической гипотезой принято считать любое предположение о законе распределения случайной величины генеральной совокупности или о значениях параметров закона распределения.

Высказанное предположение, которое подлежит проверке, обозначается H_0 и называется *основной* или *нулевой* гипотезой. Наряду с основной гипотезой в рассмотрение вводится и противоречащая ей гипотеза H_1 , которая называется *конкурирующей* или *альтернативной*. Цель проверки статистической гипотезы заключается в том, чтобы установить, не противоречит ли высказанная гипотеза H_0 имеющимся выборочным данным $\{X_1, X_2, \dots, X_n\}$.

Для проверки нулевой гипотезы формируется *статистический критерий* — специальная статистика $K(X_1, X_2, \dots, X_n)$, распределение которой в условиях нулевой гипотезы H_0 известно. По известному распределению статистического критерия определяется множество значений, которые величина K принимает с вероятностью γ , близкой к единице, то есть практически достоверно. Это множество называется *областью принятия нулевой гипотезы* H_0 . Дополнение этого множества образует *критическую область* (или *область отвержения гипотезы* H_0).

Проверка нулевой гипотезы осуществляется следующим образом. По выборочным данным вычисляется наблюдаемое значение критерия $K_n = K(X_1, X_2, \dots, X_n)$. Если значение K_n принадлежит критической области, то проверяемая гипотеза H_0 отвергается, как противоречащая выборочным данным, и принимается альтернативная

гипотеза H_1 . Если же K_n принадлежит области принятия нулевой гипотезы, то она принимается, как согласующаяся с выборочными данными. В этом случае говорят, что нулевая гипотеза принимается на уровне значимости $\alpha = 1 - \gamma$.

Принципиально важно понимание того, что статистическими методами можно лишь опровергнуть выдвинутую гипотезу H_0 , но нельзя её доказать.

Уровень значимости гипотезы α характеризует вероятность совершить ошибку первого рода, заключающуюся в напрасном отвержении верной нулевой гипотезы: $P\{H_1|H_0\} = \alpha$. Помимо этого, существует вероятность совершить ошибку второго рода, состоящую в напрасном принятии неверной нулевой гипотезы $P\{H_0|H_1\} = \beta$. Дополнительную к β величину, соответствующую вероятности недопущения ошибки второго рода $P\{H_1|H_1\} = 1 - \beta$, называют *мощностью критерия*. Заметим, что одновременное уменьшение вероятностей ошибок первого и второго рода возможно только при увеличении объёма выборки n .

Во многих системах компьютерной математики, в том числе и в **R**, для наблюдаемого значения критерия K_n определяется *достигаемый уровень значимости*, называемый также « p -значением» или « p -value», соответствующий наименьшему уровню значимости α , при котором нулевая гипотеза H_0 отвергается для данного наблюдаемого значения критерия K_n . Чем меньше значение величины p , тем увереннее отвергается нулевая гипотеза H_0 .

Важное значение в математической статистике имеет принцип двойственности при построении доверительных интервалов и проверке гипотез о значениях параметров распределения. Нетрудно убедиться в том, что при выбранном уровне надёжности γ доверительный интервал для некоторого параметра θ составляют те его значения, которые совместимы с гипотезой $H_0: \theta = \theta_n$ на уровне значимости $\alpha = 1 - \gamma$.

3.4.1. Пирсона χ^2 -критерий согласия

Пусть необходимо проверить нулевую гипотезу H_0 о том, что случайная величина X подчиняется определённому закону распределения $F_0(x)$, то есть $H_0: F(x) = F_0(x)$. Если не оговорено иное, то под альтернативной гипотезой H_1 будем понимать дополнение к нулевой, то есть $H_1: F(x) \neq F_0(x)$. Для того чтобы определить, согласуются ли результаты наблюдений с нулевой гипотезой H_0 , принято использовать критерии согласия.

Критерием согласия называется статистический критерий проверки гипотезы о соответствии эмпирического распределения вероятностей — теоретическому. Выделяют *общие критерии согласия*, применимые для проверки любых видов распределений вероятностей, и *специальные критерии*, применимые для проверки определенных групп распределений. В последнем случае при формулировании критериев согласия используются свойства функций для выбранной группы распределений.

Критерии согласия могут быть основаны на изучении разницы между теоретической плотностью распределения и гистограммой (к примеру, критерий согласия χ^2), а могут — на изучении разницы между теоретической и эмпирической функциями распределения (к примеру, критерий Колмогорова–Смирнова).

Гипотезы: Проверяется нулевая гипотеза $H_0 : F(x) = F_0(x, \theta)$ против альтернативной $H_1 : F(x) \neq F_0(x, \theta)$, где $F_0(x, \theta)$ — теоретическая функция распределения случайной величины X ; $\theta \in \mathbf{R}^m$ — m -мерный вектор в общем случае неизвестных параметров распределения X .

Статистика: Критерий согласия χ^2 , предложенный К. Пирсоном в 1900 году, основывается на анализе группированных данных. При этом область возможных значений реализации выборки $\{x_1, x_2, \dots, x_n\}$ разбивают на k непересекающихся интервалов: $x_j \in (a_0, a_k] = (a_0, a_1] \cup (a_1, a_2] \cup \dots \cup (a_{k-1}, a_k]$ и вычисляют статистику, имеющую распределение χ^2 с числом степеней свободы $k - m - 1$:

$$X_d^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \sim \chi_{k-m-1}^2, \quad n_i = \sum_{j=1}^n \mathbf{1}_{(a_{i-1}, a_i]} x_j,$$

где n_i — эмпирическая частота попаданий выборочных значений x_j в интервал $(a_{i-1}, a_i]$; p_i — теоретическая вероятность попадания значений случайной величины X в интервал $(a_{i-1}, a_i]$: $p_i = F_0(a_i, \theta_n) - F_0(a_{i-1}, \theta_n)$, где $\theta_n \in \mathbf{R}^m$ — выборочная оценка m -мерного вектора неизвестных параметров распределения θ .

Критерий: Если наблюдаемое значение статистики превосходит на заданном уровне значимости α квантиль распределения χ^2 с тем же числом степеней свободы: $X_d^2 > \chi_{\alpha, k-m-1}^2$, то нулевая гипотеза на уровне значимости α отвергается в пользу альтернативной $H_1 : F(x) \neq F_0(x, \theta)$. В противном случае

при $X_d^2 \leq \chi_{\alpha, k-m-1}^2$ говорят, что нулевая гипотеза $H_0 : F(x) = F_0(x, \theta)$ на уровне значимости α согласуется с выборочными данными.

В ряде случаев критерий согласия χ^2 может демонстрировать слабую устойчивость на выборках с *низкочастотными событиями* $n_i < 5$. Для решения этой проблемы обычно рекомендуется *объединять интервалы*, не отвечающие критерию $n_i \geq 5$, с соседними до достижения частот приемлемого уровня или использовать *равновероятное группирование*, при котором $n_i \approx \frac{n}{k}$, где $i = 1, 2, \dots, k$.

Необходимо отметить, что по действующим рекомендациям уменьшение числа степеней свободы в распределении χ^2 на число неизвестных параметров m до $k - m - 1$ оправдано лишь в том случае, когда эти параметры θ оценивались по группированным данным¹. Если же оценки параметров θ вычислялись по негруппированной реализации выборки, то действительное распределение наблюдаемой статистики будет заключено между χ_{k-m-1}^2 и χ_{k-1}^2 и при определённых допущениях будет лучше аппроксимироваться распределением χ_{k-1}^2 .

Пример 3.3. Для реализации выборки, использованной в примере 3.1, выполним проверку нулевой гипотезы $H_0 : F(x) = F_0(x, (\bar{x}, s_x))$, где $F_0(x, (\bar{x}, s_x)) = \Phi(\frac{x-\bar{x}}{s_x}) + \frac{1}{2}$ при альтернативной $H_1 = H_0$ по критерию согласия Пирсона на уровне значимости $\alpha = 0.05$.

```

1 > source("samples.r")
2 > x <- samples(n=100, seed=20100625)
3 > a <- mean(x); s <- sd(x); a; s
4 [1] 8.343084
5 [1] 1.891777
6 > range(x)
7 [1] 4.130656 14.872348
8 > b1 <- c(4:15); m1 <- table(cut(x, breaks=b1)); m1
9 (4,5] (5,6] (6,7] (7,8] (8,9] (9,10] (10,11] (11,12]
10 3 6 15 18 22 18 12 3
11 (12,13] (13,14] (14,15]
12 1 1 1

```

Назначение команд в строках [1–5] целиком аналогичны ранее указанным в примере 3.1. В строке [6–7] вычисляются наибольшее и наименьшее значения по реализации: $x_{\min} \approx 4.13$, $x_{\max} \approx 14.87$.

¹ ГОСТ Р 50.1.033-2001. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть 1: Критерии типа хи-квадрат. — М.: Госстандарт России, 2002. — 168 с.

При построении равномерного разбиения указанный диапазон следует «расширить» до ближайших целых или рациональных значений: $x_{\min} \downarrow \hat{x}_{\min}$ и $x_{\max} \uparrow \hat{x}_{\max}$ таким образом, чтобы общая длина была кратной выбранному шагу h : $\hat{x}_{\max} - \hat{x}_{\min} = kh$, а число интервалов группировки лежало бы в диапазоне: $k \in [5, 15]$. Этому соответствует разбиение целыми точками интервала: $(\hat{x}_{\min}, \hat{x}_{\max}] = (4, 15] = (4, 5] \cup (5, 6] \cup \dots \cup (14, 15]$.

В строке [8] формируется вектор граничных точек «b1» и с помощью суперпозиции функций «table(cut())» осуществляется группировка выборочных значений по указанным интервалам [9–12].

```

13 > b2 <- c(4,6:11,15); m2 <- table(cut(x, breaks=b2)); m2
14      (4,6]  (6,7]  (7,8]  (8,9]  (9,10] (10,11] (11,15]
15          9      15      18      22      18      12      6
16 > b3 <- c(-Inf,6:11,Inf); p3 <- diff(pnorm(b3,a,s))
17 > round(p3,5); sum(p3)
18 [1] 0.10775 0.13111 0.18918 0.20775 0.17365 0.11046 0.08009
19 [1] 1
20 > chisq.test(x=m2, p=p3)
21      Chi-squared test for given probabilities
22      data:  m2
23      X-squared = 1.291, df = 6, p-value = 0.9722
24 > x1 <- seq(4,15,length=300); f1 <- dnorm(x1,a,s)
25 > hist(x, breaks=b2); rug(x); lines(x1, f1)
26 > windows(); qqnorm(x, pch=3); qqline(x, lty=2)

```

Из приведённых в строках [9–12] данных видно, что первый интервал группировки и четыре последних содержат слишком мало значений: $n_1 = n_8 = 3$, $n_9 = n_{10} = n_{11} = 1$. Тогда, для соответствия условию $n_i \geq 5$, следует попытаться объединить эти интервалы с соседними: $(4, 5] \cup (5, 6] = (4, 6]$ и $(11, 12] \cup (12, 13] \cup (13, 14] \cup (14, 15] = (11, 15]$. Новый вектор граничных точек «b2» и соответствующая ему группировка выборочных значений показаны в строках [13–15].

Для подсчёта вектора теоретических частот «p3» в строке [16] используется дополнительный вектор «b3», «расширяющий» границы эмпирического разбиения на всю область определения функции $F_0(x)$. Значение «Inf» в системе **R** соответствует бесконечности, а суперпозиция функций «diff(pnorm())» по вектору заданных граничных точек «b3» вычисляет приращения функции нормального распределения $\Phi\left(\frac{x-\bar{x}}{s_x}\right) + \frac{1}{2}$, которые показаны в строке [18].

В строке [20] с помощью критерия χ^2 выполняется проверка гипотезы о соответствии эмпирических частот — теоретическим для нормального закона распределения $\mathcal{N}(\bar{x}, s_x)$. Из данных в строке [23] вид-

но, что достигаемый для исследуемой реализации выборки «х» уровень значимости «p-value = 0.9722» значительно превосходит заданное значение $\alpha = 0.05$, что позволяет сделать вывод о согласии исследуемых данных с нулевой гипотезой: $H_0 : F(x) = \Phi\left(\frac{x-\bar{x}}{s_x}\right) + \frac{1}{2}$.

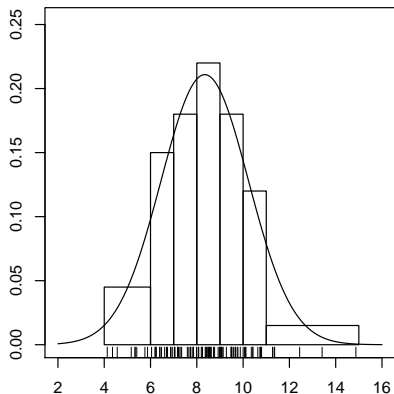


Рис. 3.5. Гистограмма $f_{n,h}(x)$ выборки значений и график плотности вероятности $f(x) = \frac{1}{1.89} \varphi\left(\frac{x-8.34}{1.89}\right)$

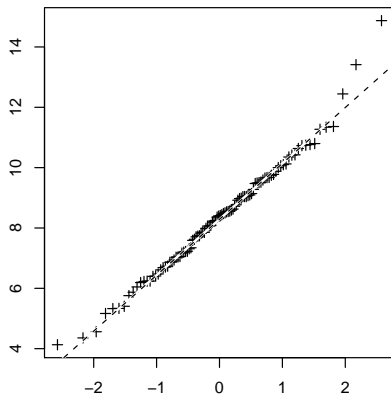


Рис. 3.6. Q-Q график для выборки значений и график функции распределения $F(x) = \Phi\left(\frac{x-8.34}{1.89}\right) + \frac{1}{2}$

В качестве иллюстрации в строках [24–25] выполняются построения гистограммы для вышеуказанной группировки выборочных данных и кривой плотности вероятности $f(x) = \frac{1}{1.89} \varphi\left(\frac{x-8.34}{1.89}\right)$, которые показаны на рис. 3.5.

Ещё одной удобной иллюстрацией к проверке гипотезы о нормальности распределения является квантиль-квантильный (Q–Q) график, построение которого реализовано в строке [26] и показано на рис. 3.6. Для построения Q–Q графика используется функция «qqnorm()», которая выполняет отображение выборочных данных на «нормальной вероятностной бумаге», где абсциссами являются теоретические, а ординатами — эмпирические квантили выборочных данных. Теоретические квантили вычисляются в предположении, что параметры нормального распределения соответствуют их несмещённым точечным оценкам, то есть $X \sim \mathcal{N}(8.34, 1.89)$. Функция «qqline()» добавляет к выборочным данным график функции нормального распределения $F(x) = \Phi\left(\frac{x-8.34}{1.89}\right) + \frac{1}{2}$, выглядящий в указанной системе координат как прямая линия. При этом использованы два параметра: «pch=3» — отображение выборочных точек символами «+»; «lty=2» — отображе-

ние Q-Q графика $F(x)$ штриховой линией.

Из приведённых иллюстраций видно, что отклонения эмпирических распределений выборочных данных от теоретических весьма незначительны, что согласуется с принятой нулевой гипотезой.

3.4.2. Критерии Колмогорова–Смирнова

Критерий согласия Колмогорова используется для проверки простой гипотезы о том, подчиняется ли данное эмпирическое распределение точно известной теоретической модели. Критерий однородности Смирнова предназначен для проверки гипотезы о том, подчиняются ли два эмпирических распределения одному и тому же закону.

Критерий согласия Колмогорова

Гипотезы: Проверяется *нулевая* гипотеза $H_0 : F(x) = F_0(x)$ против *альтернативной* $H_1 : F(x) \neq F_0(x)$, где $F_0(x)$ — теоретическая непрерывная функция распределения случайной величины X , известная с точностью до своих параметров θ .

Статистика: Рассматривается так называемая *статистика Колмогорова*, соответствующая максимальному абсолютному отклонению эмпирической функции распределения $F_n(x)$ от теоретической $F_0(x)$

$$D_n = \sup_{|x| < \infty} |F_n(x) - F_0(x)|,$$

где $\sup A$ — *точная верхняя грань* или *супремум*, обобщающий понятие максимума на случай любого упорядоченного множества A .

