А. Г. БУХОВЕЦ,
П. В. МОСКАЛЕВ

АЛГОРИТМЫ ВЫЧИСЛИТЕЛЬНОЙ СТАТИСТИКИ В СИСТЕМЕ R

Издание второе, переработанное и дополненное

ДОПУЩЕНО
УМО по образованию в области прикладной информатики
в качестве учебного пособия для студентов вузов,
обучающихся по направлению «Прикладная информатика»

САНКТ-ПЕТЕРБУРГ · МОСКВА · КРАСНОДАР
2015

ББК 22.18я73
Б 68

Буховец А. Г., Москалев П. В.

Б 68 Алгоритмы вычислительной статистики в системе R: Учебное пособие. — 2-е изд., перераб. и доп. — СПб.: Издательство «Лань», 2015. — 160 с.: ил. — (Учебники для вузов. Специальная литература).

ISBN 978-5-8114-1802-2

В учебном пособии в краткой форме излагается теоретический материал и приводятся примеры решения практических задач по разделам: линейная алгебра, теория вероятностей, методы оценивания и проверки гипотез, метод главных компонент, регрессионный и кластерный анализ с применением свободной системы статистической обработки данных и программирования R. В приложениях к настоящему пособию содержатся сведения по установке и использованию системы R, а также листинги программ, которые могут быть использованы в учебном процессе.

Учебное пособие предназначено для студентов, обучающихся по направлению «Прикладная информатика», программа которых предусматривает изучение современных средств и методов вычислительной статистики.

ББК 22.18я73

Рецензенты:
Ю. Ю. ТАРАСЕВИЧ — доктор физико-математических наук, профессор, зав. кафедрой прикладной математики и информатики Астраханского государственного университета;
М. Г. МАТВЕЕВ — доктор технических наук, профессор, зав. кафедрой информационных технологий управления Воронежского государственного университета.

Обложка
Е. А. ВЛАСОВА

Оглавление

Введение

Предлагаемое вниманию читателей учебное пособие рассчитано для студентов бакалавриата, проходящих подготовку по направлению «Прикладная информатика», которые как самостоятельно, так и под руководством преподавателя занимаются изучением методов проведения статистического анализа данных с помощью современных программных средств. В главах 1–2 настоящего пособия в краткой форме излагаются основные сведения из линейной алгебры и теории вероятностей, необходимые для адекватного понимания последующего материала. Главы 3–6 посвящены изложению методов статистического оценивания и проверки гипотез, метода главных компонент, а также основ регрессионного и кластерного анализа данных.

Материал первой и второй глав имеет справочный характер и сопровождается относительно простыми примерами, иллюстрирующими базовые свойства векторов, матриц и операций над ними, а также функций распределения и числовых характеристик случайных величин для некоторых, наиболее распространённых законов распределения.

Основной теоретический материал излагается в третьей, четвёртой, пятой, шестой главах и иллюстрируется более развёрнутыми примерами, ориентированными на практические задачи вычислительной статистики и методов статистического анализа данных. Завершается учебное пособие приложениями с описанием базовых сведений по установке и применению системы статистической обработки данных и программирования **R**, а также с листингами примеров на языке **R**, оформленными с учётом возможности 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 вре-

мя 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 быстродействия, 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 ускорялся и в скором времени возникла распределенная система хранения и распространения библиотек 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 качественную прикладную библиотеку, реализующую 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 и Microsoft Windows, а в распределённых хранилищах системы CRAN по состоянию на апрель 2014 года были доступны для свободной загрузки свыше 5400 библиотек 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 [3,10].

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* $\boldsymbol{x}$ как столбец, состоящий из n действительных чисел x_i, записанных в определённом порядке $i = 1, 2, \ldots, n$ и называемых *координатами или компонентами* вектора

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

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

1. *Сложение* векторов $\boldsymbol{x}$ и $\boldsymbol{y}$

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

2. *Умножение* вектора $\boldsymbol{x}$ на вещественное число α

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

Для этих операций справедливы следующие *свойства* векторного пространства:

1. $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$, $\alpha \boldsymbol{x} = \boldsymbol{x}\alpha$;
2. $\boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) = (\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z}$, $\alpha(\beta \boldsymbol{x}) = (\alpha\beta)\boldsymbol{x}$;
3. $\alpha(\boldsymbol{x} + \boldsymbol{y}) = \alpha\boldsymbol{x} + \alpha\boldsymbol{y}$, $(\alpha + \beta)\boldsymbol{x} = \alpha\boldsymbol{x} + \beta\boldsymbol{x}$;
4. $0\boldsymbol{x} = \boldsymbol{o}$, $\boldsymbol{x} + \boldsymbol{o} = \boldsymbol{x}$, где $\boldsymbol{o}$ — *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 $\boldsymbol{R}^n$.

Пример 1.1. В качестве примера проиллюстрируем вышеуказанные свойства векторов с помощью языка **R**.

```
1  > x <- c(8,1,9); y <- c(8,5,7)
2  > z <- c(7,1,5); o <- c(0,0,0)
3  > all(x+y == y+x); all(4*x == x*4)
4  [1] TRUE
5  [1] TRUE
6  > all(x+(y+z) == (x+y)+z); all(4*(9*x) == (4*9)*x)
7  [1] TRUE
8  [1] TRUE
9  > all(3*(x+y) == 3*x + 3*y); all((3+1)*x == 3*x+1*x)
10 [1] TRUE
11 [1] TRUE
12 > all(x+o == x); all(0*x == o)
13 [1] TRUE
14 [1] 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 «c()» поэлементно определяются значения векторов $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{o}$, 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». Функция «all()» используется для сокращённой записи результата проверки свойств вектора и возвращает истинное значение лишь при истинности значений всех элементов вектора аргумента. Нетрудно убедиться, что для приведённых исходных данных все перечисленные свойства векторного пространства выполняются.

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

1.2. Базис векторного пространства

Линейной комбинацией векторов $\boldsymbol{x}_i$, $i = 1, 2, \ldots, k$ в пространстве $\boldsymbol{R}^n$ называется выражение вида

$$\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2 + \ldots + \alpha_k \boldsymbol{x}_k = \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i.$$

Система векторов $\boldsymbol{x}_i$, $i = 1, 2, \ldots, k$ называется *линейно независимой*, если равенство

$$\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2 + \ldots + \alpha_k \boldsymbol{x}_k = 0$$

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 $\{\boldsymbol{e}_i\}$, $i = 1, 2, \ldots, n$ называется *базисом векторного пространства* $\boldsymbol{R}^n$, если любой вектор этого пространства $\boldsymbol{x}$ может быть представлен в виде линейной комбинации векторов базиса

$$\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + \ldots + x_n\boldsymbol{e}_n.$$

Это равенство называется *разложением вектора* $\boldsymbol{x}$ по базису $\{\boldsymbol{e}_i\}$, 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 $\{\boldsymbol{e}_i\}$, $i = 1, 2, \ldots, n$, совпадает с *размерностью* векторного пространства, которая обозначается как $\dim \boldsymbol{R}^n = n$.
2. Любой вектор n-мерного векторного пространства *единственным образом* раскладывается по заданному базису $\{\boldsymbol{e}_i\}$, $i = 1, 2, \ldots, n$.

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

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

Очевидно, что размерность линейного подпространства не превосходит размерности линейного пространства $\dim \boldsymbol{L} \leqslant \dim \boldsymbol{R}^n$.

Совокупность всех линейных комбинаций векторов $\{\boldsymbol{x}_i\}$, где $i = 1, 2, \ldots, k$ называется *линейной оболочкой* этих векторов.

Пример 1.2. Продолжая предыдущий пример, найдём координаты вектора $\boldsymbol{a}(-1, -1, 1)$ в базисе $\boldsymbol{x}(8, 1, 9)$, $\boldsymbol{y}(8, 5, 7)$, $\boldsymbol{z}(7, 1, 5)$ с помощью языка **R**.

Напомним, что решение этой задачи сводится к решению системы линейных алгебраических уравнений, в которой столбцы векторов базиса $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ формируют матрицу коэффициентов, а разлагаемый по базису вектор $\boldsymbol{a}$ — столбец свободных членов:

$$\begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix} \begin{pmatrix} a'_1 \\ a'_2 \\ a'_3 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}.$$