В теореме Колмогорова доказывается, что при $n \rightarrow \infty$ случайная величина $\sqrt{n}D_n$ стремится по вероятности к распределению Колмогорова

$$\lim_{n \rightarrow \infty} \mathbf{P} \{ \sqrt{n}D_n \leq t \} = K(t) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 t^2}.$$

Если объём выборки n достаточно велик, то квантиль распределения Колмогорова K_α можно приблизительно вычислить по формуле

$$K_\alpha \approx \sqrt{-\frac{1}{2} \ln \frac{\alpha}{2}}.$$

Критерий: Если наблюдаемое значение статистики превосходит на заданном уровне значимости α квантиль распределения Колмогорова: $\sqrt{n}D_n > K_\alpha$, то нулевая гипотеза на уровне значимости α отвергается в пользу альтернативной $H_1 : F(x) \neq F_0(x)$. В противном случае говорят, что нулевая гипотеза $H_0 : F(x) = F_0(x)$ на уровне значимости α согласуется с выборочными данными.

Пример 3.4. Для центрированной реализации выборки, использованной в предыдущем примере $v_i = x_i - \bar{x}$, с помощью критерия согласия Колмогорова проанализировать зависимость достигаемого уровня значимости α_p от числа степеней свободы $m \in [2, 20]$ для нулевой гипотезы $H_0 : F(v) = F_0(v)$ при альтернативной $H_1 : F(v) \neq F_0(v)$, где $F_0(v)$ — теоретическая функция распределения Стьюдента с заданным числом степеней свободы m .

```

1 > source("samples.r")
2 > x <- samples(n=100, seed=20100625)
3 > m <- 2:20; v <- x - mean(x); w <- sort(v)
4 > alpha <- sapply(m, function(k) ks.test(v, "pt", k)[[2]])
5 > plot(m, alpha, type="b", pch=20, ylim=c(0, max(alpha, 0.05)))
6 > abline(h=0.05, lty=2)
7 > windows(); plot(ecdf(v), pch=".", main="", xlab="", ylab="")
8 > rug(v); lines(w, pt(w, m[which.max(alpha)]))

```

Назначение команд в строках [1–2] целиком аналогичны ранее указанным в примере 3.1. В строке [3] задаётся вектор числа степеней свободы «m», вычисляется центрированный вектор значений выборки «v» и проводится его сортировка по возрастанию «w». Далее в строке [4] с помощью композиции функций «sapply(...ks.test()[[2]])» для каждого значения числа степеней свободы «m» по критерию согласия Колмогорова вычисляются достигаемые уровни значимости α_p для нулевой гипотезы $H_0 : F(v) = F_0(v)$ при альтернативной $H_1 : F(v) \neq F_0(v)$, где $F_0(v)$ — теоретическая функция распределения Стьюдента с заданным числом степеней свободы m .

Далее в строках [5–6] с помощью функции «plot()» строится график зависимости $\alpha_p(m)$, на котором с помощью функции «abline()» горизонтальной «h=0.05» штриховой линией «lty=2» отмечается типичный уровень значимости, используемый при проверке гипотез.

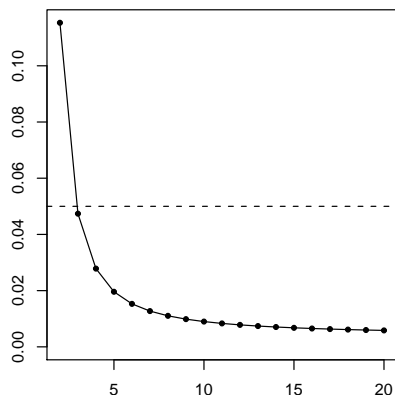


Рис. 3.7. Изменение уровня значимости α_p от числа степеней свободы m , достигаемого для нулевой гипотезы $H_0 : F(v) = F_0(v)$ при альтернативной $H_1 : F(v) \neq F_0(v)$

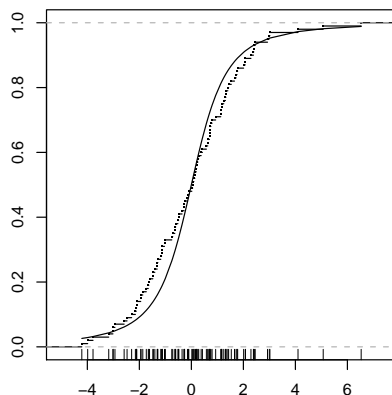


Рис. 3.8. Графики эмпирической $F_m(v)$ и наиболее близкой к ней теоретической функции распределения Стьюдента $F_0(v)$ с числом степеней свободы $m = 2$

В строках [7–8] с помощью композиций «`plot(ecdf())...`», а также «`lines(...pt())`» строятся графики эмпирической функции распределения $F_m(v)$ и теоретической функции t -распределения $F_0(v)$ с числом степеней свободы $m = 2$, соответствующим максимально достижимому уровню значимости «`m[which.max(alpha)]`».

Изучение графиков показывает, что с ростом числа степеней свободы m теоретического распределения Стьюдента уровень значимости α_p , достигаемый при проверке нулевой гипотезы $H_0 : F(v) = F_0(v)$ по критерию согласия Колмогорова падает, что на первый взгляд плохо согласуется со свойствами t -распределения. Объяснение этого кажущегося несоответствия авторы предлагают читателю найти самостоятельно.

Критерий однородности Смирнова

Гипотезы: Проверяется нулевая гипотеза $H_0 : F_1(x) = F_2(x)$ против альтернативной $H_1 : F_1(x) \neq F_2(x)$, где $F_1(x)$ и $F_2(x)$ — неизвестные теоретические функции распределения, для оценки которых используются построенные по независимым выборкам объемами n и m эмпирические функции распределения $F_n(x)$ и $F_m(x)$.

Статистика: Здесь также используется статистика Колмогорова, соответствующая максимальному абсолютному отклонению эмпирических функций распределения $F_n(x)$ и $F_m(x)$

$$D_{n,m} = \sup_{|x| < \infty} |F_n(x) - F_m(x)|.$$

В теореме Смирнова доказывается, что при $n, m \rightarrow \infty$ случайная величина $\sqrt{\frac{nm}{n+m}} D_{n,m}$ стремится по вероятности к распределению Колмогорова

$$\lim_{n,m \rightarrow \infty} \mathbf{P} \left\{ \sqrt{\frac{nm}{n+m}} D_{n,m} \leq t \right\} = K(t) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 t^2}.$$

Критерий: Если наблюдаемое значение статистики превосходит на заданном уровне значимости α квантиль распределения Колмогорова: $\sqrt{\frac{nm}{n+m}} D_{n,m} > K_\alpha$, то нулевая гипотеза на данном уровне значимости α отвергается в пользу альтернативной $H_1 : F_1(x) \neq F_2(x)$. В противном случае говорят, что нулевая гипотеза $H_0 : F_1(x) = F_2(x)$ на уровне значимости α согласуется с выборочными данными.

Пример 3.5. Для двух частей реализации выборки, использованной в примере 3.1: $\{x_i\}_n = \{u_j\}_m \cup \{v_k\}_l$, где $n = m + l$ — объёмы полной и частичных выборок, с помощью критерия однородности Смирнова на уровне значимости $\alpha = 0.05$ проверить нулевую гипотезу $H_0 : F_1(u) = F_2(v)$ при альтернативной $H_1 : F_1(u) \neq F_2(v)$.

```

1 > source("samples.r")
2 > x <- samples(n=100, seed=20100625)
3 > u <- x[1:49]; v <- x[50:100]; ks.test(u,v)
4       Two-sample Kolmogorov-Smirnov test
5 data:  u and v
6 D = 0.1949, p-value = 0.2479
7 alternative hypothesis: two-sided
8 > plot(ecdf(u), pch=25, cex=0.5, xlim=range(x)); rug(u, side=1)
9 > plot(ecdf(v), pch=24, cex=0.5, add=TRUE); rug(v, side=3)

```

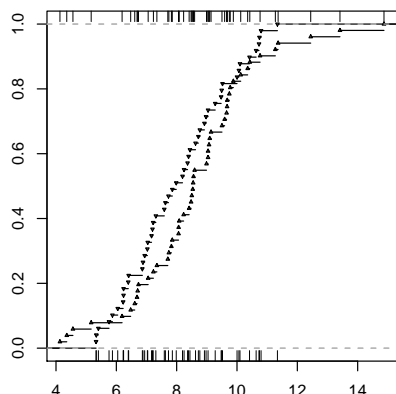


Рис. 3.9. Графики эмпирических функций распределения $F_m(u)$ и $F_l(v)$ для соответствующих частей выборки: $\{u_j\}_m$ и $\{v_k\}_l$

В строках [8–9] с помощью композиции «`plot(ecdf())`» выполняется построение эмпирических функций распределения $F_m(u)$ и $F_l(v)$, а с помощью команды «`rug()`» — отображение соответствующих этим функциям частей выборки вблизи нижней и верхней границ графика.

3.4.3. Стьюдента t -критерий значимости различий

Критерии значимости различий предполагают проверку гипотез о численных значениях известного закона распределения. Например, гипотезы о равенстве средних значений $H_0 : \bar{x} = \bar{y}$ или гипотезы о равенстве дисперсий $H_0 : \sigma_x^2 = \sigma_y^2$.

Исторически t -критерий Стьюдента получил свое название в связи с работой Уильяма Госсета, опубликованной в 1908 году в журнале «Биометрика» под псевдонимом «Student». В настоящее время, под t -критерием Стьюдента понимаются любые тесты, в которых статистика критерия имеет распределение Стьюдента.

Наиболее часто t -критерии применяются для проверки нулевой гипотезы о равенстве средних значений в двух выборках. Все разновидности критерия Стьюдента основаны на предположении о нормальности выборочных данных. Поэтому перед применением критерия Стьюдента необходимо проверить соответствующую гипотезу с помощью одного из критериев согласия.

Назначение команд в строках [1–2] соответствуют ранее указанным в примере 3.1. В строке [3] вычисляются частичные выборочные векторы $\{u_j\}_m$ и $\{v_k\}_l$, а затем «`ks.test()`» по критерию однородности Смирнова вычисляется достигаемый уровень значимости для нулевой гипотезы $H_0 : F_1(u) = F_2(v)$ при альтернативной $H_1 : F_1(u) \neq F_2(v)$.

В строке [6] показан достигаемый уровень значимости $\alpha_p = 0.2479$, сравнение которого с заданным уровнем $\alpha = 0.05$ позволяет сделать вывод об удовлетворительном согласовании нулевой гипотезы с выборочными данными.

Одновыборочный t -критерий

Гипотезы: Проверяется *нулевая* гипотеза $H_0 : \bar{x} = a$ против *альтернативных* H_1 : а) $\bar{x} \neq a$, б) $\bar{x} < a$, в) $\bar{x} > a$, где \bar{x} — среднее значение; a — заданное постоянное значение.

Статистика: При проверке нулевой гипотезы используется статистика, которая имеет распределение Стьюдента с числом степеней свободы $n - 1$

$$T_s = \frac{\bar{x} - a}{s_x} \sqrt{n} \sim t_{n-1},$$

где s_x^2 — исправленная выборочная дисперсия.

Критерий: Если наблюдаемое значение статистики T_s :

- а) *по модулю превосходит* $\frac{\alpha}{2}$ -квантиль распределения Стьюдента с числом степеней свободы $n - 1$: $|T_s| > t_{\frac{\alpha}{2}, n-1}$;
- б) *меньше* α -квантиля распределения Стьюдента с числом степеней свободы $n - 1$: $T_s < t_{\alpha, n-1}$;
- в) *больше* $(1 - \alpha)$ -квантиля распределения Стьюдента с числом степеней свободы $n - 1$: $T_s > t_{1-\alpha, n-1}$, то нулевая гипотеза $H_0 : \bar{x} = a$ на уровне значимости α *отвергается* в пользу альтернативных H_1 : а) $\bar{x} \neq a$; б) $\bar{x} < a$; в) $\bar{x} > a$.

В противном случае говорят, что нулевая гипотеза на уровне значимости α *согласуется* с выборочными данными.

Пример 3.6. Для реализации выборки, использованной в примере 3.1, проанализировать с помощью одновыборочного критерия Стьюдента зависимости для достигаемого уровня значимости α_p от значения параметра $a \in [\bar{x} - \frac{s_x}{2}, \bar{x} + \frac{s_x}{2}]$ для нулевой гипотезы $H_0 : \bar{x} = a$ при альтернативных H_1 : а) $\bar{x} \neq a$, б) $\bar{x} < a$, в) $\bar{x} > a$.

```

1 > source("samples.r")
2 > x <- samples(n=100, seed=20100625)
3 > a <- mean(x); s <- sd(x)
4 > a <- seq(a-s/2, a+s/2, length=99)
5 > p <- sapply(a, function(aa) t.test(x, mu=aa, alter="two")[[3]])
6 > pl <- sapply(a, function(aa) t.test(x, mu=aa, alter="le")[[3]])
7 > pg <- sapply(a, function(aa) t.test(x, mu=aa, alter="gr")[[3]])
8 > plot(a, p, type="l"); lines(a, pl, lty=2); lines(a, pg, lty=4)
9 > abline(h=c(0,0.05), lty=3)

```

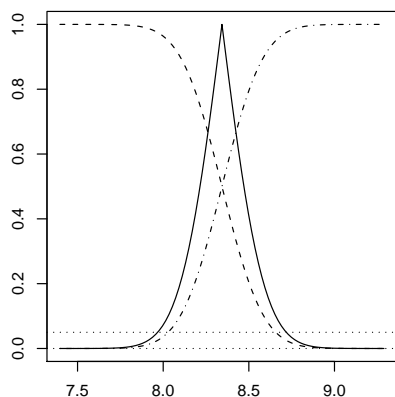



Рис. 3.10. Зависимость достигаемого уровня значимости α_p от параметра a для нулевой гипотезы $H_0: \bar{x} = a$ при альтернативных H_1 : а) $\bar{x} \neq a$, б) $\bar{x} < a$, в) $\bar{x} > a$

В строке [9] с помощью функции «`abline()`» отмечается пятипроцентный уровень значимости, позволяющий приближённо оценить размеры доверительных интервалов для a к каждой из альтернативных гипотез (а–в).

Двухвыборочный t -критерий для независимых выборок

Для применения данного критерия помимо предположения о нормальности выборочных данных, также необходимо соблюдение условия равенства дисперсий: $\sigma_x^2 = \sigma_y^2$.

Задача сравнения средних двух нормально распределённых выборок при неизвестных и неравных дисперсиях известна как проблема Беренса–Фишера. Точного решения этой задачи к настоящему времени не существует, но на практике получили распространение различные приближенные методы.

Гипотезы: Проверяется нулевая гипотеза $H_0: \bar{x} = \bar{y}$ против альтернативных: а) $H_1: \bar{x} \neq \bar{y}$, б) $H_1: \bar{x} < \bar{y}$, в) $H_1: \bar{x} > \bar{y}$, где \bar{x} , \bar{y} — средние значения.

Статистика: При проверке нулевой гипотезы используется статистика, которая имеет распределение Стьюдента с числом степе-

Назначение команд в строках [1–3] соответствуют ранее указанным в примере 3.1. В строках [4–7] вначале находится вектор значений a , а затем с помощью композиции «`sapply(...t.test())`» по t -критерию Стьюдента для каждого значения a вычисляют достигаемые уровни значимости α_p к каждой из альтернативных гипотез (а–в).

В строке [8] выполняется построение кривых $\alpha_p(a)$, соответствующих основной гипотезе $H_0: \bar{x} = a$ при альтернативных H_1 : а) $\bar{x} \neq a$, б) $\bar{x} < a$, в) $\bar{x} > a$. Указанные кривые изображаются на графике: а) сплошной, б) штриховой «`lty=2`», в) штрихпунктирной «`lty=4`» линиями.

ней свободы $m + n - 2$:

$$T_s = (\bar{x} - \bar{y}) \sqrt{\frac{mn(m+n-2)}{(m+n)((m-1)s_x^2 + (n-1)s_y^2)}} \sim t_{m+n-2},$$

где m, n — объёмы выборок $\{x_i\}_m$ и $\{y_j\}_n$; s_x^2, s_y^2 — исправленные выборочные дисперсии.