```
15 > x; y; z
16 [1] 8 1 9
17 [1] 8 5 7
18 [1] 7 1 5
19 > a <- c(-1,-1,1)
20 > d <- matrix(c(x,y,z), nrow=length(x), byrow=TRUE)
21 > if (det(d) != 0) solve(d,a) else
22 + stop("Векторы линейно зависимы!")
23 [1]  0.6666667 -0.3333333 -0.6666667
```

Так как векторы $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ уже были определены ранее 1–2, то для постановки задачи достаточно лишь убедиться в существовании одноимённых переменных 15–18 и определить вектор $\boldsymbol{a}$ 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)» с числом строк, определяемым длиной первого вектора «nrow=length(x)» 20, а условие «det(d) != 0» используется для проверки линейной независимости векторов «x,y,z», что является необходимым и достаточным условием для существования одноимённого базиса. Если же указанное условие не будет выполнено, то вместо искомых координат в строке 23 будет выдано сообщение об ошибке.

Символ «+» в начале строки 22 появляется при переносе незавершённого выражения с предыдущей строки.

Как видно из приведённого в строке 23 ответа, искомые координаты вектора $\boldsymbol{a}'$ в базисе $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ будут равны $\boldsymbol{a}'(-\frac{2}{3}, -\frac{1}{3}, -\frac{2}{3})$.

1.3. Скалярное произведение векторов

Скалярным произведением векторов $\boldsymbol{x}$ и $\boldsymbol{y}$ называется число (скаляр), обозначаемое как $(\boldsymbol{x}, \boldsymbol{y})$ или просто $\boldsymbol{xy}$ и определяемое соотношением

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

Основные *свойства* скалярного произведения:

1. $(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{y}, \boldsymbol{x})$;
2. $(\boldsymbol{x}_1 + \boldsymbol{x}_2, y) = (\boldsymbol{x}_1, \boldsymbol{y}) + (\boldsymbol{x}_2, \boldsymbol{y})$;
3. $(\alpha\boldsymbol{x}, \boldsymbol{y}) = \alpha(\boldsymbol{x}, \boldsymbol{y})$ для любого вещественного α;
4. $(\boldsymbol{x}, \boldsymbol{x}) = |\boldsymbol{x}|^2 \geqslant 0$, причём $|\boldsymbol{x}| = 0$ тогда и только тогда, когда $\boldsymbol{x} = \boldsymbol{o}$, где $|\boldsymbol{x}| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}$ — *модуль* или *длина* вектора $\boldsymbol{x}$.

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

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

Векторы $\boldsymbol{x}$ и $\boldsymbol{y}$ называются *ортогональными*, если их скалярное произведение равно нулю: $(\boldsymbol{x}, \boldsymbol{y}) = 0$.

Вещественное линейное пространство называется *евклидовым*, если в нём определено скалярное произведение элементов. В евклидовом пространстве удобно использовать базис $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$, все элементы которого взаимно ортогональны и имеют единичную длину:

$$(\boldsymbol{e}_i, \boldsymbol{e}_j) = \delta_{ij}, \quad \delta_{ij} = \begin{cases} 1, & \text{если } i = j; \\ 0, & \text{если } i \neq j, \end{cases}$$

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

$$\boldsymbol{x} = \sum_{i=1}^{n} (\boldsymbol{x}, \boldsymbol{e}_i)\boldsymbol{e}_i.$$

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 (МНК), а также для объяснения свойств МНК-оценок.

Пример 1.3. В продолжение предыдущего примера выясним ортогональность вектора $\boldsymbol{a}$ с базисом $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ с помощью языка **R**.

```
24 > a
25 [1]  -1 -1  1
26 > d
27        [,1] [,2] [,3]
28 [1,]    8    1    9
29 [2,]    8    5    7
30 [3,]    7    1    5
31 > as.vector(d%*%a)
32 [1]  0 -6 -3
```

Для проверки ортогональности вектора $\boldsymbol{a}$ с векторами указанного базиса потребуется вычислить три скалярных произведения: $(\boldsymbol{x}, \boldsymbol{a})$, $(\boldsymbol{y}, \boldsymbol{a})$, $(\boldsymbol{z}, \boldsymbol{a})$. 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 первая пара векторов: $(\boldsymbol{x}, \boldsymbol{a}) = 0 \;\Rightarrow\; \boldsymbol{x}(8, 1, 9) \perp \boldsymbol{a}(-1, -1, 1)$.

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:

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

Составляющие матрицу числа a_{ij}, где $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, 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 $\boldsymbol{R}^m$.

Главной диагональю квадратной матрицы порядка n называется совокупность элементов: a_{ij}, $i = j = 1, 2, \ldots, n$. 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

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

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

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

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 **R**.

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

```
[3,]    2   -7   -8
> matrix(round(runif(9, min=-9, max=9)), nrow=3) -> B; B
       [,1] [,2] [,3]
[1,]    5   -6   -7
[2,]   -5    6    9
[3,]    3    4   -5
> matrix(round(runif(9, min=-9, max=9)), nrow=3) -> C; C
       [,1] [,2] [,3]
[1,]    0    4   -2
[2,]    9    5    8
[3,]    0   -4    6
> matrix(0, nrow=3, ncol=3) -> O
> all(A+B == B+A); all(7*A == A*7)
[1] TRUE
[1] TRUE
> all((A+B)+C == A+(B+C)); all(3*(4*A) == (3*4)*A)
[1] TRUE
[1] TRUE
> all(3*(A+B) == 3*A+3*B); all((3+4)*A == 3*A+4*A)
[1] TRUE
[1] TRUE
> all(A+O == A); all(0*A == O)
[1] TRUE
[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 девяти 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 нулевое значение 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, ранг

Транспонированием матрицы A называется операция, в результате которой меняются местами строки и столбцы матрицы при со-

хранении порядка их следования. Полученная в результате этого матрица называется *транспонированной* и обозначается:

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

Свойства операции транспонирования:

1. $(A^{\mathsf{T}})^{\mathsf{T}} = A$;
2. $(A + B)^{\mathsf{T}} = A^{\mathsf{T}} + B^{\mathsf{T}}$.

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

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

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. $(AB)^{\mathsf{T}} = B^{\mathsf{T}} A^{\mathsf{T}}$.

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: если $AB = BA$, 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 $\boldsymbol{x}$ и $\boldsymbol{y}$ также может быть записано в виде матричного произведения: $(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{y}$.

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 $\operatorname{rank} A$ или $\mathrm{r}(A)$. Из определения очевидно, что $0 \leqslant \operatorname{rank} A \leqslant \min(n, m)$. Для нулевой матрицы полагают, что $\operatorname{rank} A = 0$.

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 > library(Matrix)
55 > c(rankMatrix(A), rankMatrix(B), rankMatrix(C))
56 [1] 3 3 3
```

Для определения ранга матриц A, B, C в строках 54–55 загружается библиотека «Matrix» и трижды вызывается функция «rankMatrix», определяющая ранг передаваемой в качестве аргумента матрицы. Как видно из строки 56, ранги матриц A, B, C оказались равными их порядку: $\operatorname{rank} A = \operatorname{rank} B = \operatorname{rank} C = 3$.

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 $i, j = 1, 2, \ldots, n$; 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: $\det A^{\mathsf{T}} = \det A$;
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*:

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

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

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

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

Для прямоугольных матриц существует *псевдообращение Мура–Пенроуза*. Матрица A^{+} называется псевдообратной для прямоугольной матрицы A, если удовлетворяет равенствам: а) $AA^{+}A = A$; б) $A^{+}AA^{+} = A^{+}$; в) $AA^{+} = (AA^{+})^{\mathsf{T}}$; г) $A^{+}A = (A^{+}A)^{\mathsf{T}}$.

Основные *свойства псевдообращений* прямоугольных матриц:

1. $(A^{+})^{+} = A$;
2. $(A^{\mathsf{T}})^{+} = (A^{+})^{\mathsf{T}}$;
3. $(\alpha A)^{+} = \alpha^{-1}A^{+}$ для $\alpha \neq 0$;
4. $(A^{\mathsf{T}}A)^{+}A^{\mathsf{T}} = A^{\mathsf{T}}(AA^{\mathsf{T}})^{+} = A^{+}$.

Взаимосвязь обращения и псевдообращения зависит от линейной независимости строк и/или столбцов матрицы A. В частности:

1. Если линейно независимы столбцы матрицы A, то это означает, что произведение $A^{\mathsf{T}}A$ будет невырожденной квадратной матрицей, а псевдообращение будет эквивалентно $A^{+} = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$;
2. Если линейно независимы строки матрицы A, то это означает, что невырожденной квадратной матрицей будет произведение AA^{T}, а псевдообращение будет эквивалентно $A^{+} = A^{\mathsf{T}}(AA^{\mathsf{T}})^{-1}$;
3. Если линейно независимы как строки, так и столбцы матрицы A, то это означает, что невырожденной квадратной будет матрица A, а псевдообращение будет эквивалентно её обращению $A^{+} = A^{-1}$;
4. Если определено произведение матриц AB, причём у матрицы A линейно независимы столбцы, а у матрицы B — строки, то для псевдообращения их произведения будет верно $(AB)^{+} = B^{+}A^{+}$.

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

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

$$(A - \lambda I)\boldsymbol{x} = \boldsymbol{o}.$$

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 $\boldsymbol{x}_i$ 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**.