Критерий: Если наблюдаемое значение статистики T_s :

- а) по модулю превосходит $\frac{\alpha}{2}$ -квантиль распределения Стьюдента с числом степеней свободы $m + n - 2$: $|T_s| > t_{\frac{\alpha}{2}, m+n-2}$;
- б) меньше α -квантиля распределения Стьюдента с числом степеней свободы $m + n - 2$: $T_s < t_{\alpha, m+n-2}$;
- в) больше $(1 - \alpha)$ -квантиля распределения Стьюдента с числом степеней свободы $n - 1$: $T_s > t_{1-\alpha, m+n-2}$, то нулевая гипотеза $H_0: \bar{x} = \bar{y}$ на уровне значимости α отвергается в пользу альтернативных H_1 : а) $\bar{x} \neq \bar{y}$, б) $\bar{x} < \bar{y}$, в) $\bar{x} > \bar{y}$.

В противном случае говорят, что нулевая гипотеза на уровне значимости α согласуется с выборочными данными.

Пример 3.7. Для двух частей реализации выборки, использованной в примере 3.1: $\{x_i\}_n = \{u_j\}_m \cup \{v_k\}_l$, где $n = m + l$ — объёмы полной и частичных выборок, с помощью двухвыборочного t -критерия Стьюдента на уровне значимости $\alpha = 0.05$ проверить нулевую гипотезу $H_0: \bar{u} = \bar{v}$ при альтернативных H_1 : а) $\bar{u} \neq \bar{v}$, б) $\bar{u} < \bar{v}$, в) $\bar{u} > \bar{v}$.

```

1 > source("samples.r")
2 > x <- samples(n=100, seed=20100625)
3 > u <- x[1:49]; v <- x[50:100]
4 > t.test(u,v, var.equal=TRUE, alter="two")
5       Two Sample t-test
6 data:  u and v
7 t = -1.4731, df = 98, p-value = 0.1439
8 alternative hypothesis: true difference in means is not equal to 0
9 95 percent confidence interval:
10  -1.3007923  0.1923783
11 sample estimates:
12 mean of x mean of y
13  8.060439  8.614646
14 > t.test(u,v, var.equal=TRUE, alter="le")
15       Two Sample t-test
16 data:  u and v
17 t = -1.4731, df = 98, p-value = 0.07196
18 alternative hypothesis: true difference in means is less than 0
19 95 percent confidence interval:
20  -Inf 0.07051633

```

```

21 sample estimates:
22 mean of x mean of y
23 8.060439 8.614646
24 > t.test(u,v, var.equal=TRUE, alter="gr")
25 Two Sample t-test
26 data: u and v
27 t = -1.4731, df = 98, p-value = 0.928
28 alternative hypothesis: true difference in means is greater than 0
29 95 percent confidence interval:
30 -1.178930 Inf
31 sample estimates:
32 mean of x mean of y
33 8.060439 8.614646

```

Назначения команд в строках [1–2] соответствуют ранее указанным в примере 3.1. В строке [3] вычисляются частичные выборочные векторы $\{u_j\}_m$ и $\{v_k\}_l$.

Функция «`t.test()`» в строках [4], [14] и [24] по двухвыборочному t -критерию Стьюдента вычисляет достигаемый уровень значимости для нулевой гипотезы $H_0: \bar{u} = \bar{v}$ при альтернативных H_1 : а) $\bar{u} \neq \bar{v}$, б) $\bar{u} < \bar{v}$, в) $\bar{u} > \bar{v}$.

В строках [7], [17] и [27] показаны достигаемые уровни значимости для соответствующих пар нулевой и альтернативных гипотез: а) $\alpha_p = 0.1439$, б) $\alpha_p = 0.07196$, в) $\alpha_p = 0.928$. Сравнение достигаемых уровней с заданным $\alpha = 0.05$ позволяет в случаях (а) и (б) сделать выводы об удовлетворительном, а в случае (в) — о хорошем согласовании нулевой гипотезы с выборочными данными.

3.4.4. Фишера F -критерий значимости различий

F -критерий Фишера применяется для проверки гипотезы о равенстве дисперсий. Критерий Фишера может применяться как самостоятельно, так и перед проверкой гипотез о равенстве средних с помощью критерия Стьюдента. Если гипотеза о равенстве дисперсий принимается, то для сравнения средних можно выбрать более мощный критерий. Критерий Фишера основан на дополнительных предположениях о независимости и нормальности выборочных данных. Поэтому перед применением критерия Фишера необходимо проверить соответствующую гипотезу с помощью одного из критериев согласия.

Гипотезы: Проверяется нулевая гипотеза $H_0: \sigma_x^2 = \sigma_y^2$ против альтернативных: а) $H_1: \sigma_x^2 \neq \sigma_y^2$, б) $H_1: \sigma_x^2 > \sigma_y^2$, где σ_x^2, σ_y^2 — дисперсии.

Статистика: При проверке нулевой гипотезы используется статистика, которая имеет распределение Фишера с числом степеней свободы $\frac{m-1}{n-1}$:

$$F_s = \frac{s_x^2}{s_y^2} \sim F_{\frac{m-1}{n-1}},$$

где m, n — объёмы выборок $\{x_i\}_m$ и $\{y_j\}_n$; s_x^2, s_y^2 — исправленные выборочные дисперсии.

Критерий: Если наблюдаемое значение статистики F_s :

а) меньше $\frac{\alpha}{2}$ -квантиля или больше $(1 - \frac{\alpha}{2})$ -квантиля распределения Фишера с числом степеней свободы $\frac{m-1}{n-1}$: $F_s < F_{\frac{\alpha}{2}, \frac{m-1}{n-1}}$ или $F_s > F_{1-\frac{\alpha}{2}, \frac{m-1}{n-1}}$;

б) больше $(1 - \alpha)$ -квантиля распределения Фишера с числом степеней свободы $\frac{m-1}{n-1}$: $F_s > F_{1-\alpha, \frac{m-1}{n-1}}$,

то нулевая гипотеза $H_0 : \sigma_x^2 = \sigma_y^2$ на уровне значимости α отвергается в пользу альтернативных H_1 : а) $\sigma_x^2 \neq \sigma_y^2$, б) $\sigma_x^2 > \sigma_y^2$. В противном случае говорят, что нулевая гипотеза на уровне значимости α согласуется с выборочными данными.

Пример 3.8. Для двух частей реализации выборки, использованной в примере 3.1: $\{x_i\}_n = \{u_j\}_m \cup \{v_k\}_l$, где $n = m + l$ — объёмы полной и частичных выборок, с помощью F -критерия Фишера на уровне значимости $\alpha = 0.05$ проверить нулевую гипотезу $H_0 : \sigma_u^2 = \sigma_v^2$ при альтернативных H_1 : а) $\sigma_u^2 \neq \sigma_v^2$, б) $\sigma_u^2 < \sigma_v^2$, в) $\sigma_u^2 > \sigma_v^2$.

```

1 > source("samples.r")
2 > x <- samples(n=100, seed=20100625)
3 > u <- x[1:49]; v <- x[50:100]
4 > var(u); var(v)
5 [1] 2.701471
6 [1] 4.339142
7 > var.test(u,v)
8      F test to compare two variances
9 data:  u and v
10 F = 0.6226, num df = 48, denom df = 50, p-value = 0.1015
11 alternative hypothesis: true ratio of variances is not equal to 1
12 95 percent confidence interval:
13  0.3538423 1.0990592
14 sample estimates:
15 ratio of variances
16  0.622582
17 > var.test(u,v, alter="le")
18      F test to compare two variances
19 data:  u and v

```

```

20 F = 0.6226, num df = 48, denom df = 50, p-value = 0.05075
21 alternative hypothesis: true ratio of variances is less than 1
22 95 percent confidence interval:
23 0.000000 1.002103
24 sample estimates:
25 ratio of variances
26 0.622582
27 > var.test(u,v, alter="gr")
28 F test to compare two variances
29 data: u and v
30 F = 0.6226, num df = 48, denom df = 50, p-value = 0.9493
31 alternative hypothesis: true ratio of variances is greater than 1
32 95 percent confidence interval:
33 0.3878247 Inf
34 sample estimates:
35 ratio of variances
36 0.622582

```

Назначения команд в строках [1–2] соответствуют ранее указанным в примере 3.1. В строках [3–4] вычисляются частичные выборочные векторы $\{u_j\}_m$ и $\{v_k\}_l$, а с помощью функции «var()» — соответствующие им исправленные выборочные дисперсии: $s_u^2 \approx 2.70$ и $s_v^2 \approx 4.34$.

Функция «var.test()» в строках [7], [17] и [27] по F -критерию Фишера вычисляет достигаемый уровень значимости для нулевой гипотезы $H_0: \sigma_u^2 = \sigma_v^2$ при альтернативных H_1 : а) $\sigma_u^2 \neq \sigma_v^2$, б) $\sigma_u^2 < \sigma_v^2$, в) $\sigma_u^2 > \sigma_v^2$.

В строках [10], [20] и [30] показаны достигаемые уровни значимости для соответствующих пар нулевой и альтернативных гипотез: а) $\alpha_p = 0.1015$, б) $\alpha_p = 0.05075$, в) $\alpha_p = 0.9493$. Сравнение достигаемых уровней с заданным $\alpha = 0.05$ позволяет в случаях (а) и (в) сделать выводы об удовлетворительном и хорошем, а в случае (б) — о неудовлетворительном согласовании нулевой гипотезы с выборочными данными.

3.4.5. Однофакторный дисперсионный анализ

Дисперсионный анализ предназначен для оценки влияния одного или нескольких факторов (качественных величин) на количественную случайную величину. В случае, когда рассматривается влияние только одного качественного признака, имеющего конечное число уровней градаций, дисперсионный анализ называется *однофакторным*.

Предположим, что одна и та же случайная величина X с одинаковой точностью измеряется при k различных значениях фактора. Если

анализируемый фактор оказывает существенное влияние на X , то наблюдения на одном уровне будут значимо отличаться от наблюдений на других уровнях, и, следовательно, средние значения на разных уровнях будут различными. И наоборот, если фактор не оказывает влияние на рассматриваемую случайную величину, то средние значения X на различных уровнях будут статистически незначимо отличаться друг от друга.

Представим результаты наблюдений в виде таблицы:

i	n_i	x_{ij}
1	n_1	$x_{11}, x_{12}, \dots, x_{1n_1}$
2	n_2	$x_{21}, x_{22}, \dots, x_{2n_2}$
\dots	\dots	\dots
k	n_k	$x_{k1}, x_{k2}, \dots, x_{kn_k}$

где i — уровни фактора; $j = 1, 2, \dots, n_i$ — номера наблюдений на i -ом уровне; n_i — количество наблюдений на i -ом уровне; x_{ij} — наблюдаемые значения.

При проведении дисперсионного анализа предполагается выполнение следующих условий:

1. Результаты наблюдений x_{ij} — это независимые случайные величины, то есть $\text{cov}(x_{ij}, x_{lm}) = 0$, где $i \neq l$ и/или $j \neq m$;
2. Совокупности наблюдаемых значений $\{x_{i1}, x_{i2}, \dots, x_{in_i}\}$ на каждом уровне i нормально распределены: $\mathcal{N}(a_i, \sigma_i^2)$, где a_i, σ_i^2 — среднее и дисперсия i -го уровня;
3. Дисперсии распределений на всех уровнях $i = 1, 2, \dots, k$ одинаковы: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \text{const}$.

Гипотеза: С учётом выдвинутых условий формулируется нулевая гипотеза о равенстве средних всех уровней $H_0 : a_1 = a_2 = \dots = a_k$ при альтернативной, что хотя бы одно из указанных равенств нарушается $H_1 : \exists a_l \neq a_m$, где $l \neq m$.

Статистика: Рассмотрим следующие величины:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^k n_i \bar{x}_i, \quad n = \sum_{i=1}^k n_i,$$

где \bar{x}_i — средние значения i -го уровня; \bar{x} — общее среднее значение всех n величин.

$$Q_t = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2, Q_d = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2, Q_r = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$

где Q_t — сумма квадратов отклонений отдельных наблюдений x_{ij} от общего среднего \bar{x} ; Q_d — сумма квадратов отклонений средних значений уровней \bar{x}_i от общей средней \bar{x} , которая характеризует различия между средними значениями отдельных уровней и определяется влиянием рассматриваемого фактора; Q_r — сумма квадратов отклонений отдельных наблюдений x_{ij} от средних значений своего уровня \bar{x}_i , которая обусловлена наличием неучтённых факторов и называется остаточным рассеянием или суммой квадратов внутри групп.

Можно доказать, что имеет место равенство $Q_t = Q_d + Q_r$, причём, левая часть равенства имеет $n - 1$ степень свободы, первое слагаемое в правой части — $(k - 1)$ степень свободы, а второе — $(n - k)$, и каждая сумма квадратов, делённая на соответствующее число степеней свободы, будет представлять несмещённую оценку дисперсии случайной величины X . При этом, величина $\frac{1}{n-1} Q_t$ в любом случае является несмещённой оценкой дисперсии X , а величины $\frac{1}{k-1} Q_d$ и $\frac{1}{n-k} Q_r$ — только в рамках гипотезы о равенстве средних значений уровней фактора, то есть при отсутствии влияния исследуемого фактора на случайную величину X . Тогда при согласии с нулевой гипотезой $H_0 : a_1 = a_2 = \dots = a_k$ статистика F_s будет иметь распределение Фишера с числами степеней свободы числителя $k - 1$, и знаменателя $n - k$:

$$F_s = \frac{\frac{1}{k-1} Q_d}{\frac{1}{n-k} Q_r} = \frac{(n-k) \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2}{(k-1) \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2} \sim F_{\frac{k-1}{n-k}}.$$

Критерий: Гипотеза H_0 принимается, если $F_s < F_{\alpha, \frac{k-1}{n-k}}$, и отвергается в противном случае, где $F_{\alpha, \frac{k-1}{n-k}}$ — квантиль уровня α распределения Фишера с указанными выше числами степеней свободы.

Вышеуказанный выбор критической области $F_s \geq F_{\alpha, \frac{k-1}{n-k}}$ определяется тем, что при выполнении альтернативной гипотезы H_1 статистика F_s неограниченно возрастает с ростом объёма выборки n .