```
57 > det(A) - det(t(A))
58 [1] 3.410605e-13
59 > det(A) - det(t(A)) < 1e-6
60 [1] TRUE
```

Важной особенностью функции «det()», вычисляющей определитель матрицы, является приближенный характер получаемых результатов, что видно из 57–58. Запись вида «3.410605e-13» означает весьма близкое, но не равное нулю число, соответствующее заданной предельно допустимой погрешности вычисления определителя: $3.410605 \cdot 10^{-13}$.

В связи с этим, в строке 59 вместо проверки логического равенства, соответствующего первому свойству определителей, мы вычисляем разность между правой и левой частями равенства с последующей проверкой уровня полученной погрешности.

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

В строках 59–81 иллюстрируются основные свойства определителей. Записи вида «A[,1]» и «A[,2]» в 65 означают обращения к первому и второму столбцам матрицы «A», а запись вида «A[,c(2,1,3)]» в 68 — перестановку первого и второго её столбцов.

Символ «+» в начале строк 75 и 79 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 так далее.

Выражения вида «det(A[-1,-1])» в строках 74–75 и 78–79 означают определитель матрицы A без первой строки и первого столбца, то есть минор к элементу a_{11}. Таким образом, в строках 74–75 записано разложение определителя матрицы A по первой строке, а в строках 78–79 записана сумма произведений элементов первой строки матрицы A на алгебраические дополнения к элементам её второй строки.

```
82 > sum(A%*%solve(A) - diag(3)) < 1e-6
83 [1] TRUE
84 > sum(solve(A%*%B) - solve(B)%*%solve(A)) < 1e-6
85 [1] TRUE
86 > sum(t(solve(A)) - solve(t(A))) < 1e-6
87 [1] TRUE
88 > det(solve(A)) - det(A)^-1 < 1e-6
89 [1] TRUE
```

В строках 82–89 иллюстрируются основные свойства обратных матриц. Функция «diag(3)» в строке 82 используется для получения единичной матрицы третьего порядка. Для вычисления обратной матрицы используется та же функция «solve()», что и для решения системы линейных алгебраических уравнений, но только с одним аргументом: 82, 84, 86, 88. В тех случаях, когда результат предполагал появление нулевой матрицы, использовалась её свёртка с помощью функции суммирования «sum()»: 82, 84, 86.

```
90  > D <- matrix(round(runif(12, min=-9, max=9)), nrow=3); D
91       [,1] [,2] [,3] [,4]
92  [1,]    0   -1    3    8
93  [2,]    7   -4   -6   -2
94  [3,]   -4   -4   -1   -4
95  > library(MASS)
96  > sum(ginv(ginv(D)) - D) < 1e-6
97  [1] TRUE
98  > sum(ginv(t(D)) - t(ginv(D))) < 1e-6
99  [1] TRUE
100 > sum(ginv(3*D) - ginv(D)/3) < 1e-6
101 [1] TRUE
```

```
> sum(ginv(t(D)%*%D)%*%t(D) - ginv(D)) < 1e-6
[1] TRUE
> sum(t(D)%*%ginv(D%*%t(D)) - ginv(D)) < 1e-6
[1] TRUE
> sum(solve(t(D)%*%D)%*%t(D) - ginv(D)) < 1e-6
Ошибка в solve.default(t(D) %*% D) :
  система вычислительно сингулярная: величина, обратная к числу
  обусловленности, равна 1.73312e-17
> det(t(D)%*%D) < 1e-6
[1] TRUE
> sum(t(D)%*%solve(D%*%t(D)) - ginv(D)) < 1e-6
[1] TRUE
> sum(solve(A) - ginv(A)) < 1e-6
[1] TRUE
> sum(ginv(A%*%D) - ginv(D)%*%ginv(A)) < 1e-6
[1] TRUE
```

Основные свойства псевдообращения Мура–Пенроуза иллюстрируются в строках 96–117 на примере прямоугольной матрицы «D» 90–94 и квадратной матрицы «A». Для вычисления псевдообращения Мура–Пенроуза используется функция «ginv()» из библиотеки «MASS» 95. Аналогично предыдущему листингу в тех случаях, когда результат предполагал появление нулевой матрицы, использовалась её свёртка с помощью функции суммирования 96, 98, ..., 106, 112, 114, 116. Обратите внимание на ошибку в 106–109 при попытке обращения вырожденного произведения матриц $(D^{\mathsf{T}}D)^{-1}$, возникающую из-за линейно зависимых строк матрицы D 110–111.

```
> matrix(c(eigen(A)$values, eigen(B)$values, eigen(C)$values), 3)
                     [,1]          [,2]                [,3]
[1,] -7.53094+0.000000i 11.508598+0i  6.722784+3.699489i
[2,]  1.76547+6.890168i -6.521250+0i  6.722784-3.699489i
[3,]  1.76547-6.890168i  1.012652+0i -2.445567+0.000000i
> eigen(B)$vectors
            [,1]       [,2]       [,3]
[1,] -0.70253778 -0.2447187 0.88289227
[2,]  0.71025837  0.4995737 0.04098567
[3,]  0.04442655 -0.8309867 0.46778351
```

Для вычисления собственных значений матриц A, B, C в 118 использованы функции «eigen()$values», а для поиска собственных векторов в 123 — функция «eigen()$vectors». Как видно из результатов расчёта 119–122, матрицы A и C имеют комплексно-сопряжённые собственные значения, а вещественные линейно-независимые собственные векторы есть только у матрицы B 124–127.

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 анализа 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 $\overline{A}$, 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 события 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 $\{\omega_1, \omega_2, \ldots, \omega_n\}$ 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 $\mathbf{P}\{\omega_i\} = p_i$. Значения этой функции, выражающие меру возможности осуществления элементарного события ω_i, называются его *вероятностью*. При этом имеют место следующие *свойства вероятности*: $\mathbf{P}\{\omega_i\} \in (0, 1)$, $\mathbf{P}\{\emptyset\} = 0$, $\mathbf{P}\{\Omega\} = 1$.

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

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

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

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

2.2. Условная вероятность 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*:

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

Событие A полагают *не зависимым* от B, если $\mathbf{P}\{A|B\} = \mathbf{P}\{A\}$. 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

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

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. Одномерные случайные величины

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: $0 \leqslant F(x) \leqslant 1$;
2. Функция распределения — *неубывающая* функция: $F(x_2) \geqslant F(x_1)$, 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: $\mathbf{P}\{\alpha \leqslant X < \beta\} = F(\beta) - F(\alpha)$.

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 $\sum\limits_{i=1}^{n} p_i = 1$.

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*:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}.$$

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

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

$$\int\limits_{-\infty}^{\infty} f(x)\mathrm{d}x = 1\,;$$

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

$$\mathbf{P}\{\alpha \leqslant X \leqslant \beta\} = \int\limits_{\alpha}^{\beta} f(x)\mathrm{d}x\,;$$

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

$$F(x) = \int\limits_{-\infty}^{x} f(t)\mathrm{d}t\,.$$

2.4. Многомерные случайные величины

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

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

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

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

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

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

Если рассмотрению подлежит только часть компонент вектора $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)^{\mathsf{T}}$, где $k < n$, то используется *частная* (*маргинальная*) *функция распределения*:

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

а также *частная* (*маргинальная*) *плотность распределения*:

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

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

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

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

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

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

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

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

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

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

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

$$\mathbf{E}(X) = \int\limits_{-\infty}^{\infty} x\, f(x) \mathrm{d}x \,.$$

Основные *свойства* математического ожидания:

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

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

$$\mathbf{D}(X) = \mathbf{E}(X - \mathbf{E}(X))^2.$$

Можно показать, что верна *универсальная формула дисперсии*

$$\mathbf{D}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2.$$

Для нахождения *дисперсии* дискретной случайной величины используют формулу

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

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

$$\mathbf{D}(X) = \int\limits_{-\infty}^{\infty} (x - \mathbf{E}(X))^2 f(x) \mathrm{d}x = \int\limits_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x - \mathbf{E}(X)^2.$$

Основные *свойства* дисперсии:

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

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

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

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

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

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* $x_{\frac{3}{4}}$ *квартилями*. Легко увидеть, что второй квартиль 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

$$\operatorname{cov}(X,Y) = \sigma_{XY} = \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))) = \\ = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y).$$

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*: $|\rho_{XY}| \leqslant 1$;
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* $\boldsymbol{X}$ в рассмотрение вводятся многомерные аналоги числовых характеристик.

Для случайного вектора $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^{\mathsf{T}}$ характеристикой центра группирования будет *вектор средних значений*

$$\mathbf{E}(X) = (\mathbf{E}(X_1), \mathbf{E}(X_2), \ldots, \mathbf{E}(X_n))^{\mathsf{T}}.$$

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

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

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

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

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

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

2.6. Некоторые 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 $k = 0, 1, \ldots, n$ с вероятностями, определяемыми формулой Бернулли

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

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:

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

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

$$\mathbf{E}(X) = np, \quad \mathbf{D}(X) = npq.$$

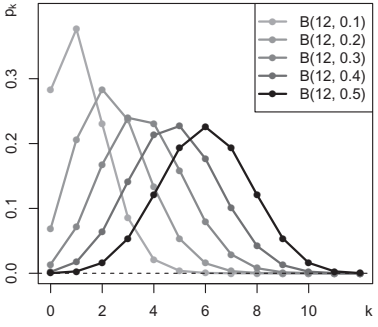


Рис. 2.1. Биномиальное распределение вероятностей p_k

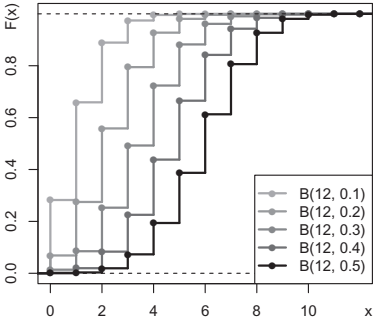


Рис. 2.2. Функция биномиального распределения $F(x)$

На рис. 2.1 и 2.2 показаны примеры построения графиков вероятностей p_k и функции распределения $F(x)$ биномиально распределённой случайной величины $X \sim \mathcal{B}(n,p)$ при $n = 12$ и $p = \frac{1}{10}, \frac{2}{10}, \dots, \frac{5}{10}$.

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