Пример 3.9. В ходе исследования были значения количественного признака для трёх различных уровней качественного признака (фактора). Используя методику однофакторного дисперсионного анализа, требуется определить: значимо ли влияние изменения качественного признака на величину признака количественного?

```

1 > D = c(4.0,4.5,4.3,5.6,4.9,5.4,3.8,3.7,4.0)
2 > B = c(4.5,4.9,5.0,5.7,5.5,5.6,4.7,4.5,4.7)
3 > S = c(5.4,4.9,5.6,5.8,6.1,6.3,5.5,5.0,5.0)
4 > adhf = stack(data.frame(D,B,S)); adhf
5     values ind
6     1     4.0 D
7     2     4.5 D
8     3     4.3 D
9     4     5.6 D
10    5     4.9 D
11    6     5.4 D
12    7     3.8 D
13    8     3.7 D
14    9     4.0 D
15   10     4.5 B
16   11     4.9 B
17   12     5.0 B
18   13     5.7 B
19   14     5.5 B
20   15     5.6 B
21   16     4.7 B
22   17     4.5 B
23   18     4.7 B
24   19     5.4 S
25   20     4.9 S
26   21     5.6 S
27   22     5.8 S
28   23     6.1 S
29   24     6.3 S
30   25     5.5 S
31   26     5.0 S
32   27     5.0 S
33 > anova(lm(values ~ ind, data=adhf))
34     Analysis of Variance Table
35     Response: values
36           Df Sum Sq Mean Sq F value    Pr(>F)
37     ind       2  4.9119  2.45593   7.7578 0.002519 **
38 Residuals  24  7.5978  0.31657
39     ---
40     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

В строках 1-3 вводятся векторы значений выборочных данных. Вектор «D» соответствует замерам количественного признака на пер-

вом, вектор «В» — на втором, а вектор «S» — на третьем уровне качественного признака.

Далее в строке [4] с помощью композиции «`stack(data.frame())`» формируется таблица исходных данных, показанная в строках [5–32], а в строке [33] с помощью композиции «`anova(lm())`» проводится её дисперсионный анализ.

Для разделения столбцов значений на качественные признаки и количественные при проведении дисперсионного анализа используется запись вида «`value~ind`», где столбец «`value`» соответствует количественному вектору, а столбец «`ind`» — качественному.

В строках [37–38] приведены данные по межгрупповым «`ind`» и внутригрупповым «`Residuals`» дисперсиям. Столбец «`Df`» содержит данные по числам степеней свободы, столбцы «`Sum Sq`» и «`Mean Sq`» — данные по суммам квадратов отклонений и дисперсиям наблюдений, столбец «`F value`» содержит наблюдаемое значение F -статистики, а столбец «`Pr(>F)`» — вероятность того, что межгрупповая дисперсия не превышает внутригрупповую.

Как показывает анализ дисперсий, вероятность того, что изменение уровней качественного признака значимо влияет на величину количественного, составляет примерно 99.75%.

Контрольные вопросы

1. Что называют генеральной совокупностью?
2. Что такое выборка (выборочная совокупность)? Что называют объёмом выборки?
3. Что называют статистикой для заданной выборки?
4. Запишите определения эмпирических функций распределения и плотности распределения. Приведите примеры их графиков.
5. Напишите формулы для вычисления основных выборочных характеристик: среднего, дисперсии, ковариации, коэффициента корреляции.
6. Какую оценку называют точечной. Поясните содержание требований, предъявляемых к точечным оценкам (состоятельность, несмещённость, эффективность).
7. Какая точечная оценка является состоятельной, несмещённой и эффективной для математического ожидания генеральной совокупности?

8. Какие точечные оценки для дисперсии генеральной совокупности являются смещёнными и несмещёнными? Являются ли эти оценки состоятельными?
9. Напишите формулы точечных оценок ковариации и коэффициента корреляции.
10. Что называют доверительной вероятностью и доверительным интервалом для неизвестного параметра θ ?
11. Какую статистику используют при построении доверительного интервала для параметра a случайной величины $X \sim \mathcal{N}(a, \sigma)$? По какому закону распределена эта статистика?
12. Какую статистику используют при построении доверительного интервала для параметра σ^2 случайной величины $X \sim \mathcal{N}(a, \sigma)$? По какому закону распределена эта статистика?
13. Что такое статистическая гипотеза? Какие статистические гипотезы называют: основными или альтернативными, сложными или простыми?
14. Что называют статистическим критерием и его уровнем значимости при проверке статистической гипотезы?
15. В чем заключается ошибка первого рода? Что такое уровень значимости (p -уровень)?
16. Какое множество называют критическим? В чем заключается ошибка второго рода? Что называют мощностью критерия?
17. В чем состоит принцип двойственности при построении доверительных интервалов и проверке гипотез о значениях параметров распределения?
18. Какие статистические критерии называют критериями согласия?
19. Каким образом проверяется гипотеза о виде распределения непрерывной случайной величины по χ^2 -критерию Пирсона?
20. Для проверки каких гипотез используются критерии Колмогорова–Смирнова? Какую статистику используют эти критерии?
21. Какие статистические критерии называют критериями значимости различий?
22. Для проверки каких гипотез используются одно- и двухвыборочные t -критерии Стьюдента? Какую статистику используют эти критерии?

-
23. Для проверки каких гипотез используется Фишера F -критерий значимости различий? Какую статистику использует этот критерий?
 24. Какие задачи являются объектом исследования в дисперсионном анализе?
 25. В каком случае дисперсионный анализ называют одно- и многофакторным?
 26. Каковы предпосылки однофакторного дисперсионного анализа?
 27. Как формулируются основная и альтернативная гипотезы однофакторного дисперсионного анализа?

Глава 4

Начала регрессионного анализа

4.1. Основные понятия регрессионного анализа

4.1.1. Зависимые и независимые переменные

Регрессионный анализ исследует и оценивает связь между *зависимой или объясняемой* переменной и *независимыми или объясняющими* переменными. Зависимую переменную иногда называют *результативным признаком*, а объясняющие переменные — *предикторами, регрессорами или факторами*.

Обозначим зависимую переменную y , а независимые — x_1, x_2, \dots, x_k . При $k = 1$ имеется только одна независимая переменная x и регрессия называется *парной*. При $k > 1$ имеется множество независимых переменных x_1, x_2, \dots, x_k и регрессия называется *множественной*.

Рассмотрим построение простейшей регрессионной модели

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

где y — *зависимая* случайная переменная; x — *независимая* детерминированная переменная; β_0, β_1 — *постоянные* параметры уравнения; ε — *случайная* переменная, называемая также *ошибкой*.

Будем считать, что истинная зависимость между x и y — линейная, то есть существует некоторая зависимость $y = \beta_0 + \beta_1 x$. Задача регрессионного анализа заключается в получении оценок коэффициентов β_0, β_1 .

Величина слагаемого ε , соответствует отклонению эмпирических данных от прямой регрессии и может быть связана с ошибками измерений, неверно выбранной формой зависимости между переменными x и y и другими причинами.

Вид зависимости обычно выбирают графически, проверяя качество моделей на контрольной выборке, либо используя априорные соображения.

Для оценивания параметров $\beta_0, \beta_1, \dots, \beta_k$ обычно применяют *метод наименьших квадратов* (МНК). Однако существуют и другие методы оценки: метод максимального правдоподобия, метод наименьших модулей и тому подобное.

4.1.2. Оценка параметров уравнения регрессии

Пусть имеется n наблюдений, тогда уравнение регрессии можно переписать в виде:

$$y_i = b_0 + b_1 x_i + e_i, \quad i = 1, 2, \dots, n.$$

Будем рассматривать случайное слагаемое ϵ как последовательность n случайных величин: e_1, e_2, \dots, e_n .

Метод наименьших квадратов сводится к тому, чтобы получить такие оценки b_0, b_1 параметров β_0, β_1 , при которых минимизируется сумма квадратов отклонений e_i фактических значений признака y_i от теоретических $\hat{y}_i = b_0 + b_1 x_i$:

$$Q_e(b_0, b_1) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 \rightarrow \min.$$

Для минимизации функции $Q_e(b_0, b_1)$ приравняем к нулю её частные производные $\frac{\partial Q_e}{\partial b_0}$ и $\frac{\partial Q_e}{\partial b_1}$:

$$\begin{cases} -2 \sum_{i=1}^n y_i + 2nb_0 + 2b_1 \sum_{i=1}^n x_i = 0; \\ -2 \sum_{i=1}^n y_i x_i + 2b_0 \sum_{i=1}^n x_i + 2b_1 \sum_{i=1}^n x_i^2 = 0. \end{cases}$$

После преобразований получим *систему нормальных уравнений МНК*:

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i; \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i. \end{cases}$$

Решая систему нормальных уравнений, находим b_0, b_1 :

$$b_0 = \frac{1}{n} \sum_{i=1}^n y_i - b_1 \frac{1}{n} \sum_{i=1}^n x_i, \text{ где } b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2},$$

или в компактной форме: $b_0 = \bar{y} - b_1 \bar{x}$, где $b_1 = \frac{\text{cov}(x, y)}{s_x^2}$.

Коэффициент b_1 называется *выборочным коэффициентом регрессии*. Если независимую переменную x увеличить на единицу, то новое значение зависимой переменной $y(x+1)$ будет равно $y(x) + b_1$.

Коэффициент b_0 численно равен значению результирующего признака y при нулевом значении фактора x .

4.1.3. Оценка качества выборочного уравнения регрессии

Уравнение выборочной регрессии имеет вид $y = b_0 + b_1 x$. Обозначим $\hat{y}_i = b_0 + b_1 x_i$ — *расчётное значение* зависимой переменной y , вычисленное при значении независимой переменной $x = x_i$. Тогда $e_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$ — *остатки*, характеризующие отклонения наблюдаемых значений зависимой переменной от расчётных. Заметим, что полная сумма отклонений e_i будет равна нулю при любых выборочных значениях y_i и, следовательно, не может быть использована для оценки качества уравнения регрессии. Это свойство является одним из важнейших оптимизационных свойств МНК-оценок.

В связи с этим при оценке качества выборочного уравнения регрессии используются следующие суммы квадратов отклонений:

$$Q_t = \sum_{i=1}^n (y_i - \bar{y})^2; \quad Q_r = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2; \quad Q_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2,$$

где Q_t — общая сумма квадратов отклонений значений зависимой переменной от её выборочного среднего значения; Q_r — сумма квадратов отклонений расчётных значений зависимой переменной от её выборочного среднего значения; Q_e — сумма квадратов отклонений y_i от линии регрессии, обычно называемая суммой квадратов остатков или ошибок.

Величину $\sqrt{\frac{Q_e}{n-2}}$ называют *средней квадратической погрешностью* или *ошибкой* уравнения регрессии.

Между приведёнными выше суммами квадратов существует связь: $Q_t = Q_r + Q_e$, которая и позволяет характеризовать качество постро-

енного уравнения регрессии. Уравнение регрессии считается тем лучше, чем больше сумма квадратов, обусловленная регрессией Q_r , по сравнению с суммой квадратов остатков Q_e . В этом случае уравнение регрессии воспроизводит большую часть суммы квадратов отклонений зависимой переменной от её среднего значения и может быть использовано в практических приложениях.

Для того чтобы формализовать это представление используется коэффициент детерминации:

$$R^2 = \frac{Q_r}{Q_t} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad R^2 \in [0, 1]$$

причём, чем ближе коэффициент детерминации R^2 к единице, тем выше качество полученного уравнения регрессии. Максимальное значение коэффициента детерминации $R^2 = 1$ достигается в том случае, когда все остатки $e_i = 0$, а уравнение прямой регрессии проходит точно через все точки y_i .

Таким образом, значение коэффициента детерминации R^2 можно интерпретировать как долю общей дисперсии зависимой переменной y , которая будет объяснена (воспроизведена) с помощью уравнения регрессии.

4.1.4. Проверка значимости уравнения регрессии

В рассмотренном выше подходе не учитываются статистические свойства эмпирического материала. Найденные по методу наименьших квадратов коэффициенты b_0 и b_1 являются так называемыми МНК-оценками истинных коэффициентов β_0 и β_1 . Эти оценки являются случайными величинами, зависящими как от реализации выборки (x_i, y_i) , так и от её объёма n .

Использование МНК накладывает ряд ограничений на поведение случайной составляющей ε уравнения регрессии: $y = \beta_0 + \beta_1 x + \varepsilon$. Обычно эти ограничения формулируются в следующем виде:

1. Математические ожидания всех случайных составляющих равны нулю: $M(\varepsilon_i) = 0$, где $i = 1, 2, \dots, n$. Практически это условие означает, что случайная составляющая ε не вносит систематического смещения в значения зависимой переменной y ;

2. Дисперсии всех случайных составляющих *равны друг другу*¹: $D(\varepsilon_i) = \sigma^2$, где $i = 1, 2, \dots, n$. Практически это условие означает, что все наблюдаемые значения зависимой переменной y_i измерены с одинаковой точностью;
3. Различные случайные составляющие ε_i *не коррелируют* друг с другом: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$, где $i, j = 1, 2, \dots, n$. Практически это условие означает, что ошибки при различных наблюдениях независимы. Данное условие часто заменяют предположением о независимости распределения случайной составляющей ε_j и значений величины X , то есть $\text{cov}(x_i, \varepsilon_j) = 0$.
4. Случайные составляющие ε_i *распределены по нормальному закону*: $\varepsilon_i \sim \mathcal{N}(0, \sigma)$, где $i = 1, 2, \dots, n$. При выполнении этого условия уравнение регрессии называется *нормальной* (классической) *линейной регрессионной моделью*.

Условия 1–3 называют *условиями Гаусса–Маркова*, а соответствующая им теорема утверждает, что при выполнении данных условий МНК-оценки параметров уравнения регрессии будут *несмещёнными, состоятельными и эффективными*.

Отметим, что сам метод оценивания параметров не требует соблюдения условия о нормальности распределения случайной составляющей, но это предположение становится необходимым для построения доверительных интервалов МНК-оценок и проверки значимости уравнения в целом. Именно в этих условиях МНК-оценки неизвестных параметров уравнения регрессии обладают ясными статистическими свойствами.

В частности, можно показать, что МНК-оценки b_0, b_1 для параметров нормальной линейной регрессионной модели β_0, β_1 будут иметь нормальные распределения: $b_0 \sim \mathcal{N}(\beta_0, \sigma_{b_0})$, $b_1 \sim \mathcal{N}(\beta_1, \sigma_{b_1})$, где

$$\sigma_{b_0} = \frac{\sigma \sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n \sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \sigma_{b_1} = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Знание законов распределения оценок параметров уравнения регрессии необходимо для построения их доверительных интервалов: $\beta_0 \in (b_0^-, b_0^+)$ и $\beta_1 \in (b_1^-, b_1^+)$ и проверки статистических гипотез.

¹ Выполнение данного условия называют *гомогенностью дисперсии* или *гомоскедастичностью*, а его невыполнение — *гетероскедастичностью*.

Однако, следует иметь в виду, что значение параметра σ в общем случае не является известным, а поэтому вместо точных значений $\sigma(b_0)$ и $\sigma(b_1)$ могут быть использованы лишь их выборочные оценки s_{b_0} и s_{b_1} . Тогда стандартизация выборочных оценок b_0 и b_1 будет приводить не к стандартному нормальному распределению, а к распределению Стьюдента с числом степеней свободы $n - 2$:

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2}, \quad \frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}.$$

Полученную статистику можно использовать для проверки простой гипотезы $H_0 : \beta_0 = b_0$ при альтернативной $H_1 : \beta_0 \neq b_0$. Если известен уровень надёжности γ , то определена соответствующая квантиль $t_{\gamma, n-2}$ и при выполнении соотношения

$$P \left\{ \left| \frac{b_0 - \beta_0}{s_{b_0}} \right| < t_{\gamma, n-2} \right\} = \gamma$$

можно сделать вывод о принятии гипотезы $H_0 : \beta_0 = b_0$, а разрешив вероятностное неравенство — получить доверительный интервал для оценки параметра β_0 с заданной надёжностью γ

$$I_\gamma(\beta_0) = (b_0 - t_{\gamma, n-2}s_{b_0}, b_0 + t_{\gamma, n-2}s_{b_0}).$$

Аналогичные рассуждения приводят к доверительному интервалу для оценки параметра β_1 с заданной надёжностью γ

$$I_\gamma(\beta_1) = (b_1 - t_{\gamma, n-2}s_{b_1}, b_1 + t_{\gamma, n-2}s_{b_1}).$$

Любое значение b_1 из этого интервала будет совместно с гипотезой $H_0 : \beta_1 = b_1$ на уровне значимости $\alpha = 1 - \gamma$. Поэтому, если доверительный интервал содержит нулевое значение, то это будет означать, что имеющиеся данные не позволяют, в частности, отвергнуть гипотезу $H_0 : \beta_1 = 0$. В этом случае построенное уравнение регрессии признается незначимым, то есть принимается утверждение, что связь между переменными x и y в реальности отсутствует, а то, что наблюдается в эксперименте, является случайной особенностью данной выборки. Надёжность такого утверждения соответствует γ , а вероятность ошибочности $\alpha = 1 - \gamma$.

Отметим, что именно гипотеза $H_0 : \beta_1 = 0$ при альтернативной $H_1 : \beta_1 \neq 0$ представляет наибольший практический интерес. Наблюдаемая в критерии статистика при этом имеет вид: $\left| \frac{b_1}{s_{b_1}} \right| \sim t_{n-2}$ и используется для проверки значимости уравнения регрессии.

4.1.5. Точечный и интервальный прогнозы по уравнению регрессии

Уравнение регрессии, полученное в результате анализа эмпирических данных, может быть использовано для прогнозирования значений зависимой переменной y при заданных значениях независимой переменной x путём подстановки этих значений в уравнение: $\hat{y} = b_0 + b_1 x$. Поскольку оценки b_0 и b_1 являются случайными величинами, то вычисленное с их участием расчётное значение \hat{y} также будет являться случайной величиной. Причём в условиях нормальной линейной регрессии прогнозируемая величина будет иметь нормальное распределение: $y \sim \mathcal{N}(\hat{y}, k_{\hat{y}}\sigma_{\hat{y}})$, где

$$k_{\hat{y}} = \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Заменяя неизвестное значение параметра $\sigma_{\hat{y}}$ его оценкой

$$s_{\hat{y}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

получим доверительный интервал для y с заданной надёжностью γ :

$$I_{\gamma}(y) = (\hat{y} - t_{\gamma, n-2} k_{\hat{y}} s_{\hat{y}}, \hat{y} + t_{\gamma, n-2} k_{\hat{y}} s_{\hat{y}}).$$

Пакеты статистической обработки данных обычно отображают интервальные прогнозы для расчётных значений зависимой переменной в виде двух гипербол, расположенных выше и ниже построенной линии регрессии.

Если требуется оценить индивидуальное расчётное значение \dot{y} , то в оценке дисперсии необходимо дополнительно учитывать дисперсию самого наблюдения. В этом случае доверительный интервал принимает вид:

$$I_{\gamma}(\dot{y}) = (\hat{y} - t_{\gamma, n-2} k_{\dot{y}} s_{\dot{y}}, \hat{y} + t_{\gamma, n-2} k_{\dot{y}} s_{\dot{y}}),$$

где

$$k_{\dot{y}} = \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Пример 4.1. Для заданных в файле «regression1.csv» векторов выборочных данных x_i и y_i построить линейную модель парной регрессии и проверить её качество.