```
1 > source("plot2graph.r")
2 > k <- seq(0, n <- 12)
3 > p <- seq(5)/10
4 > pk <- sapply(p, function(i) dbinom(k, n, i))
5 > Fk <- sapply(p, function(i) pbinom(k, n, i))
6 > lt <- sapply(p, function(i) sprintf("B(%.0f, %.3g)", n, i))
7 > pmfPlot(k, pk, lt)
8 > cmfPlot(k, Fk, lt)
```

Команда «source("plot2graph.r")» в строке [1] производит загрузку из указанного файла исходного кода, содержащего функции для построения графиков распределений случайных величин.

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

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

Функции «pmfPlot()» и «cmfPlot()» определены в пользовательской библиотеке «plot2graph.r» и производят построение графиков вероятностей и функций распределения дискретной случайной величины по переданным векторам абсцисс «k» и ординат «pk» или «Fk». Полный текст функций из файла «plot2graph.r» приведён в Приложении Б.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 $k = 0, 1, \ldots, \infty$ 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 $n \to \infty$, а $p \to 0$ 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:

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

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

$$\mathbf{E}(X) = \mathbf{D}(X) = \lambda = np.$$

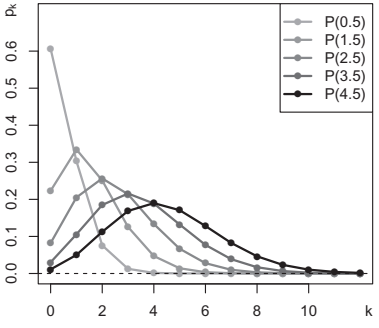


Рис. 2.3. Пуассоновское распределение вероятностей p_k

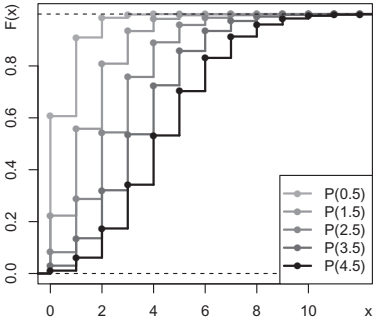


Рис. 2.4. Функция пуассоновского распределения $F(x)$

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 $\lambda = \frac{1}{2}, \frac{3}{2}, \dots, \frac{9}{2}$.

Пример 2.2. Продолжая предыдущий пример, построим графики вероятностей p_k и функции распределения $F(x)$ для пуассоновской случайной величины $X \sim \mathcal{P}(\lambda)$ при вышеуказанных значениях параметра λ.

```
> lambda <- seq(5, 45, 10)/10
> pk <- sapply(lambda, function(i) dpois(k, i))
> Fk <- sapply(lambda, function(i) ppois(k, i))
> lt <- sapply(lambda, function(i) sprintf("P(%.3g)", i))
> pmfPlot(k, pk, lt)
> cmfPlot(k, Fk, lt)
```

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 $k = 0, 1, \ldots, \infty$ 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:

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

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

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

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

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

На рис. 2.5 и 2.6 показаны примеры построения графиков вероятностей p_k и функции распределения $F(x)$ геометрически распределённой случайной величины $X \sim \mathcal{G}(p)$ при $p = \frac{1}{10}, \frac{2}{10}, \ldots, \frac{5}{10}$.

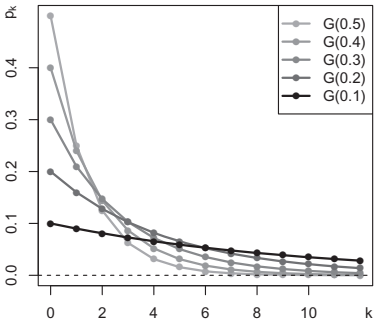


Рис. 2.5. Геометрическое распределение вероятностей p_k

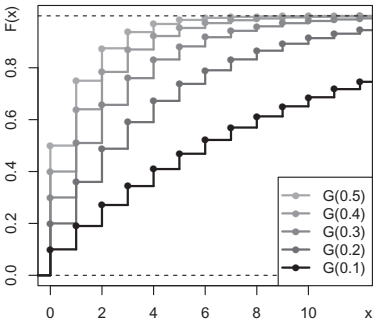


Рис. 2.6. Функция геометрического распределения $F(x)$

Пример 2.3. Продолжая предыдущий пример, построим графики вероятностей p_k и функции распределения $F(x)$ для геометрически распределённой случайной величины $X \sim \mathcal{G}(p)$ при вышеуказанных значениях параметра p.

```
> p <- seq(5, 1)/10
> pk <- sapply(p, function(i) dgeom(k, i))
> Fk <- sapply(p, function(i) pgeom(k, i))
> lt <- sapply(p, function(i) sprintf("G(%.3g)", i))
> pmfPlot(k, pk, lt)
> cmfPlot(k, Fk, lt)
```

2.6.4. Равномерное распределение

Простейшим из непрерывных распределений является равномерное распределение, возникающее при обобщении понятия n равновероятных случайных событий на случай $n \to \infty$. 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 величи-

ны, подчиняющейся равномерному закону $X \sim \mathcal{U}(a, b)$, имеет вид:

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

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

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

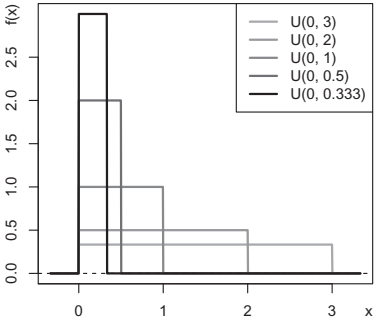


Рис. 2.7. Плотность равномерного распределения $f(x)$

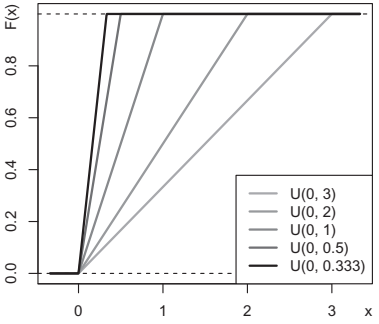


Рис. 2.8. Функция равномерного распределения $F(x)$

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: $a = 0$, $b = \frac{1}{3}, \frac{1}{2}, 1, 2, 3$.

Пример 2.4. Продолжая предыдущий пример, построим графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для равномерно распределённой случайной величины $X \sim \mathcal{U}(a, b)$ при вышеуказанных значениях параметров a и b.

```
> a <- 0; b <- c(seq(3,2), 1/seq(3))
> x <- seq(a-min(b), max(b)+min(b), len=900)
> fx <- sapply(b, function(t) dunif(x, a, t))
> Fx <- sapply(b, function(t) punif(x, a, t))
```

```
> lt <- sapply(b, function(t) sprintf("U(%.3g, %.3g)", a, t))
> pdfPlot(x, fx, lt)
> cdfPlot(x, Fx, lt)
```

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:

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

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:

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

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

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

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 $\lambda = \frac{1}{2}, 1, 2, 3, 4$.

Пример 2.5. Продолжая предыдущий пример, построим графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для показательно распределённой случайной величины $X \sim \mathcal{E}(\lambda)$ при вышеуказанных значениях параметра λ.