```

1  > dat <- read.table("regression1.csv", head=TRUE)
2  > attach(dat); dat
3      x      y
4  1  2.36  1.12
5  2  2.67  0.46
6  3  2.98  0.19
7  4  3.30 -0.27
8  5  3.61 -0.85
9  6  3.93 -0.79
10 7  4.24 -1.17
11 8  4.56 -1.88
12 9  4.87 -1.62
13 10 5.18 -1.25
14 11 5.50 -1.04
15 > fit <- lm(y ~ x); summary(fit)
16 Call:
17 lm(formula = y ~ x)
18 Residuals:
19      Min       1Q   Median       3Q      Max
20 -0.7473 -0.2662 -0.1076  0.2487  0.8165
21 Coefficients:
22             Estimate Std. Error t value Pr(>|t|)
23 (Intercept)  2.3786      0.5900  4.032 0.002965 **
24 x           -0.7700      0.1456 -5.287 0.000502 ***
25 ---
26 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
27 Residual standard error: 0.4799 on 9 degrees of freedom
28 Multiple R-squared:  0.7565,    Adjusted R-squared:  0.7294
29 F-statistic: 27.96 on 1 and 9 DF,  p-value: 0.0005022
30 > xin <- seq(0.9*min(x), 1.1*max(x), length=100)
31 > pre <- predict(fit, data.frame(x=xin), interval="confidence")
32 > plot(dat, pch=3); points(mean(x), mean(y))
33 > matplot(xin, pre, type="l", lty=c(1,2,2), add=TRUE)
34 > windows(); par(mfrow=c(2,1))
35 > plot(x, fit[[2]], pch=4); abline(h=0)
36 > qqnorm(fit[[2]], pch=4); qqline(as.vector(fit[[2]]))
37 > detach(dat)

```

Функция «read.table()» в строке [1] считывает выборочные данные, по-умолчанию сохранённые в файле «regression1.csv» из текущей папки, а функция «attach()» в строке [2] делает эти данные доступными для расчётов под именами: «x» и «y». По своей структуре загружаемый csv-файл является упорядоченным по столбцам текстовым файлом, первая строка которого содержит имена выборочных

векторов, а последующие строки — их значения. Загруженные имена и значения показаны в строках [3–14]. Заметим, что при использовании данных, загружаемых с функцией «attach()», после проведения расчётов рекомендуется очищать память с использованием «detach()», что сделано в строке [37].

Функция «lm()» в строке [15] на основе приведённых выборочных данных рассчитывает параметры линейной модели вида « $y \sim x$ », что соответствует уравнению парной регрессии: $y_i = b_0 + b_1 x_i + e_i$. Сводка основных результатов расчёта выводится в строках [16–29] с помощью функции «summary()». В строках [23–24] показаны оценки коэффициентов выборочного уравнения регрессии: $b_0 = 2.38$, $b_1 = -0.77$, а также соответствующие значения стандартных ошибок и вероятности отклонения гипотез о равенстве полученных оценок истинным значениям: $P\{\beta_0 \neq b_0\} = 0.003$, $P\{\beta_1 \neq b_1\} = 0.0005$. С учётом значения выборочного коэффициента детерминации $R^2 = 0.76$ в строке [28] качество построенного уравнения регрессии можно охарактеризовать как высокое.

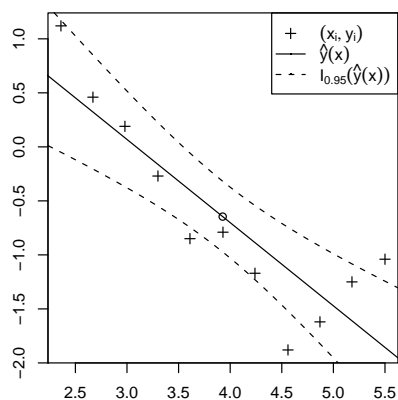


Рис. 4.1. Выборочные значения (x_i, y_i) и доверительные интервалы $I_{0.95}(\hat{y})$ для уравнения регрессии $\hat{y} = 2.38 - 0.77x$

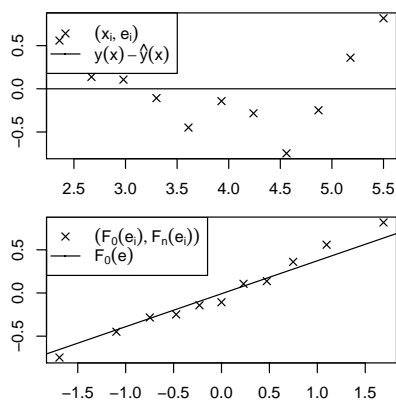


Рис. 4.2. Центрированный относительно \hat{y} график остатков e_i . Q-Q график остатков e_i и функции распределения $F_0(e) = \Phi\left(\frac{e}{0.46}\right) + \frac{1}{2}$

График к полученной линейной модели парной регрессии показан на рис. 4.1. Для построения этого рисунка в строках [30–33] используются функции: «predict(...interval="confidence")» — вычисление границ 0.95-доверительных интервалов для уравнения регрессии;

«plot(...pch=3)» — отображение выборочных значений (x_i, y_i) , используя символы «+»; «points()» — отображение точки (\bar{x}, \bar{y}) , используя символ «o»; «matplot()» — отображение графика выборочного уравнения регрессии $\hat{y} = 2.38 - 0.77x$, а также верхней и нижней границ его доверительных интервалов $I_{0.95}(\hat{y})$, используя сплошную и две штриховые линии: «lty=c(1,2,2)».

После построения модели и проверки качества полезно провести анализ распределения её остатков e_i , показанных на рис. 4.2 сверху. Это можно сделать с помощью Q-Q графика, показанного на рис. 4.2 снизу. Для построения этих графиков в строках [34–36] используются функции: «windows()» — создание нового графического окна; «par(mfrow=c(2,1))» — разбиение графического окна на две части по вертикали; «plot(...pch=4)» — отображение остатков e_i , используя символ «x»; «abline(h=0)» — отображение горизонтальной линии на нулевом уровне; «qqnorm()» — отображение на Q-Q графике остатков e_i для исходных данных линейной модели; «qqline()» — отображение на Q-Q графике функции нормального распределения $F_0(e) = \Phi(\frac{e}{0.46}) + \frac{1}{2}$, где значение 0.46 соответствует исправленному выборочному среднему квадратическому отклонению остатков s_e .

4.2. Модели множественной линейной регрессии

Множественный регрессионный анализ является развитием парного анализа в случае, когда зависимая переменная связана с более чем одной независимой переменной. Модель парной регрессии даёт хороший результат в том случае, когда влиянием других факторов на объект исследования можно пренебречь. Например, если коэффициент детерминации для построенного уравнения регрессии близок к единице: $R^2 \geq 0.8$. Однако в практических задачах такие ситуации являются скорее исключением, чем правилом. Поэтому модели множественной линейной регрессии имеют довольно широкое распространение.

4.2.1. Метод наименьших квадратов для множественной регрессии

Рассмотрим регрессионное уравнение, в котором определяется линейная связь зависимой переменной y от k независимых переменных

x_1, x_2, \dots, x_k :

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon.$$

Пусть проведено n наблюдений, в результате которых получены следующие эмпирические наборы данных:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

Все использованные обозначения соответствуют по смыслу введённым ранее. Основная задача будет заключаться в том, чтобы получить такие оценки b_i параметров β_i , где $i = 0, 1, \dots, k$, при которых сумма квадратов отклонений e_i фактических значений признака y_i от расчётных \hat{y}_i была бы минимальна:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_{1i} - \dots - b_k x_{ki})^2 = \sum_{i=1}^n e_i^2 \rightarrow \min.$$

Рассмотрим следующие векторы и матрицы:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ 1 & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_k \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix}.$$

Столбцами матрицы X являются векторы $X_s = (x_{1s}, x_{2s}, \dots, x_{ns})$, где $s = 0, 1, \dots, k$, соответствующие независимым переменным x_1, x_2, \dots, x_k . Каждый элемент матрицы x_{ij} представляет собой результат i -го наблюдения для j -го признака, а первый единичный столбец соответствует значениям некоторой фиктивной переменной, используемой для большего удобства.

Тогда система уравнений для определения оценок параметров линейной модели множественной регрессии b_0, b_1, \dots, b_k в матричной форме примет вид

$$Y = Xb + e,$$

а подлежащая минимизации сумма квадратов отклонений

$$\sum_{i=1}^n e_i^2 = e^T e = (Y - Xb)^T (Y - Xb) \rightarrow \min.$$

Решение такой задачи базируется на простых геометрических соображениях. Рассмотрим в качестве примера модель линейной регрессии для двух наблюдений: $Y = X\beta + \varepsilon$, где $Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ —

векторы наблюдений, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$ — вектор случайной составляющей. Этой модели соответствуют построения, показанные на рис. 4.3.

Очевидно, что векторы X , $X\beta$ и Xb взаимно коллинеарны, при этом $\varepsilon = Y - X\beta$. Тогда оценку b параметра β следует выбирать таким образом, чтобы модуль оценки e вектора ε был минимальным, откуда вытекает требование ортогональности векторов: $e \perp Xb$.

Так как необходимым и достаточным условием ортогональности двух векторов является равенство нулю их скалярного произведения, то в результате получим систему уравнений в матричной форме:

$$X^T(Y - Xb) = 0.$$

Выполнив соответствующие преобразования, приходим к общей системе нормальных уравнений метода наименьших квадратов

$$X^T X b = X^T Y.$$

Если матрица системы $X^T X$ невырожденная, то система нормальных уравнений будет иметь искомое решение

$$b = (X^T X)^{-1} X^T Y.$$

Оценки b вектора β , полученные при решении указанной системы нормальных уравнений, как и в случае парной регрессии, называются *МНК-оценками* или оценками, полученными по методу наименьших квадратов.

Знание значений МНК-оценок b позволяет вычислять расчётные значения зависимой переменной \hat{Y}

$$\hat{Y} = Xb = X(X^T X)^{-1} X^T Y.$$

Заметим, что геометрически вектор \hat{Y} является наилучшей аппроксимацией вектора Y с помощью линейной комбинации векторов X_i , где $i = 1, 2, \dots, k$.

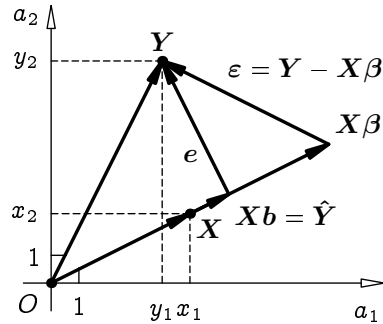


Рис. 4.3. Геометрическая интерпретация модели линейной регрессии на плоскости $a_1 a_2$

4.2.2. Статистические свойства МНК-оценок множественной регрессии

Теорема Гаусса–Маркова: Предположим, что :

1. Дана определённая ранее модель *множественной линейной регрессии*: $Y = X\beta + \epsilon$;
2. Здесь X — детерминированная *матрица*, имеющая максимальный ранг $k + 1$; практически это означает линейную независимость векторов-столбцов матрицы X , откуда следует *невыврожденность* матрицы $X^T X$;
3. $M(\epsilon) = 0$, $M(\epsilon\epsilon^T) = \sigma^2 I$, где I — единичная матрица k -го порядка; первое условие означает *однородность дисперсии* всех случайных составляющих ϵ_i , а второе — *отсутствие корреляции* случайных составляющих для различных наблюдений.

Тогда МНК-оценки $b = (X^T X)^{-1} X^T Y$ будут *несмещёнными* и *эффективными* в классе линейных несмещённых оценок.

Отметим, что матрицы вида $X^T X$ играют весьма важную роль как при построении МНК-оценок, так и при определении их значимости и точности. Например, если векторы выборочных данных *стандартизированы*: $M(X_i) = 0$, $D(X_i) = 1$, то матрица $X^T X$ будет соответствовать матрице *выборочных коэффициентов корреляции*, а в качестве *несмещённой оценки параметра σ^2* используют величину $e^T e$, нормированную по числу степеней свободы: $s^2 = \frac{e^T e}{n-k-1}$.

Можно доказать, что при выполнении условий теоремы Гаусса–Маркова и нормальности распределения случайной составляющей ϵ *оценки параметров b и s будут независимыми*.

В дальнейшем, для оценки значимости коэффициентов уравнения регрессии и построения доверительных интервалов будем использовать матрицу $\text{cov}(b) = s^2 (X^T X)^{-1}$, квадратные корни элементов главной диагонали которой называются *стандартными ошибками коэффициентов уравнения регрессии*.

4.2.3. Оценка качества уравнения множественной регрессии

Как и в случае парной регрессии, качество полученного уравнения будем оценивать по той доли изменчивости зависимой переменной Y , которая объясняется построенным уравнением. С учётом того, что $\epsilon = Y - \hat{Y}$ и $\epsilon \perp \hat{Y}$ запишем следующие равенства:

$$\|Y\|^2 = Y^T Y = ((Y - \hat{Y}) + \hat{Y})^T ((Y - \hat{Y}) + \hat{Y}) = \|Y - \hat{Y}\|^2 + \|\hat{Y}\|^2.$$

Полученное разложение суммы квадратов можно непосредственно увидеть на рис. 4.3 в качестве аналога теоремы Пифагора.

Учитывая, что показанные на рисунке данные были стандартизированы, разложение суммы квадратов в общем случае будет иметь вид: $Q_t = Q_e + Q_r$, где $Q_t = Y Y^T - n \bar{Y}^2$ — общая сумма квадратов отклонений Y относительно среднего \bar{Y} ; $Q_e = Y Y^T - b^T X^T Y$ — сумма квадратов отклонений Y , относительно расчётных значений по уравнению регрессии \hat{Y} ; $Q_r = b^T X^T Y - n \bar{Y}^2$ — сумма квадратов отклонений расчётных значений \hat{Y} относительно среднего \bar{Y} или остаточная сумма квадратов. Все использованные обозначения соответствуют ранее введённым.

Тогда коэффициент детерминации R^2 , определяется так же, как и в случае парной регрессии:

$$R^2 = \frac{Q_r}{Q_t} = \frac{b^T X^T Y - n \bar{Y}^2}{Y Y^T - n \bar{Y}^2}.$$

Свойства коэффициента детерминации R^2 аналогичны сформулированным ранее. Коэффициент R^2 показывает качество подгонки регрессионной модели к наблюдаемым значениям Y .

Если $R^2 = 0$, то $\|e\|^2 = \|Y\|^2$, то есть весь разброс величины Y соответствует случайным отклонениям, называемым ошибками. В этом случае $\hat{Y} = \bar{Y}$ и это значит, что построенное уравнение регрессии не следует использовать, так как оно не улучшает предсказание по сравнению с тривиальным прогнозом. Если же $R^2 = 1$, то этот случай соответствует $\|e\|^2 = 0$ и, следовательно, имеет место точное соответствие, при котором все эмпирические точки лежат на регрессионной гиперплоскости.

Таким образом, R^2 характеризует тесноту связи набора независимых признаков или факторов: X_1, X_2, \dots, X_k с зависимой переменной Y , то есть оценивает степень тесноты их связи. При этом можно показать, что коэффициент детерминации в случае линейной модели с точностью до знака равен выборочному коэффициенту корреляции между наблюдаемыми величинами Y и расчётными \hat{Y} , то есть $|R| = |r_{Y\hat{Y}}|$.

Замечание: Величину $R = \sqrt{R^2}$ в случае множественной регрессионной модели называют ещё и коэффициентом *множественной корреляции*. Такой подход позволяет обобщить и распространить понятие связи на совокупности переменных.

Недостатком коэффициента детерминации R^2 , ограничивающим его применение, является то, что при добавлении новых независимых переменных его значение всегда возрастает, хотя это и не означает улучшения качества модели как таковой. Чтобы избежать этой ситуации предлагается использовать коэффициент детерминации R_a^2 , скорректированный по числу степеней свободы:

$$R_a^2 = 1 - \frac{(n-1)(1-R^2)}{(n-k-1)} = \frac{(n-1) e^T e}{(n-k-1) Y^T Y}.$$

В отличие от R^2 при введении в модель новых независимых переменных скорректированный коэффициент R_a^2 может уменьшаться в том случае, когда эти переменные не оказывают существенного влияния на зависимую переменную. При различном количестве независимых переменных использование R_a^2 для сравнения регрессий является более корректным. Однако в этом случае величину R_a^2 уже не следует интерпретировать как меру объяснённой вариации зависимой переменной.

4.2.4. Проверка значимости уравнения множественной регрессии

Проверка значимости построенного уравнения регрессии может быть выполнена статистическими методами только в том случае, если известны законы распределения статистик, участвующих в построении самого уравнения. Наиболее распространённым и привлекательным является предположение о нормальности распределения случайной составляющей: $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

В этом случае полученные оценки параметров строятся как линейные комбинации нормально распределённых независимых случайных величин. При этом обычно ссылаются на известную теорему, утверждающую, что любая линейная комбинация независимых нормально распределённых случайных величин будет иметь нормальное распределение.

Регрессионная модель при выполнении предположения о нормальности ϵ называется *классической нормальной линейной моделью* множественной регрессии.

В целом значимость уравнения регрессии обычно понимается как существование такой зависимости, в которой на величину Y оказывает влияние хотя бы одна независимая переменная X_i . И наоборот, уравнение регрессии считается незначимым, если все переменные X_i

не связаны с Y , то есть не оказывают на неё никакого влияния. В этом случае вся изменчивость величины Y объясняется случайной составляющей ϵ , и, как было отмечено выше, коэффициент детерминации будет равен нулю: $R^2 = 0$.

Поэтому проверка значимости регрессионной модели будет сводиться к проверке статистической гипотезы $H_0 : R^2 = 0$ при альтернативной $H_1 : R^2 \neq 0$.

Эквивалентная формулировка гипотезы для оценки значимости уравнения регрессии утверждает, что коэффициенты при всех переменных X_j равны нулю $H_0 : \beta_j = 0$ при $j = 1, 2, \dots, k$. Тогда альтернативная гипотеза будет состоять в том, что существует хотя бы одна переменная X_j , коэффициент при которой будет отличен от нуля $H_1 : \exists \beta_j \neq 0$.

Статистика критерия для проверки значимости уравнения регрессии может быть выражена через коэффициент множественной детерминации R^2 :

$$F_s = \frac{Q_r(n - k - 1)}{Q_e k} = \frac{R^2(n - k - 1)}{(1 - R^2)k}.$$

В условиях нулевой гипотезы полученная статистика F_s имеет распределение Фишера с числами степеней свободы числителя k и знаменателя $n - k - 1$. Проверка основной гипотезы осуществляется стандартным образом.

По заданному уровню значимости $\alpha = 1 - \gamma$ определяется α -квантиль распределения Фишера $F_{\alpha, \frac{k}{n-k-1}}$ с указанными числами степеней свободы. Если расчётное значение статистики F_s превышает α -квантиль: $F_s > F_{\alpha, \frac{k}{n-k-1}}$, то гипотезу о незначимости уравнения регрессии отвергают с вероятностью ошибки α , а уравнение множественной регрессии признаётся значимым с надёжностью γ и может быть использовано в практических расчётах.

Если же расчётное значение статистики F_s не превышает α -квантиль: $F_s \leq F_{\alpha, \frac{k}{n-k-1}}$, то в этом случае говорят, что имеющиеся данные не позволяют отвергнуть нулевую гипотезу на выбранном уровне значимости α , а уравнение регрессии признаётся незначимым. Другими словами, уравнение регрессии ничего, кроме случайной составляющей или ошибки, не воспроизводит, и вряд ли имеет смысл использовать его в дальнейшем.

Поскольку рассмотренная критическая статистика представлена в виде отношения двух независимых оценок дисперсий, то данный подход носит название *дисперсионного анализа уравнения регрессии*. В данном случае в качестве фактора, вызывающего разложение

оценки дисперсии, выступает построенное уравнение регрессии. Поэтому проверка гипотезы о значимости влияния фактора на полученные дисперсии равносильна проверке гипотезы о значимости самого построенного уравнения дисперсии.

4.2.5. Доверительные интервалы для b_i и \hat{y}

В условиях классической нормальной линейной модели множественной регрессии имеется возможность оценить не только значимость уравнения регрессии в целом, но и значимость влияния на зависимую величину Y каждой переменной X_k в отдельности. Эта возможность в свою очередь позволяет в дальнейшем производить отбор переменных в уравнении регрессии, исключая из рассмотрения незначимые с точки зрения исследуемой зависимости переменные.

В предположении о нормальности распределения случайных компонент $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ выборочные оценки коэффициентов уравнения регрессии b_i , $i = 0, 1, \dots, k$ будут иметь нормальные распределения, а их нормированные отклонения от истинных значений — распределения Стьюдента с числом степеней свободы $n - k - 1$:

$$b_i \sim \mathcal{N}(\beta_i, \sigma_{\beta_i}), \quad b_i^* = \frac{b_i - \beta_i}{s_{b_i}} \sim t_{n-k-1}.$$

Последнее утверждение позволяет по заданному уровню значимости $\alpha = 1 - \gamma$ проверить нулевую гипотезу $H_0 : \beta_i = b_i$ при альтернативной $H_1 : \beta_i \neq b_i$ и построить доверительный интервал для истинного значения параметра β_i . Для этого следует найти α -квантиль распределения Стьюдента $t_{\alpha, n-k-1}$ и решить вероятностное неравенство $\mathbf{P}\{|b_i^*| < t_{\alpha, n-k-1}\} = \gamma$ относительно оцениваемого значения b_i :

$$\mathbf{I}_\gamma(b_i) = (b_i - t_{\alpha, n-k-1}s_{b_i}; b_i + t_{\alpha, n-k-1}s_{b_i}),$$

где s_{b_i} — стандартная ошибка оценки коэффициента уравнения регрессии b_i : $s_{b_i}^2 = s^2(\mathbf{X}^\top \mathbf{X})^{-1}$.

Полученная оценка позволяет проверить гипотезу о равенстве нулю значения неизвестного параметра β_i . Если эта гипотеза справедлива, то выборочные оценки b_i будут отличаться от нуля лишь за счёт случайных отклонений, а доверительный интервал $\mathbf{I}_\gamma(b_i)$ будет содержать нулевое значение.

Проверка гипотезы о значимости коэффициента β_i сводится к сравнению статистики $\frac{b_i}{s_{b_i}} \sim t_{n-k-1}$ с соответствующим квантилем распределения Стьюдента.

Замечание: При использовании помимо точечных оценок b_i их интервальных аналогов имеются определённые трудности. Можно показать, что если для одного и того же уровня надёжности γ построить доверительные интервалы для каждого b_i , то общая вероятность того, что эти оценки будут соблюдаться одновременно, равняется не γ , а превышает значение $(1 - k\gamma)$.

Наряду с интервальными оценками полученных коэффициентов регрессии в условиях классической нормальной множественной регрессионной модели имеется возможность оценить точность вычисляемой зависимой переменной Y , то есть точность прогноза.

Пусть вектор $X_p = (1, x_{p1}, x_{p2}, \dots, x_{pk})$ представляет значения независимых переменных, при которых требуется определить значение зависимой переменной Y . Тогда $\hat{y}_p = b_0 + b_1 x_{p1} + \dots + b_k x_{pk}$ равняется условному математическому ожиданию (среднему значению) переменной Y при $X = X_p$. Рассуждая аналогично вышеизложенному, получим

$$I_\gamma(\hat{y}_p) = (\hat{y}_p - t_{\alpha, n-k-1} s_{\hat{y}_p}; \hat{y}_p + t_{\alpha, n-k-1} s_{\hat{y}_p}),$$

где $s_{\hat{y}_p}$ — стандартная ошибка расчётного значения \hat{y}_p ; $t_{\alpha, n-k-1}$ — двусторонняя α -квантиль распределения Стьюдента с числом степеней свободы $n - k - 1$; $s_{\hat{y}_p}^2 = s^2(X_p^\top (X^\top X)^{-1} X_p)$. Аналогичный вид имеет и доверительный интервал для индивидуального значения переменной Y , но его стандартная ошибка $s_{\hat{y}_p}$ будет вычисляться иначе: $s_{\hat{y}_p}^2 = s^2(I + X_p^\top (X^\top X)^{-1} X_p)$.

Пример 4.2. Для заданных в файле «regression2.csv» векторов выборочных данных $x_{1,i}$, $x_{2,i}$ и y_i построить линейную модель множественной регрессии и проверить её качество.