```
> lambda <- c(1/2, seq(4))
> x <- seq(-1/4, 4, len=900)
> fx <- sapply(lambda, function(t) dexp(x, t))
> Fx <- sapply(lambda, function(t) pexp(x, t))
```

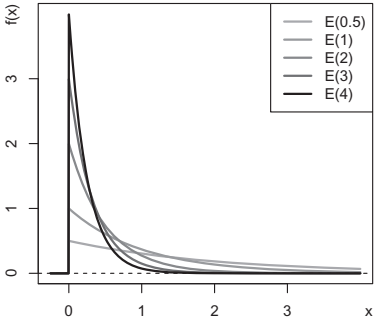


Рис. 2.9. Плотность показательного распределения $f(x)$

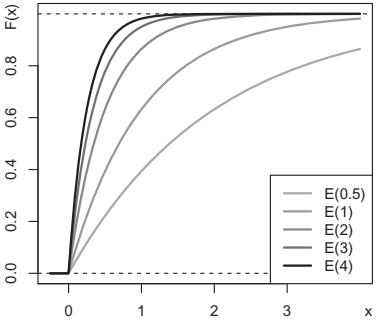


Рис. 2.10. Функция показательного распределения $F(x)$

```
> lt <- sapply(lambda, function(t) sprintf("E(%.3g)", t))
> pdfPlot(x, fx, lt)
> cdfPlot(x, Fx, lt)
```

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:

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

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

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

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

$$\mathbf{E}(X) = a, \quad \mathbf{D}(X) = \sigma^2.$$

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 $i = 1, 2, \ldots, n$ — независимые случайные величины, то $Y = \sum\limits_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum\limits_{i=1}^{n} a_i\,, \sqrt{\sum\limits_{i=1}^{n} \sigma_i^2}\right)$;
3. Если $X_i \sim \mathcal{N}(a_i, \sigma_i)$, при $i = 1, 2, \ldots, n$ — зависимые случайные величины, то $Y = \sum\limits_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum\limits_{i=1}^{n} a_i\,, \sqrt{\sum\limits_{i=1}^{n} \sigma_i^2 + \sum\limits_{i<j} \rho_{ij}\sigma_i\sigma_j}\right)$.

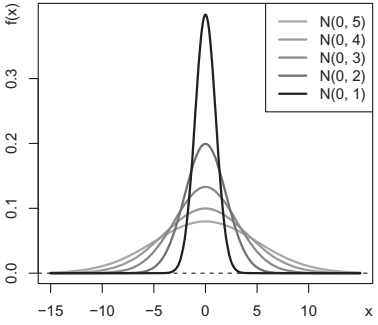


Рис. 2.11. Плотность нормального распределения $f(x)$

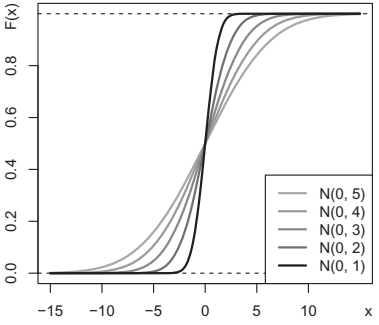


Рис. 2.12. Функция нормального распределения $F(x)$

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: $a = 0$, $\sigma = 1, 2, \ldots, 5$.

Пример 2.6. Продолжая предыдущий пример, построим графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для нормально распределённой случайной величины $X \sim \mathcal{N}(a, \sigma)$ при вышеуказанных значениях параметров a и σ.

```
> a <- 0; sigma <- seq(5, 1)
> x <- seq(a-3*max(sigma), a+3*max(sigma), len=300)
> fx <- sapply(sigma, function(t) dnorm(x, a, t))
> Fx <- sapply(sigma, function(t) pnorm(x, a, t))
> lt <- sapply(sigma, function(t) sprintf("N(%.3g, %.3g)", a, t))
> pdfPlot(x, fx, lt)
> cdfPlot(x, Fx, lt)
```

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:

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

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

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

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

где $\Phi(x)$ — функция Лапласа.

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

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

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: $a = 0$, $\sigma = \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1$.

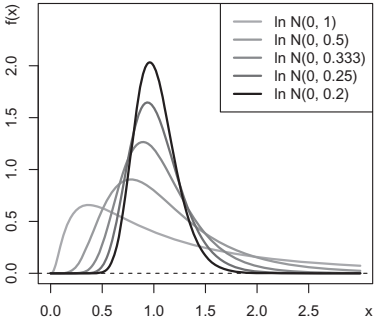


Рис. 2.13. Плотность логарифмически нормального распределения $f(x)$

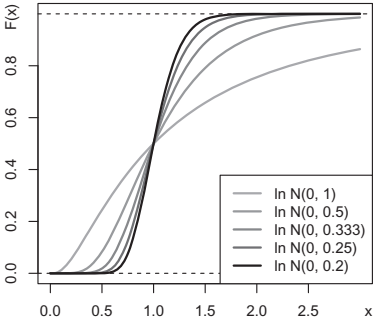


Рис. 2.14. Функция логарифмически нормального распределения $F(x)$

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

```
> a <- 0; sigma <- 1/seq(5)
> x <- seq(a, a+3*max(sigma), len=300)
> fx <- sapply(sigma, function(t) dlnorm(x, a, t))
> Fx <- sapply(sigma, function(t) plnorm(x, a, t))
> lt <- sapply(sigma, function(t) sprintf("ln N(%.3g, %.3g)", a, t))
> pdfPlot(x, fx, lt)
> cdfPlot(x, Fx, lt)
```

2.6.8. Пирсона χ^2-распределение

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

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

Плотность распределения χ^2 выражается формулой:

$$f(x) = \begin{cases} 0\,, & \text{если } x \leqslant 0; \\ \frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n-2}{2}} e^{-\frac{x}{2}}, & \text{если } 0 < x; \end{cases} \quad \Gamma(p) = \int\limits_0^\infty t^{p-1} e^{-t}\,\mathrm{d}t,$$

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

Математическое ожидание и дисперсия распределения χ^2_n имеют вид:

$$\mathbf{E}(\chi^2_n) = n, \quad \mathbf{D}(\chi^2_n) = 2n.$$

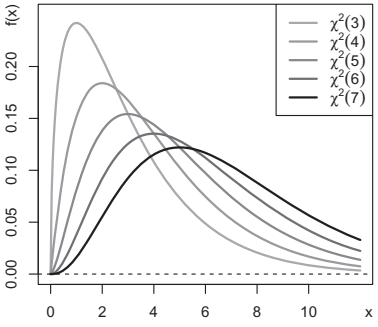


Рис. 2.15. Плотность χ^2-распределения $f(x)$

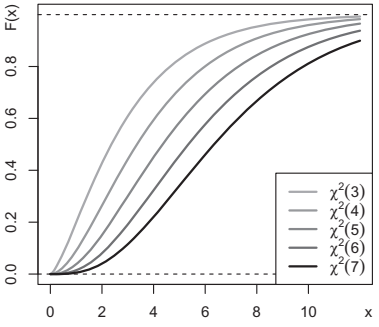


Рис. 2.16. Функция χ^2-распределения $F(x)$

На рис. 2.15 и 2.16 показаны примеры построения графиков плотности вероятности $f(x)$ и функции распределения $F(x)$ случайной величины $X \sim \chi^2_n$ при числе степеней свободы: $n = 3, 4, \ldots, 7$.

Пример 2.8. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ для случайной величины $X \sim \chi^2_n$ при вышеуказанных значениях параметра n.

```
> n <- seq(3, 7)
> x <- seq(0, 12, len=300)
> fx <- sapply(n, function(i) dchisq(x, i))
> Fx <- sapply(n, function(i) pchisq(x, i))
> lt <- sapply(n, function(i) parse(text=sprintf("chi^2*(%.1g)", i)))
> pdfPlot(x, fx, lt)
> cdfPlot(x, Fx, lt)
```

2.6.9. Стьюдента t-распределение

Если случайные величины $Z \sim \mathcal{N}(0,1)$ и $U \sim \chi^2_n$ — 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:

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

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

Математическое ожидание и дисперсия t-распределения выражаются формулами:

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

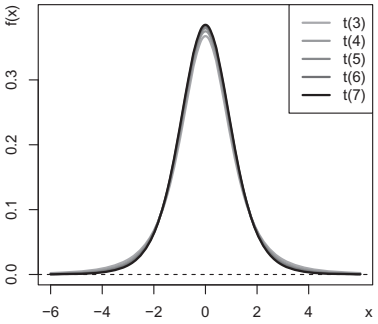


Рис. 2.17. Плотность t-распределения $f(x)$

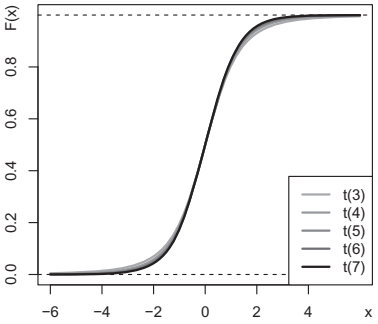


Рис. 2.18. Функция t-распределения $F(x)$

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 $n = 3, 4, \ldots, 7$.

Пример 2.9. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ случайной величины $X \sim t_n$ при вышеуказанных значениях параметра n.

```
> n <- seq(3, 7)
> x <- seq(-6, 6, len=300)
> fx <- sapply(n, function(i) dt(x, i))
> Fx <- sapply(n, function(i) pt(x, i))
> lt <- sapply(n, function(i) sprintf("t(%.0f)", i))
> pdfPlot(x, fx, lt)
> cdfPlot(x, Fx, lt)
```

2.6.10. Фишера F-распределение

Если случайные величины $U \sim \chi^2_m$ и $V \sim \chi^2_n$ — 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:

$$f(x) = \begin{cases} 0\,, & \text{если } x \leqslant 0\,; \\ \dfrac{\sqrt{\frac{(mx)^m n^n}{(mx+n)^{m+n}}}}{xB\left(\frac{m}{2}, \frac{n}{2}\right)}, & \text{если } x > 0\,; \end{cases} \qquad B(u,v) = \int\limits_0^1 t^{u-1}(1-t)^{v-1}\mathrm{d}t,$$

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

Математическое ожидание и дисперсия F-распределения выражается формулами:

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

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 одинаковых значениях степеней свободы $m = n = 3, 4, \ldots, 7$.

[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 $F_{m,n}$.

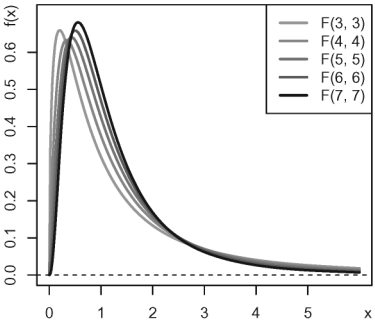


Рис. 2.19. Плотность F-распределения $f(x)$

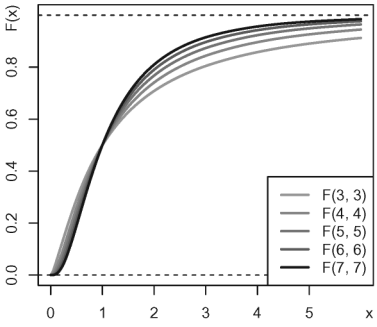


Рис. 2.20. Функция F-распределения $F(x)$

Пример 2.10. Продолжая предыдущий пример, построим вышеприведённые графики плотности вероятностей $f(x)$ и функции распределения $F(x)$ случайной величины $X \sim F_{\frac{m}{n}}$ при вышеуказанных значениях параметров m и n.

```
> m <- n <- seq(3, 7); k <- seq_along(m)
> x <- seq(0, 6, len=300)
> fx <- sapply(k, function(i) df(x, m[i], n[i]))
> Fx <- sapply(k, function(i) pf(x, m[i], n[i]))
> lt <- sapply(k, function(i) sprintf("F(%.0f, %.0f)", m[i], n[i]))
> pdfPlot(x, fx, lt)
> cdfPlot(x, Fx, lt)
```

Контрольные вопросы

1. Что называется случайным событием и пространством элементарных исходов? Дайте определения достоверного и невозможного событий.
2. Дайте определения суммы и произведения событий.
3. Сформулируйте теоремы сложения вероятностей для совместных и несовместных событий, а также умножения вероятностей для зависимых и независимых событий.
4. Сформулируйте теорему о полной вероятности и запишите формулу Байеса.

5. Дайте определение случайной величины и закона её распределения. Перечислите типы случайных величин.
6. Дайте определение математического ожидания случайной величины и перечислите его свойства.
7. Дайте определение и перечислите свойства дисперсии и среднего квадратического отклонения случайной величины.
8. Дайте определение плотности вероятности случайной величины и перечислите его свойства.
9. Как определяется двумерная случайная величина и закон её распределения.
10. Приведите определения и перечислите свойства условного математического ожидания и дисперсии случайной величины.
11. Дайте определение ковариации и коэффициента корреляции системы случайных величин и сформулируйте их свойства.
12. Сформулируйте законы распределения дискретных случайных величин: биномиальный, геометрический и распределения Пуассона? Как найти числовые характеристики этих распределений?
13. Сформулируйте законы распределения непрерывных случайных величин: равномерный, показательный? Как найти числовые характеристики этих распределений?
14. Какие случайные величины называют нормально и логнормально распределёнными? Как найти числовые характеристики этих распределений?
15. 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

Методы оценивания и проверки гипотез

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 приблизительной, 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 $\{x_1, x_2, \ldots, x_n\}$ — *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 $\{x_i\}$, $i = 1, 2, \ldots, n$. 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 $g(x_1, x_2, \ldots, x_n)$. Очевидно, что если рассматривать элементы выборки как независимые одинаково распределённые случайные

величины $x_i \sim X \sim F(x)$, то и статистика будет случайной величиной, имеющей свой закон распределения $g(x_1, x_2, \ldots, x_n) \sim G(x)$.

Упорядоченные по неубыванию элементы выборки называются *вариационным рядом* $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$:

$$\min(x_i) = x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)} = \max(x_i).$$

3.2. Точечные оценки параметров распределения

Любые характеристики случайной величины X, полученные по её выборке $\{x_1, x_2, \ldots, x_n\}$, называются *выборочными* или *эмпирическими*. *Статической 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 $\{x_1,\ x_2,\ \ldots,\ x_n\}$ — случайная выборка, тогда $\theta_n(x_1, x_2, \ldots, x_n)$ — выборочная оценка некоторого параметра θ. Оценка θ_n называется *несмещённой*, если для любого фиксированного n верно, что $\mathbf{E}(\theta_n) = \theta$. 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

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

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 ожида-

ния $\mathbf{E}(X)$ и дисперсии $\mathbf{D}(X)$ являются *выборочное среднее* $\overline{x}$ и *исправленная выборочная дисперсия* s_x^2:

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

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

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

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

где $\overline{x}$ и $\overline{y}$ — 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

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

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* $f_{n,h}(x)$:

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

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

где точки вокруг обозначения вероятности $\cdot\mathbf{P}\cdot$ указывают на поточечную сходимость по вероятности гистограммы $f_{n,h}(x)$ к функции

плотности вероятности при выполнении условий $nh \to \infty, h \to 0$; $h = \text{const}$ — 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:

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

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:

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

Обобщая предыдущую формулу, найдём *выборочный квантиль порядка* p:

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

где $[a]$ — целая часть числа a. При анализе распределений с большими выбросами для характеристики центра распределения вместо

выборочного среднего $\overline{x}$ 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 квартилями: $x_{\frac{3}{4},n} - x_{\frac{1}{4},n}$.

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**.

```
> source("samples.r")
> n <- 100; x <- samples(n, seed=20100625); x
  [1] 10.50  6.20  5.55  7.30 10.57  8.50  5.53  5.87  9.61 10.80
 [11]  6.70  9.42 10.53  8.05  7.44 10.47  5.53  8.72  7.06  7.12
 [21]  6.82  9.77  8.74 10.18  6.10  9.53  6.44  7.89  9.23  6.96
 [31]  8.05  6.59  9.61  8.44  6.14  8.12  8.04  6.86  7.89  7.27
 [41]  9.14  8.37  9.26  9.06  7.04  9.32  7.16  7.24  5.55  7.53
 [51]  9.60  7.73  8.82  9.98 10.68  8.58  6.56  8.38  7.93  9.11
 [61]  5.26  7.70  8.20  4.24  9.79 10.93  9.19 14.35 13.44  9.91
 [71]  8.42 10.00  8.73  8.17  9.03  9.17  9.31 12.74  7.38  4.82
 [81]  9.76  8.72  8.27  8.78  7.24 11.23  8.83 10.55  6.63  8.48
 [91]  6.12  6.21  8.01  6.75  9.11  6.99  9.95  7.09  4.46  8.41
```

Команда «`source("samples.r")`» в первой строке листинга загружает вспомогательную функцию «`samples()`», вызываемую во второй строке и обеспечивающую генерирование выборочных значений случайной величины X. Полный текст функции «`samples()`» приведён в приложении Б.2. Переменная «`n<-100`» задаёт объём выборки, а параметр «`seed=20100625`» устанавливает начальное состояние генератора псевдослучайных чисел, формирующего эту выборку. Если вы пожелаете получить другую последовательность выборочных значений, то укажите другое значение этого параметра [1]

```
> a1 <- mean(x); s1 <- sd(x); c(a1, s1)
[1] 8.295500 1.815452
> quantile(x)
    0%    25%    50%    75%   100%
 4.240  7.055  8.375  9.345 14.350
```

[1] При организации индивидуальной работы студентов значение параметра «`seed`» может быть указано преподавателем.

В строках 13–17 вычисляются основные выборочные характеристики: среднее значение $\overline{x} \approx 8.30$, исправленное среднее квадратическое отклонение $s_x \approx 1.82$, а также квартили: $x_{0,n} \approx 4.24$, $x_{\frac{1}{4},n} \approx 7.06$, $x_{\frac{1}{2},n} \approx 8.38$, $x_{\frac{3}{4},n} \approx 9.35$, $x_{1,n} \approx 14.35$. Заметим, что квартили уровней 0 и 1 соответствуют минимальному и максимальному элементам выборки: $x_{0,n} = \min(x_i)$, $x_{1,n} = \max(x_i)$.

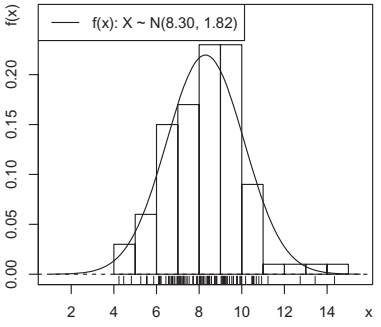


Рис. 3.1. Гистограмма $f_{n,h}(x)$ выборки значений и график плотности вероятности $f(x)$ случайной величины $X \sim \mathcal{N}(8.30, 1.82)$

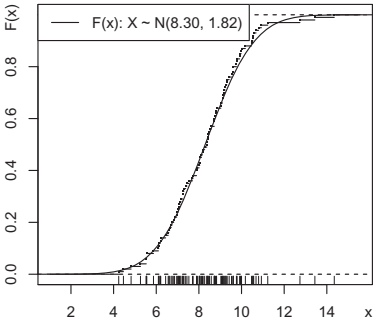


Рис. 3.2. Эмпирическая функция распределения $F_n(x)$ выборки и график функции распределения $F(x)$ с.в. $X \sim \mathcal{N}(8.30, 1.82)$