```

1 > dat <- read.table("regression2.csv", head=TRUE)
2 > attach(dat); dat
3       y    x1  x2
4    1 12.2 4795 69
5    2   7.6 6962 82
6    3 10.4 6571 87
7    4   9.9 4249 92
8    5 15.7 9540 23
9    6 14.0 3488 31
10   7 12.7 4888 55
11   8 10.5 6237 81
12   9 15.1 2997 65
13  10 10.6 2990 98

```

```

14 11 15.2 1748 100
15 12 17.2 2128 69
16 > fit <- lm(y ~ x1 + x2); summary(fit)
17 Call:
18 lm(formula = y ~ x1 + x2)
19 Residuals:
20      Min       1Q   Median       3Q      Max
21 -2.9490 -1.1543 -0.2731  1.0857  2.8351
22 Coefficients:
23             Estimate Std. Error t value Pr(>|t|)
24 (Intercept) 22.3218043  2.9415849   7.588 3.37e-05 ***
25 x1          -0.0007869  0.0003039  -2.590  0.0292 *
26 x2          -0.0847747  0.0281108  -3.016  0.0146 *
27 ---
28 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
29 Residual standard error: 2.115 on 9 degrees of freedom
30 Multiple R-squared:  0.5596,    Adjusted R-squared:  0.4617
31 F-statistic: 5.717 on 2 and 9 DF,  p-value: 0.02497
32 > x1i <- seq(0.9*min(x1), 1.1*max(x1), len=100)
33 > x2i <- seq(0.9*min(x2), 1.1*max(x2), len=100)
34 > pre <- predict(fit,data.frame(x1=x1i,x2=x2i),interval="confidence")
35 > windows(); plot(x1, y, pch=3)
36 > matplot(x1i, pre, type="l", lty=c(1,2,2), add=TRUE)
37 > windows(); plot(x2, y, pch=3)
38 > matplot(x2i, pre, type="l", lty=c(1,2,2), add=TRUE)
39 > windows(); par(mfrow=c(3,1))
40 > plot(x1, fit[[2]], pch=4); abline(h=0, lty=1)
41 > plot(x2, fit[[2]], pch=4); abline(h=0, lty=1)
42 > qqnorm(fit[[2]]); qqline(as.vector(fit[[2]]))
43 > detach(dat)

```

Функция «`read.table()`» в строке [1] считывает выборочные данные, по-умолчанию сохранённые в файле «`regression2.csv`» из текущей папки, а функция «`attach()`» в строке [2] делает эти данные доступными для расчётов под именами: «`x1`», «`x2`» и «`y`». Структура загруженных из csv-файла данных показана в строках [3–15]. Заметим, что при использовании загружаемых данных с функцией «`attach()`», после проведения расчётов рекомендуется очищать память с использованием «`detach()`», как это сделано в строке [43].

В данном примере мы будем использовать модель множественной линейной регрессии: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Функция «`lm()`» в строке [16] на основе приведённых выборочных данных рассчитывает параметры линейной модели «`y ~ x1 + x2`», которая соответствует регрессии: $y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + e_i$. Сводка основных результатов выводится в строках [17–31] с помощью функции «`summary()`». В строках [24–26] показаны оценки коэффициентов уравнения регрессии: $b_0 = 22.32$, $b_1 = -0.0008$, $b_2 = -0.08$, соот-

ветствующие значения стандартных ошибок и вероятности отклонения гипотез о равенстве полученных оценок истинным значениям: $\mathbf{P}\{\beta_0 \neq b_0\} = 0.00003$, $\mathbf{P}\{\beta_1 \neq b_1\} = 0.03$, $\mathbf{P}\{\beta_2 \neq b_2\} = 0.01$. С учётом значения скорректированного выборочного коэффициента детерминации $R_a^2 = 0.46$, показанного в строке [30], качество уравнения регрессии можно охарактеризовать как умеренное.

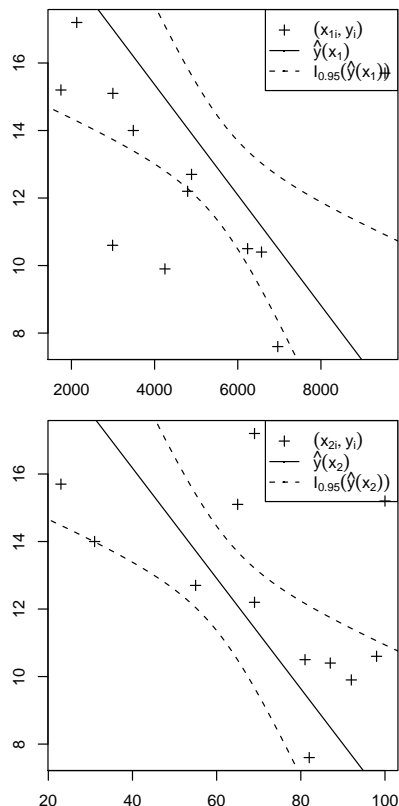


Рис. 4.4. Выборочные значения $(x_{1,i}, y_i)$ и $(x_{2,i}, y_i)$, доверительные интервалы $I_{0.95}(\hat{y})$ для уравнения регрессии $\hat{y} = 22.32 - 0.0008x_1 - 0.08x_2$

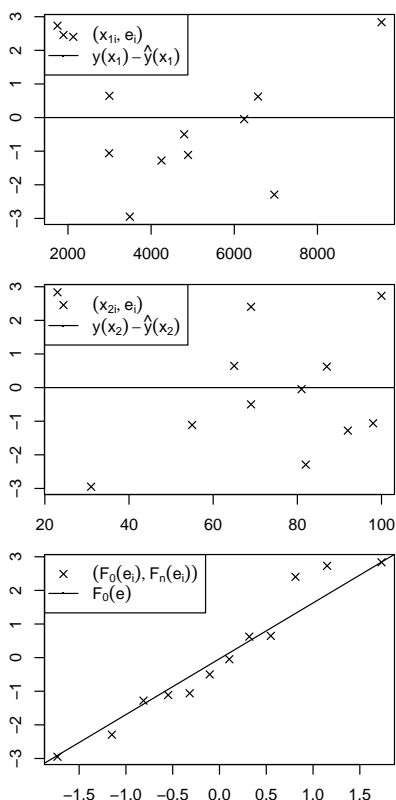


Рис. 4.5. Центрированные относительно \hat{y} графики остатков e_i . Q-Q графики для остатков e_i и функции распределения $F_0(e) = \Phi\left(\frac{e}{1.91}\right) + \frac{1}{2}$

Графики к полученной модели множественной регрессии показана-

ны на рис. 4.4. Для их построения в строках [32–38] и [39–41] используются функции: «`predict(...interval="confidence")`» — вычисление границ 0.95-доверительных интервалов для уравнения регрессии; «`plot(...pch=3)`» — отображение элементов выборки $(x_{1,i}, y_i)$ и $(x_{2,i}, y_i)$, используя символы «+»; «`matplot(...lty=c(1,2,2))`» — отображение графика выборочного уравнения регрессии $\hat{y}(x)$, а также верхней и нижней границ доверительного интервала $I_{0.95}(\hat{y})$, используя сплошную и две штриховые линии; «`windows()`» — создание нового графического окна.

После построения модели и проверки качества полезно провести анализ распределения её остатков e_i , показанных на рис. 4.5 сверху и в середине. Это можно сделать с помощью Q–Q графика, показанного на рис. 4.5 снизу. Для построения этих графиков в строках [39–42] используются функции: «`windows()`» — создание нового графического окна; «`par(mfrow=c(3,1))`» — разбиение графического окна на три части по вертикали; «`plot(...pch=4)`» — отображение остатков e_i , используя символ «x»; «`abline(h=0)`» — отображение горизонтальной линии на нулевом уровне; «`qqnorm()`» — отображение на Q–Q графике остатков e_i для исходных данных линейной модели; «`qqline()`» — отображение на Q–Q графике функции нормального распределения $F_0(e) = \Phi(\frac{e}{1.91}) + \frac{1}{2}$, где значение 1.91 соответствует исправленному выборочному среднему квадратическому отклонению остатков s_e .

Контрольные вопросы

1. Какую зависимость называют регрессионной? В чем отличие регрессионной зависимости от функциональной?
2. Как формулируется задача регрессионного анализа? Из каких соображений выбирается форма регрессионной зависимости?
3. Какой вид имеет линейная регрессионная модель? Как называются переменные, представленные в модели?
4. Какой метод используется для оценки параметров уравнения регрессии? Запишите формулы для МНК-оценок парной регрессии.
5. Как оценивается качество построенного уравнения регрессии? Приведите формулу для расчёта коэффициента детерминации.
6. Сформулируйте условия теоремы Гаусса–Маркова. Какими свойствами будут обладать оценки коэффициентов в случае выполнения этих условий?

7. Как производится проверка значимости построенного уравнения регрессии? Какой критерий при этом используется?
8. Запишите линейную регрессионную модель с k независимыми переменными. Как выглядит система уравнений множественной линейной регрессии в матричной форме?
9. Из каких соображений получается система нормальных уравнений для определения оценок параметров уравнения регрессии? Запишите в матричной форме систему нормальных уравнений.
10. Как проводится дисперсионный анализ для определения значимости уравнения множественной регрессии?
11. Как проверяется значимость коэффициентов уравнения регрессии?
12. Приведите формулы для расчёта доверительного интервала прогнозного значения в случае индивидуальных значений зависимой переменной. В чем отличие случая построения прогноза для функции регрессии?

Литература

1. The R Project for Statistical Computing [Сайт] //URL: <http://www.r-project.org/>
2. The Comprehensive R Archive Network [Сайт] //URL: <http://www.cran.r-project.org/>
3. Воеводин В. В., Воеводин Вл. В. Энциклопедия линейной алгебры. Электронная система ЛИНЕАЛ. — СПб.: ВХВ-Петербург, 2006. — 544 с.
4. БОРОДИН А. Н. Элементарный курс теории вероятностей и математической статистики. — СПб.: Лань, 2004. — 256 с.
5. Кивзун А. И., Горяинова Е. Р., Наумов А. В., Сиротин А. Н. Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами. — М.: ФИЗМАТЛИТ, 2002. — 224 с.
6. Айвазян С. А., Мхитарян В. С. Прикладная статистика. Основы эконометрики. Т.1. — М.: ЮНИТИ-ДАНА, 2001. — 656 с.
7. Айвазян С. А. Прикладная статистика. Основы эконометрики. Т.2. — М.: ЮНИТИ-ДАНА, 2001. — 432 с.
8. КРАМЕР Г. Математические методы статистики. — М.: Мир, 1975. — 648 с.
9. АНДЕРСОН Т. Введение в многомерный статистический анализ. — М.: ФИЗМАТГИЗ, 1963. — 500 с.

Приложение А

Введение в систему R

А.1. Принципы взаимодействия с R

Система статистической обработки данных и программирования **R** ориентирована на использование интерфейса командной строки. Обработка данных в системе **R** представляет собой последовательность команд для загрузки исходных данных, вычислений и текстового или графического вывода полученных результатов. Такая последовательность может быть сформирована пользователем как с помощью командной строки (интерактивный режим), так и из текстового файла (пакетный режим), а текстовые или графические результаты вычислений могут быть выведены на экран и/или записаны в соответствующие файлы.

Для пользователя, привыкшего к графическому интерфейсу, подобный подход может показаться неудобным и устаревшим, но, к счастью, это лишь широко распространённое заблуждение. После отработки основных навыков эффективность обработки данных с использованием клавиатуры и интерфейса командной строки оказываются не ниже, а выше, чем с помощью мыши и графического интерфейса. Одна из причин состоит в том, что вынести в меню и на пиктограммы сотни функций, применяемых в статистическом анализе крайне затруднительно, если вообще возможно, а командная строка **R** принимает любую комбинацию функций, корректную с точки зрения интерпретатора.

А.1.1. Установка и запуск системы R

Общие принципы работы с системой **R** мало зависят от того, под управлением какой операционной системы эта работа выполняется. Однако, существуют детали в установке, настройке и использовании **R**, которые существенным образом зависят от выбранной операционной системы и используемого программного обеспечения.

Подробные инструкции по установке **R** для семейств операционных систем: GNU/Linux и Apple Mac OS X можно найти на сайте про-

екта [1]. Для установки **R** на компьютере с операционной системой семейства Microsoft Windows необходимо загрузить исполняемый файл вида «R-2.11.1-win32.exe», запустить его и следовать инструкциям программы-установщика. После завершения установки на компьютере появится папка, по-умолчанию располагаемая по адресу «C:\Program Files\R\R-2.11.1» или «G:\R\R-2.11.1», где имя «R-2.11.1» будет соответствовать номеру устанавливаемой версии **R**.

Для запуска **R** можно воспользоваться ярлыком на рабочем столе или найти соответствующий раздел в меню «Пуск». В обоих случаях происходит запуск исполняемого файла, по-умолчанию расположенного по адресу «C:\Program Files\R\R-2.11.1\bin\Rgui.exe» или «G:\R\R-2.11.1\bin\Rgui.exe» и загружающего систему **R**.

Заметим, что для упрощения работы с системой **R** можно изменить в ярлыке **R** путь к рабочей папке с установленного по-умолчанию «C:\Program Files\R\R-2.11.1\bin» или «G:\R\R-2.11.1\bin» на папку, действительно содержащую необходимые файлы, например на: «H:\Лабораторные работы в R».

А.1.2. Интерфейс системы R

Главное окно в системе **R** с заголовком «RGui» имеет строку меню и панель инструментов, а в его рабочей области размещаются все остальные окна. В зависимости от выбранного в данный момент рабочего окна и состояния системы **R** строка меню и строка инструментов главного окна меняются и содержат команды системы **R**, актуальные в данный момент для выбранного рабочего окна.

Основным рабочим окном в системе **R** является «R Console». Все команды, вводимые пользователем в этом окне, отмечаются красным цветом и располагаются в самой нижней строке, начинающейся с символа «>», а все выводимые системой **R** ответы отмечаются синим цветом. Вспомогательная информация системы выводится с начала строки, а числовые векторы предваряются символами «[1]». Если выводимый на экран вектор длиннее одной строки, то не уместившиеся элементы вектора переносятся на следующую строку и предваряются символами «[k]», где k соответствует номеру первого в данной строке элемента.

Графическая информация отображается в отдельных окнах с заголовками: «R Graphics: Device k (ACTIVE)», где $k = 2, 3, \dots$ соответствует номеру данного окна в системе **R**, а статус «(ACTIVE)» означает, что все графические команды системы **R** будут влиять на содержимое именно этого окна. Если в рабочей области открыто бо-

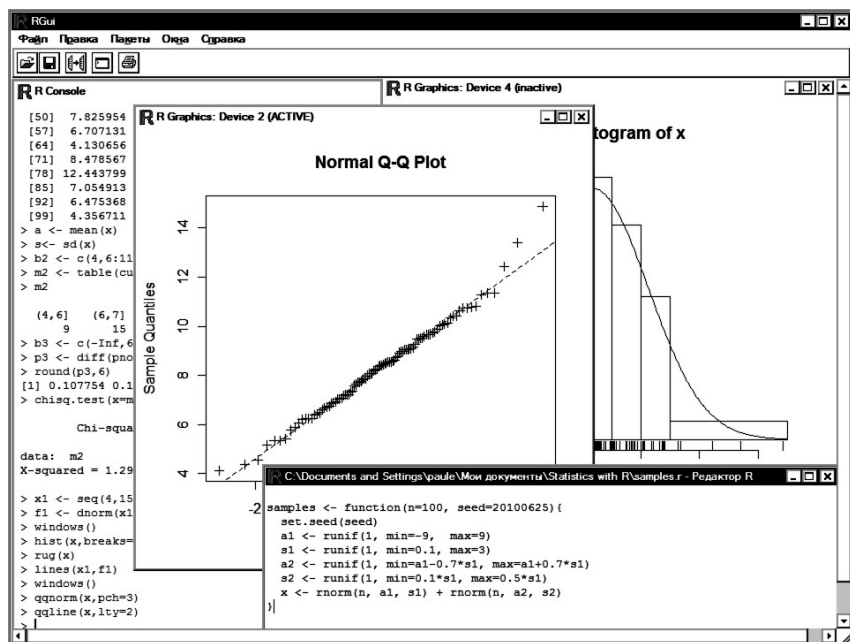


Рис. А.1. Типичное представление рабочей области R в операционной системе Microsoft Windows XP

лее одного графического окна, то статус «(ACTIVE)» может иметь только одно из них, а статус всех остальных окон — «(inactive)».

А.1.3. Справочная информация по R

R — это свободное программное обеспечение, которое пользователи вольны распространять и использовать по собственному усмотрению при соблюдении условий Открытого лицензионного соглашения GNU, известного под аббревиатурой «GNU GPL version 2.1». Для получения актуальных ссылок на полный текст данной лицензии используйте команду «`license()`». R — это проект, в котором участвует множество разработчиков. Для получения контактных данных и другой информации о сообществе, принявшем участие в разработке данной версии R, используйте команду «`contributors()`».

Практически вся справочная информация в системе R представлена на английском языке. Исключение составляют тексты в строке

меню и наиболее распространённые диагностические сообщения об ошибках при выполнении команд системы.

Для доступа к общей справочной информации о системе рекомендуется использовать раздел меню «Справка», а для получения краткой справки по командам и функциям **R** можно использовать «help(full)», где «full» — полное имя искомой команды или функции. В случае, если точное имя функции неизвестно, или когда требуется найти часть слова в наименовании или в описании функции можно использовать команды «help.search("part")», где «part» — часть имени или описания искомой функции.

Помимо этого **R** содержит программы для демонстрации различных возможностей системы. Для получения списка и запуска доступных демонстрационных программ используйте команду «demo()».



А.1.4. Работа с файловой системой в R

При работе с системой **R** важное значение имеет папка, которую **R** считает рабочей. Для того чтобы увидеть полный путь к рабочей папке, можно использовать команду «getwd()», а для того чтобы увидеть её содержимое — команды «dir()» или «dir("Путь/к папке")», при необходимости увидеть содержимое какой-либо другой папки.

Если в процессе работы с **R** возникает необходимость в изменении рабочей папки, то следует использовать команду «setwd("Путь/к новой/рабочей папке")». Заметим, что в качестве разделителя вложенных папок при записи пути в системе **R** используется символ «/», в отличие от символа «\», используемого по-умолчанию в операционных системах Microsoft Windows.

Для пакетного исполнения команд **R**, записанных в простом текстовом файле с именем «script.r», который расположен в рабочей папке, необходимо использовать команду «source("script.r")».

А.1.5. Сохранение данных и выход из R

Для выхода из системы **R** можно использовать графический интерфейс: кнопку закрытия окна или команду меню «Файл|Выход», а можно ввести «q()» в командной строке и согласиться или отказаться от сохранения образа рабочей области **R**, то есть всех объектов в оперативной памяти, а также истории введённых команд. Для вызова в командной строке **R** ранее введённой команды и перемещения по истории команд можно использовать клавиши  и .

Сохранение образа рабочей области полезно в том случае, если обработка данных ещё не завершена, но пользователь вынужден сделать более или менее длительный перерыв, например, в конце лабораторного занятия. Если рабочая область была сохранена при выходе из R, то при следующей загрузке системы её состояние будет восстановлено и пользователь сможет продолжить работу.

В том случае, если ведётся параллельная обработка нескольких наборов данных, то для сохранения существующих данных, очистки оперативной памяти и загрузки новых данных без выхода из R, можно воспользоваться командами: `«save.image(file="oldName.RData")»`, `«rm(list=ls())»` и `«load(file="newName.RData")»`.

Приложение В

Листинги программ

В этом разделе приводятся исходные тексты всех программ, подготовленных для пакетного исполнения. Для запуска любой из приведённых ниже программ её следует записать как текстовый файл в любой доступной для записи папке и ввести в командной строке **R**: «`setwd("путь/к папке/с файлом")`» и «`source("имя файла")`». Обратите внимание, что в качестве разделителей для вложенных папок используются символы прямой косой черты «/» так же, как при записи адреса ресурса в сети Интернет.

В.1. Наиболее распространённые распределения

Для построения графиков различных законов распределения дискретных и непрерывных случайных величин используются функции из файла «`probGraph.r`», листинг которого показан ниже. Поэтому, при запуске всех программ из данного раздела файл «`probGraph.r`» должен находиться в вашей рабочей папке, путь к которой можно увидеть с помощью команды «`getwd()`». Содержимое рабочей папки можно увидеть с помощью команды «`dir()`», а выбрать другую рабочую папку — командой «`setwd("путь/к новой/папке")`».

Визуализация законов распределения

```
1  # Имя файла: probGraph.r
2  #####
3  # Основные обозначения: x, P, F, f - значения дискретных и
4  # непрерывных с.в., а также их вероятности, функции распределения
5  # и плотности вероятности; ltext, cpal - подписи к графикам и
6  # цвет графиков; kk, k - вспомогательные переменные.
7  #####
8  # Графики распределений вероятностей дискретных с.в.
9  #####
10 dgraph <- function(x, P, ltext) { windows()
11     kk <- seq(dim(P)[2]); cpal <- rainbow(max(kk))
12     plot(x, P[,1], col=cpal[1], type="o", pch=1, lwd=1,
```



```

13     xlim=c(0.9*min(x), 1.1*max(x)), ylim=c(0, 1.1*max(P)),
14     xlab="", ylab="", main="")
15     for(k in kk[-1]) lines(x, P[,k],
16         col=cpal[k], type="o", pch=k, lwd=1)
17     legend("topright", col=cpal, pch=kk, lwd=1, legend=ltext)
18 }
19 # Графики функций распределения дискретных с.в.
20 #####
21 pgraph <- function(x, F, ltext) { windows()
22     kk <- seq(dim(F)[2]); cpal <- rainbow(max(kk))
23     plot(stepfun(x, c(0, F[,1])), col=cpal[1], pch=1, lwd=1,
24         xlim=c(0.9*min(x), 1.1*max(x)), ylim=c(0, 1.1*max(F)),
25         xlab="", ylab="", main="")
26     for(k in kk[-1]) lines(stepfun(x, c(0, F[,k])),
27         col=cpal[k], pch=k, lwd=1)
28     legend("bottomright", col=cpal, pch=kk, lwd=1, legend=ltext)
29 }
30 # Графики плотностей вероятностей непрерывных с.в.
31 #####
32 cgraph <- function(x, f, ltext) { windows()
33     kk <- seq(dim(f)[2]); cpal <- rainbow(max(kk))
34     plot(x, f[,1], col=cpal[1], type="l", lwd=1,
35         xlim=c(0.9*min(x), 1.1*max(x)), ylim=c(0, 1.1*max(f)),
36         xlab="", ylab="", main="")
37     for(k in kk[-1]) lines(x, f[,k], col=cpal[k], lwd=1)
38     legend("topright", col=cpal, lwd=1, legend=ltext)
39 }
40 # Графики функций распределения непрерывных с.в.
41 #####
42 fgraph <- function(x, F, ltext) { windows()
43     kk <- seq(dim(F)[2]); cpal <- rainbow(max(kk))
44     plot(x, F[,1], col=cpal[1], type="l", lwd=1,
45         xlim=c(0.9*min(x), 1.1*max(x)), ylim=c(0, 1.1*max(F)),
46         xlab="", ylab="", main="")
47     for(k in kk[-1]) lines(x, F[,k], col=cpal[k], lwd=1)
48     legend("bottomright", col=cpal, lwd=1, legend=ltext)
49 }

```

Биномиальное распределение

```

1 # Графики вероятностей и функции с.в.,
2 # распределённой по биномиальному закону:
3 #####
4 # n, p - параметры биномиального распределения; x - возможные
5 # значения с.в.; P, F - значения вероятностей и функции
6 # распределения с.в.; l - подписи к графикам.
7 #####
8 source("probGraph.r")
9 p <- seq(1, 7, 2)/10; n <- 12; x <- seq(0, n)
10 P <- sapply(p, function(pp) dbinom(x, n, pp))

```

```

11 F <- sapply(p, function(pp) pbinom(x, n, pp))
12 l <- sapply(p, function(pp) sprintf("B(%.0f, %.3g)", n, pp))
13 dgraph(x, P, l); pgraph(x, F, l)

```

Распределение Пуассона