```
> x1 <- seq(a1-4*s1, a1+4*s1, length=n); range(x1)
[1]  1.033693 15.557307
> f1 <- dnorm(x1, a1, s1); F1 <- pnorm(x1, a1, s1)
> ltext <- sprintf("X ~ N(%.2f, %.2f)",a1, s1); ltext
[1] "X ~ N(8.30, 1.82)"
> hist(x, breaks=12, freq=FALSE, xlab="x", ylab="f(x)")
> rug(x); lines(x1, f1); abline(h=0, lty=2)
> legend("topleft", lty=1, legend=paste("f(x):", ltext))
> windows()
> plot(ecdf(x), pch=".", xlab="x", ylab="F(x)")
> rug(x); lines(x1, F1); abline(h=c(0,1), lty=2)
> legend("topleft", lty=1, legend=paste("F(x):", ltext))
```

В строках 18–29 выполняются построения гистограммы $f_{n,h}(x)$ и эмпирической функции распределения $F_n(x)$ для выборки 100 значений случайной величины X, показанные на 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: $a \approx 8.30$, $\sigma \approx 1.82$.

В строке 18 с помощью оценок параметров a и σ формируется вектор абсцисс «x1». Наименьшее и наибольшее значения вектора «x1» отображаются с помощью функции «range()». В строке 20 с помощью функций «dnorm()» и «pnorm()» формируются соответствующие абсциссам «x2» векторы ординат теоретических функций плотности вероятности «f1» и функции распределения «F1» случайной величины $X \sim \mathcal{N}(8.30, 1.82)$.

В строке 21 с помощью функции «sprintf()» и оценок параметров «a1» и «s1» формируется строка «ltext», содержащая описание предполагаемого закона распределения случайной величины X, показанное в 22.

В строке 23 с помощью функции «hist()» выполняется построение гистограммы для вектора выборочных значений «x». Параметр «breaks=12» позволяет указать желательное число интервалов группировки, отталкиваясь от которого алгоритм выбирает оптимальное разбиение выборочной совокупности; так, в приведённом примере, оптимальное число интервалов оказалось равным $k = 11$. Параметр «freq=FALSE» указывает, что при построении гистограммы масштаб по оси ординат должен соответствовать не абсолютной частоте выборочных значений n_k, а плотности их относительной частоты $\frac{n_k}{nh_k}$, где $h_k = h = \text{const}$, что обеспечивает нормировку гистограммы по площади: $\frac{1}{nh}\sum_{k=1}^{11} n_k = 1$. Параметры «xlab» и «ylab» позволяют установить подписи, отображаемые для осей абсцисс и ординат графика.

Функция «windows()» в строке 26 открывает новое графическое окно, а функция «plot(ecdf(x)...)» в строке 27 строит в новом 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()».

Функции «rug(x)» в строках 24 и 28 отображают над осью абсцисс метки, соответствующие координатам выборочных значений «x». Функции «lines()» строят кривые, соответствующие теоретическим функциям плотности вероятности $f(x)$ и распределения $F(x)$, а функции «abline(h=..., lty=2)» — горизонтальные штриховые линии, пересекающие ось ординат в указанных точках.

Функции «legend("topleft", ...)» в строках 25 и 29 отображают в левом верхнем углу графика надпись, определяемую параметром «legend=paste(..., ltext)», где функция «paste(..., ltext)» объединяет строки «"f(x):"» и «"F(x):"» с переменной «ltext», которая была показана ранее в 22. Параметр «lty=1» указывает, что для отображения теоретических функций распределения $f(x)$ и $F(x)$ на рис. 3.1 и 3.2 были использованы сплошные линии.

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 $\theta_n^\pm$ — функции случайной выборки, такие, что верно равенство

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

Тогда интервал $\mathbf{I}_\gamma(\theta) = (\theta_n^-, \theta_n^+)$ 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

$$\mathbf{P}\left\{\theta \in (\theta_n^-, \theta_n^+)\right\} \geqslant \gamma.$$

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 $\mathbf{E}(X)$ 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

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

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

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

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

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

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

$$\mathbf{I}_\gamma(a) = \left(\overline{x} - \frac{s_x}{\sqrt{n}} \cdot t_{\frac{1+\gamma}{2},n-1},\ \ \overline{x} - \frac{s_x}{\sqrt{n}} \cdot t_{\frac{1-\gamma}{2},n-1} \right),$$

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 $\mathbf{D}(X)$ 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 $n-1$

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

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

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

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

$$\mathbf{I}_\gamma\left(\sigma^2\right) = \left(s_x^2 \cdot \frac{(n-1)}{\chi^2_{\frac{1+\gamma}{2},n-1}},\ \ s_x^2 \cdot \frac{(n-1)}{\chi^2_{\frac{1-\gamma}{2},n-1}} \right),$$

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

Пример 3.2. Построить с помощью **R** реализации доверительных интервалов для математического ожидания $\mathbf{E}(X)$ и дисперсии $\mathbf{D}(X)$ стандартной нормально распределённой случайной величины $X \sim \mathcal{N}(0,1)$ при различных значениях доверительной вероятности и объёма выборок на интервалах: $\gamma \in [0{,}95; 0{,}999]$ и $n \in [100, 1000]$.

```
> set.seed(20100625)
> n <- seq(100, 1000, 20)
> g <- seq(0.95, 0.995,, length(n))
> ciE <- function(x,n,g) mean(x)-sd(x)/sqrt(n)*qt((1+c(g,0,-g))/2,n-1)
> ciD <- function(x,n,g) sd(x)^2*(n-1)/qchisq((1+c(g,0,-g))/2,n-1)
```