```

1 # Графики вероятностей и функции с.в.,
2 # распределённой по закону Пуассона:
3 # # # # # # # # # # # # # # # # # #
4 # a - параметр распределения Пауссона; x - возможные
5 # значения с.в.; P, F - значения вероятностей и функции
6 # распределения с.в.; l - подписи к графикам.
7 # # # # # # # # # # # # # # # # # #
8 source("probGraph.r")
9 a <- seq(0.5, 3.5, 1)
10 P <- sapply(a, function(aa) dpois(x, aa))
11 F <- sapply(a, function(aa) ppois(x, aa))
12 l <- sapply(a, function(aa) sprintf("P(%.3g)", aa))
13 dgraph(x, P, l); pgraph(x, F, l)

```

Геометрическое распределение

```

1 # Графики вероятностей и функции с.в.,
2 # распределённой по геометрическому закону:
3 # # # # # # # # # # # # # # # # # #
4 # p - параметр геометрического распределения; x - возможные
5 # значения с.в.; P, F - значения вероятностей и функции
6 # распределения с.в.; l - подписи к графикам.
7 # # # # # # # # # # # # # # # # # #
8 source("probGraph.r")
9 p <- seq(3, 9, 2)/10
10 P <- sapply(p, function(pp) dgeom(x, pp))
11 F <- sapply(p, function(pp) pgeom(x, pp))
12 l <- sapply(p, function(pp) sprintf("G(%.3g)", pp))
13 dgraph(x, P, l); pgraph(x, F, l)

```

Равномерное распределение

```

1 # Графики плотности вероятностей и функции
2 # равномерно распределённой с.в.:
3 # # # # # # # # # # # # # # # # # #
4 # a, b - параметры равномерного распределения; x - возможные
5 # значения с.в.; f, F - значения плотности вероятностей и функции
6 # распределения с.в.; l - подписи к графикам.
7 # # # # # # # # # # # # # # # # # #
8 source("probGraph.r")
9 a <- 0; b <- c(1/4, 1/2, 1, 2); x <- seq(-1/4, 2+1/4, len=300)

```



```

9 | a <- 0; s <- c(1/4, 1/2, 1, 2); x <- seq(0, 6, len=300)
10 | f <- sapply(s, function(ss) dlnorm(x, a, ss))
11 | F <- sapply(s, function(ss) plnorm(x, a, ss))
12 | l <- sapply(s, function(ss) sprintf("ln N(%.3g, %.3g)", a, ss))
13 | cgraph(x, f, l); fgraph(x, F, l)

```

Распределение Пирсона

```

1 | # Графики плотности вероятностей и функции с.в.,
2 | # распределённой по закону Пирсона:
3 | #####
4 | # k - параметр распределения Пирсона; x - возможные
5 | # значения с.в.; f, F - значения плотности вероятностей и функции
6 | # распределения с.в.; l - подписи к графикам.
7 | #####
8 | source("probGraph.r")
9 | k <- c(2, 3, 4, 5); x <- seq(0, max(k), len=300)
10 | f <- sapply(k, function(kk) dchisq(x, kk))
11 | F <- sapply(k, function(kk) pchisq(x, kk))
12 | l <- sapply(k, function(kk) sprintf("chi^2(%.0f)", kk))
13 | cgraph(x, f, l); fgraph(x, F, l)

```

Распределение Стьюдента

```

1 | # Графики плотности вероятностей и функции с.в.,
2 | # распределённой по закону Стьюдента:
3 | #####
4 | # k - параметр распределения Стьюдента; x - возможные
5 | # значения с.в.; f, F - значения плотности вероятностей и функции
6 | # распределения с.в.; l - подписи к графикам.
7 | #####
8 | source("probGraph.r")
9 | k <- c(2, 3, 4, 300); x <- seq(-6, 6, len=300)
10 | f <- sapply(k, function(kk) dt(x, kk))
11 | F <- sapply(k, function(kk) pt(x, kk))
12 | l <- sapply(k, function(kk) sprintf("t(%.0f)", kk))
13 | cgraph(x, f, l); fgraph(x, F, l)

```

Распределение Фишера

```

1 | # Графики плотности вероятностей и функции с.в.,
2 | # распределённой по закону Фишера:
3 | #####
4 | # k1, k2 - параметры распределения Фишера; x - возможные
5 | # значения с.в.; f, F - значения плотности вероятностей и функции
6 | # распределения с.в.; l - подписи к графикам.
7 | #####

```

В.2. Основы математической статистики

Генерирование реализации случайной выборки

Основные выборочные характеристики

```
1 # Вычисление основных характеристик реализации случайной
2 # выборки с законом распределения, близким к нормальному:
3 # # # # # # # # # # # # # # # # # # # # # # # # # # # #
4 # n - объём, х - реализация случайной выборки;
5 # seed - начальное состояние генератора п.с.ч.;
```

```

6 # a1, s1 - выборочные среднее и среднеквадратическое отклонение;
7 # x2 - регулярный вектор абсцисс в интервале [a1-4*s1, a1+4*s1];
8 # f1, F1 - оценки плотности вероятности и функции нормального
9 #           распределения ~ N(a1,s1).
10 #####
11 source("samples.r")
12 n <- 100; x <- samples(n); a1 <- mean(x); s1 <- sd(x)
13 x2 <- seq(a1-4*s1, a1+4*s1, len=n)
14 f1 <- dnorm(x2, a1, s1); F1 <- pnorm(x2, a1, s1)
15 ltext <- sprintf("X~N(%.2f, %.2f)",a1,s1)
16 windows(); hist(x, breaks="Scott", xlim=range(x2), ylim=c(0,
17               1.2*max(f1)), freq=FALSE, main="", xlab="", ylab="")
18 rug(x); lines(x2, f1); box()
19 legend("topleft", lty=1, legend=paste("f(x):",ltext))
20 windows(); plot(ecdf(x), pch=".", xlim=range(x2),
21               main="", xlab="", ylab="")
22 rug(x); lines(x2, F1)
23 legend("topleft", lty=1, legend=paste("F(x):",ltext))

```

Интервальные оценки параметров распределения

```

1 # Построение реализаций доверительных интервалов для
2 # параметров нормально распределённой случайной величины:
3 #####
4 # n, x - объём и реализация случайной выборки; g - доверительная
5 # вероятность; ci{M,D}{n,g} - границы доверительных интервалов для
6 # MX и DX при постоянной доверительной вероятности и объёме выборки.
7 #####
8 set.seed(20100625)
9 n <- seq(100, 1000, 20); g <- seq(0.95, 0.995, length(n))
10 ciM <- function(x,n,g) mean(x)-sd(x)/sqrt(n)*qt((1+c(g,0,-g))/2,n-1)
11 ciD <- function(x,n,g) sd(x)^2*(n-1)/qchisq((1+c(g,0,-g))/2,n-1)
12 ciMn <- sapply(n, function(nn) ciM(rnorm(nn), nn, g[1]))
13 ciDn <- sapply(n, function(nn) ciD(rnorm(nn), nn, g[1]))
14 ciMg <- sapply(g, function(gg) ciM(rnorm(n[1]), n[1], gg))
15 ciDg <- sapply(g, function(gg) ciD(rnorm(n[1]), n[1], gg))
16 txtMn <- sprintf("MX(n,g=%.3g)",g[1])
17 txtDn <- sprintf("DX(n,g=%.3g)",g[1])
18 txtMg <- sprintf("MX(g,n=%.0f)",n[1])
19 txtDg <- sprintf("DX(g,n=%.0f)",n[1])
20 ci_graph <- function(x, y, point, text) { windows()
21   plot(range(x), range(y), type="n", xlab="", ylab="")
22   for(j in seq(length(y))) {
23     lines(x[,j], rep(y[j],3), lwd=2)
24     points(x[2,j], y[j], pch=16, lwd=2) }
25   legend("topright", legend=text, bg="white")
26   abline(v=point, lty=2, lwd=2) }
27 ci_graph(ciMn, n, 0, txtMn)
28 ci_graph(ciMg, g, 0, txtMg)
29 ci_graph(ciDn, n, 1, txtDn)

```

```
30 | ci_graph(ciDg, g, 1, txtDg)
```

```

3  #####
4  # x, u, v - исходная реализация случайной выборки и две её части.
5  #####
6  source("samples.r")
7  x <- samples(); u <- x[1:49]; v <- x[50:100]; print(ks.test(u,v))
8  plot(ecdf(u), pch=25, cex=0.5, xlim=range(x)); rug(u, side=1)
9  plot(ecdf(v), pch=24, cex=0.5, add=TRUE); rug(v, side=3)

```

Одновыборочный t -критерий значимости различий

```

1  # Зависимость достигаемого уровня значимости от параметра а
2  # для реализации случайной выборки по одновыборочному t-критерию
3  # значимости различий для нулевой гипотезы о соответствии выборочного
4  # среднего вышеуказанному значению параметра а:
5  #####
6  # x - реализация случайной выборки; a, s - выборочные среднее и
7  # среднеквадратическое отклонение; p, pl, pg - достигаемые уровни
8  # значимости при альтернативных гипотезах о неравенстве,
9  # меньшем или большем истинном значении.
10 #####
11 source("samples.r"); x <- samples()
12 a <- mean(x); s <- sd(x); a <- seq(a-s/2, a+s/2, length=99)
13 p <- sapply(a, function(aa) t.test(x, mu=aa, alter="two")[[3]])
14 pl <- sapply(a, function(aa) t.test(x, mu=aa, alter="le")[[3]])
15 pg <- sapply(a, function(aa) t.test(x, mu=aa, alter="gr")[[3]])
16 plot(a, p, type="l"); lines(a, pl, lty=2); lines(a, pg, lty=4)
17 abline(h=c(0,0.05), lty=3)

```

Двухвыборочный t -критерий значимости различий

```

1  # Проверка гипотезы о равенстве средних двух частей реализации
2  # случайной выборки по двухвыборочному t-критерию:
3  #####
4  # x, u, v - исходная реализация случайной выборки и две её части.
5  #####
6  source("samples.r")
7  x <- samples(); u <- x[1:49]; v <- x[50:100]
8  print(t.test(u,v, var.equal=TRUE, alter="two"))
9  print(t.test(u,v, var.equal=TRUE, alter="le"))
10 print(t.test(u,v, var.equal=TRUE, alter="gr"))

```

Фишера F -критерий значимости различий

```

1  # Проверка гипотезы о равенстве дисперсий двух частей реализации
2  # случайной выборки по двухвыборочному F-критерию:
3  #####
4  # x, u, v - исходная реализация случайной выборки и две её части.

```



```

5  #####
6  source("samples.r")
7  x <- samples(); u <- x[1:49]; v <- x[50:100]
8  print(var.test(u,v, alter="two"))
9  print(var.test(u,v, alter="le"))
10 print(var.test(u,v, alter="gr"))

```

Однофакторный дисперсионный анализ

```

1  # Проверка значимости влияния изменений качественного признака
2  # на величину количественного признака с помощью методики
3  # однофакторного дисперсионного анализа.
4  #####
5  # D, B, S - значения количественного признака, наблюдаемые
6  # на каждом из уровней качественного признака;
7  # adhf - вспомогательная таблица.
8  #####
9  D = c(4.0,4.5,4.3,5.6,4.9,5.4,3.8,3.7,4.0)
10 B = c(4.5,4.9,5.0,5.7,5.5,5.6,4.7,4.5,4.7)
11 S = c(5.4,4.9,5.6,5.8,6.1,6.3,5.5,5.0,5.0)
12 adhf = stack(data.frame(D,B,S))
13 print(anova(lm(values ~ ind, data=adhf)))

```

В.3. Начала регрессионного анализа

Для решения задач регрессионного анализа в данном разделе используются показанные ниже эмпирические данные, которые сохраняются в файлах «regression1.csv» и «regression2.csv». Поэтому, при запуске программ из данного раздела требуемый csv-файл должен находиться в вашей рабочей папке, путь к которой можно увидеть с помощью команды «getwd()». Содержимое рабочей папки можно увидеть с помощью команды «dir()», а выбрать другую рабочую папку — командой «setwd("путь/к новой/папке")».

Содержимое файлов данных

```

1  # Имя файла:
2  # regression1.csv | regression2.csv
3  #####
4  x      y      | y      x1     x2
5  2.36   1.12   | 12.2   4795   69
6  2.67   0.46   | 7.6    6962   82
7  2.98   0.19   | 10.4   6571   87
8  3.30  -0.27   | 9.9    4249   92
9  3.61  -0.85   | 15.7   9540   23

```

10	3.93	-0.79		14.0	3488	31
11	4.24	-1.17		12.7	4888	55
12	4.56	-1.88		10.5	6237	81
13	4.87	-1.62		15.1	2997	65
14	5.18	-1.25		10.6	2990	98
15	5.50	-1.04		15.2	1748	100
16				17.2	2128	69

Модель парной линейной регрессии

```

1  # Для заданных в файле "regression1.csv"
2  # выборочных векторов x и y построить линейную модель
3  # парной регрессии и проверить её качество:
4  #####
5  # dat, x, y - таблица и векторы выборочных данных; fit -
6  # линейная модель парной регрессии; xin - вектор абсцисс;
7  # pre - матрица размера 3*n, содержащая ординаты уравнения
8  # регрессии и границы его 0.95-доверительных интервалов.
9  #####
10 dat <- read.table("regression1.csv", head=TRUE)
11 attach(dat); print(dat)
12 fit <- lm(y ~ x); print(summary(fit))
13 xin <- seq(0.9*min(x), 1.1*max(x), length=100)
14 pre <- predict(fit, data.frame(x=xin), interval="confidence")
15 windows(); plot(dat, pch=3); points(mean(x), mean(y))
16 matplot(xin, pre, type="l", lty=c(1,2,2), add=TRUE)
17 windows(); par(mfrow=c(2,1))
18 plot(x, fit[[2]], pch=4); abline(h=0)
19 qqnorm(fit[[2]], pch=4); qqline(as.vector(fit[[2]]))
20 detach(dat)

```

Модель множественной линейной регрессии

```

1  # Для заданных в файле "regression2.csv"
2  # выборочных векторов x1, x2 и y построить линейную модель
3  # множественной регрессии и проверить её качество:
4  #####
5  # dat, x1, x2, y - таблица и векторы выборочных данных; fit -
6  # линейная модель множественной регрессии; x1i, x2i - векторы абсцисс;
7  # pre - матрица размера 3*n, содержащая ординаты уравнения регрессии
8  # и границы его 0.95-доверительных интервалов.
9  #####
10 dat <- read.table("regression2.csv", head=TRUE)
11 attach(dat); print(dat)
12 fit <- lm(y ~ x1 + x2); print(summary(fit))
13 x1i <- seq(0.9*min(x1), 1.1*max(x1), len=100)
14 x2i <- seq(0.9*min(x2), 1.1*max(x2), len=100)
15 pre <- predict(fit, data.frame(x1=x1i,x2=x2i), interval="confidence")

```

```
16 windows(); plot(x1, y, pch=3)
17 matplot(x1i, pre, type="l", lty=c(1,2,2), add=TRUE)
18 windows(); plot(x2, y, pch=3)
19 matplot(x2i, pre, type="l", lty=c(1,2,2), add=TRUE)
20 windows(); par(mfrow=c(3,1))
21 plot(x1, fit[[2]], pch=4); abline(h=0, lty=1)
22 plot(x2, fit[[2]], pch=4); abline(h=0, lty=1)
23 qqnorm(fit[[2]], pch=4); qqline(as.vector(fit[[2]]))
24 detach(dat)
```

Учебное издание

Буховец Алексей Георгиевич
Москалев Павел Валентинович
Богатова Вера Павловна
Бирючинская Татьяна Яковлевна

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В СИСТЕМЕ R

Учебное пособие
издаётся в авторской редакции

Редактор: А. Г. Буховец
Корректор: С. А. Дубова
Вёрстка: П. В. Москалев

Подписано в печать 02.11.2010 г. Формат $60 \times 84 \frac{1}{16}$.
Гарнитура «Concrete». Печать офсетная.
Бумага офсетная. Объем 7,6 п.л. Тираж 140 экз.
Заказ № 4593

Федеральное государственное образовательное учреждение
высшего профессионального образования
«Воронежский государственный аграрный университет
им. К. Д. Глинки» Типография ФГОУ ВПО ВГАУ
394087 г. Воронеж, ул. Мичурина, 1