```
> ciEn <- sapply(n, function(nn) ciE(rnorm(nn), nn, g[1]))
> ciDn <- sapply(n, function(nn) ciD(rnorm(nn), nn, g[1]))
> ciEg <- sapply(g, function(gg) ciE(rnorm(n[1]), n[1], gg))
> ciDg <- sapply(g, function(gg) ciD(rnorm(n[1]), n[1], gg))
> textEn <- parse(text=sprintf("EX*(list(n,gamma==%.3g))",g[1]))
> textDn <- parse(text=sprintf("DX*(list(n,gamma==%.3g))",g[1]))
> textEg <- parse(text=sprintf("EX*(list(gamma,n==%.0f))",n[1]))
> textDg <- parse(text=sprintf("DX*(list(gamma,n==%.0f))",n[1]))
> ci_graph <- function(x, y, point, text, xlb, ylb) { windows()
+   plot(range(x), range(y), type="n", xlab=xlb, ylab=ylb)
+   for(j in seq(length(x))) {
+     lines(rep(x[j],3), y[,j])
+     points(x[j], y[2,j], pch=20) }
+   legend("topright", lty=1, pch=20, legend=text, bg="white")
+   abline(h=point, lty=2) }
> ci_graph(n, ciEn, 0, textEn, "n", "EX")
> ci_graph(n, ciDn, 1, textDn, "n", "DX")
> ci_graph(g, ciEg, 0, textEg, expression(gamma), "EX")
> ci_graph(g, ciDg, 1, textDg, expression(gamma), "DX")
```

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

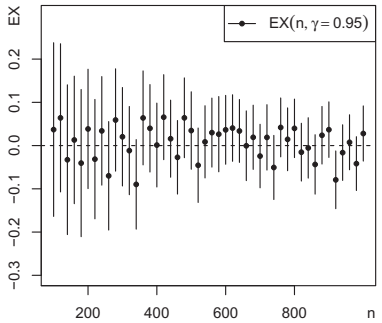
В строках 4–5 определяются функции «ciE» и «ciD», которые с заданной вероятностью «g» вычисляют границы доверительных интервалов для $\mathbf{E}(X)$ и $\mathbf{D}(X)$ по выборке «x» заданного объёма «n».

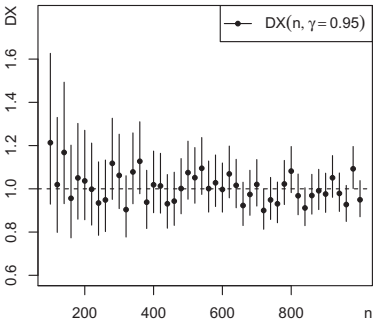
Далее, в 6–7 с помощью функции «sapply()» для каждого значения объёма выборки $n = 100, 120, \ldots, 1000$ с фиксированной вероятностью $\gamma_1 = 0{,}95$ строятся реализации доверительных интервалов для $\mathbf{E}(X)$ и $\mathbf{D}(X)$ случайной величины $X \sim \mathcal{N}(0, 1)$, показанные на рис. 3.3 и 3.4 в верхнем ряду.

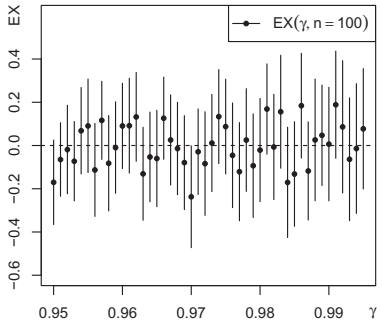
Аналогично, в 8–9 с помощью функции «sapply()» для каждого значения доверительной вероятности $\gamma = 0{,}95;\ 0{,}951;\ \ldots;\ 0{,}995$ при фиксированном объёме выборки $n_1 = 100$ строятся реализации доверительных интервалов для $\mathbf{E}(X)$ и $\mathbf{D}(X)$ случайной величины $X \sim \mathcal{N}(0, 1)$, 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.







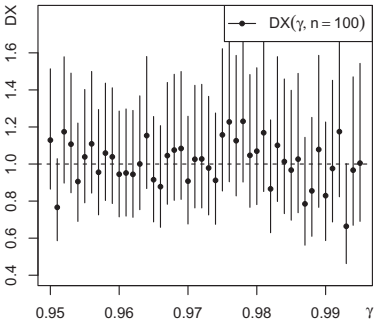


Рис. 3.3. Реализации доверительных интервалов $\mathbf{E}(X)$ для $n = 100$, 120, ... 1000 и $\gamma = 0{,}95$; 0,951; ...; 0,995 при $X \sim \mathcal{N}(0, 1)$

Рис. 3.4. Реализации доверительных интервалов $\mathbf{D}(X)$ для $n = 100$, 120, ... 1000 и $\gamma = 0{,}95$; 0,951; ...; 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 «`rep(x[j],3)`» и ординат «`y[,j]`» рисует пару горизонтальных линий с точкой между ними. Абсцисса точки соответствует j-ой точечной оценке, а отрезки горизонтальных линий — j-му доверительному интервалу для $\mathbf{E}(X)$ или $\mathbf{D}(X)$.

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 — горизонтальную штриховую линию «`lty=2`», пересекающую ось ординат в точке «`h=point`», соответствующей истинным значениям оцениваe-

мых величин $\mathbf{E}(X) = 0$ и $\mathbf{D}(X) = 1$.

В строках 21–24 описанная функция используется для непосредственного построения реализаций доверительных интервалов $\mathbf{E}(X)$ и $\mathbf{D}(X)$ при изменении $n = 100, 120, \ldots 1000$ и фиксированном значении $\gamma = 0{,}95$, а также при фиксированном значении $n = 100$ и изменении $\gamma = 0{,}95; 0{,}951; \ldots; 0{,}995$, где $X \sim \mathcal{N}(0, 1)$.

Из графиков, показанных на рис. 3.3 и 3.4, хорошо видно, что доверительные интервалы и для $\mathbf{E}(X)$ и для $\mathbf{D}(X)$ представляют собой коллинеарные оси ординат случайные векторы, длина которых уменьшается по мере увеличения объёма выборки n и уменьшения доверительной вероятности γ, а основания расположены в окрестности истинных значений оцениваемых величин $\mathbf{E}(X) = 0$ и $\mathbf{D}(X) = 1$.

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 $\{X_1, X_2, \ldots, X_n\}$.

Для проверки нулевой гипотезы формируется *статистический критерий* — специальная статистика $K(X_1, X_2, \ldots, X_n)$, 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 $K_n = K(X_1, X_2, \ldots, X_n)$. 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: $\mathbf{P}\{H_1|H_0\} = \alpha$. Помимо этого, существует вероятность совершить *ошибку второго рода*, состоящую в напрасном принятии неверной нулевой гипотезы $\mathbf{P}\{H_0|H_1\} = \beta$. Дополнительную к β величину, соответствующую вероятности недопущения ошибки второго рода $\mathbf{P}\{H_1|H_1\} = 1 - \beta$, 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 $H_0 : F(x) = F_0(x, \boldsymbol{\theta})$ против *альтернативной* $H_1 : F(x) \neq F_0(x, \boldsymbol{\theta})$, где $F_0(x, \boldsymbol{\theta})$ — теоретическая функция распределения случайной величины X; $\boldsymbol{\theta} \in \mathbf{R}^m$ — 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 $\{x_1, x_2, \ldots, x_n\}$ разбивают на k непересекающихся интервалов: $x_j \in (a_0, a_k] = (a_0, a_1] \cup (a_1, a_2] \cup \ldots \cup (a_{k-1}, a_k]$ и вычисляют статистику, имеющую распределение χ^2 с числом степеней свободы $k - m - 1$:

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

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 $(a_{i-1}, a_i]$: $p_i = F_0(a_i, \boldsymbol{\theta}_n) - F_0(a_{i-1}, \boldsymbol{\theta}_n)$, где $\boldsymbol{\theta}_n \in \mathbf{R}^m$ — выборочная оценка m-мерного вектора неизвестных параметров распределения $\boldsymbol{\theta}$.

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

при $X_d^2 \leqslant \chi^2_{\alpha,k-m-1}$ говорят, что нулевая гипотеза $H_0 : F(x) = F_0(x, \boldsymbol{\theta})$ 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 $n_i \geqslant 5$, с соседними до достижения частот приемлемого уровня или использовать *равновероятное группирование*, при котором $n_i \approx \frac{n}{k}$, где $i = 1, 2, \ldots, k$.

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 $\boldsymbol{\theta}$ оценивались по группированным данным. Если же оценки параметров $\boldsymbol{\theta}$ вычислялись по негруппированной реализации выборки, то действительное распределение наблюдаемой статистики будет заключено между χ^2_{k-m-1} и χ^2_{k-1} и при определённых допущениях будет лучше аппроксимироваться распределением χ^2_{k-1}.

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

```
1  > source("samples.r")
2  > x <- samples(seed=20100625)
3  > a <- mean(x); s <- sd(x); c(a, s)
4  [1] 8.295500 1.815452
5  > range(x)
6  [1]  4.24 14.35
7  > b1 <- c(4:15); m1 <- table(cut(x, breaks=b1)); m1
8    (4,5]   (5,6]   (6,7]   (7,8]   (8,9]  (9,10] (10,11] (11,12]
9        3       6      15      17      23      23       9       1
10 (12,13] (13,14] (14,15]
11       1       1       1
```

Назначение команд в строках 1–4 целиком аналогичны ранее указанным в примере 3.1. В строке 5–6 вычисляются наибольшее и наименьшее значения по реализации: $x_{\min} \approx 4{,}24$, $x_{\max} \approx 14{,}35$.

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: $(\hat{x}_{\min}, \hat{x}_{\max}] = (4, 15] = (4, 5] \cup (5, 6] \cup \ldots \cup (14, 15]$.

В строке 7 формируется вектор граничных точек «b1» и с помощью суперпозиции функций «table(cut())» осуществляется группировка выборочных значений по указанным интервалам 8–11.

```
12 > b2 <- c(4, 6:11, 15); m2 <- table(cut(x, breaks=b2)); m2
13   (4,6]   (6,7]   (7,8]   (8,9]  (9,10] (10,11] (11,15]
14       9      15      17      23      23       9       4
15 > b3 <- c(-Inf, 6:11, Inf); p3 <- diff(pnorm(b3, a, s))
16 > round(p3, 5); sum(p3)
17 [1] 0.10304 0.13470 0.19761 0.21566 0.17509 0.10574 0.06815
18 [1] 1
19 > chisq.test(x=m2, p=p3)
20 Chi-squared test for given probabilities
21 data: m2
22 X-squared = 3.939, df = 6, p-value = 0.6849
23 > x1 <- seq(4, 15, length=300); f1 <- dnorm(x1, a, s)
24 > hist(x, breaks=b2); rug(x); lines(x1, f1)
25 > windows(); qqnorm(x, pch=3); qqline(x, lty=2)
```

Из приведённых в строках 8–11 данных видно, что первый интервал группировки и четыре последних содержат слишком мало значений: $n_1 = 3$, $n_8 = n_9 = n_{10} = n_{11} = 1$. Тогда, для соответствия условию $n_i \geqslant 5$, 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 12–14.

Для подсчёта вектора теоретических частот «p3» в строке 15 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 $\Phi(\frac{x-\overline{x}}{s_x}) + \frac{1}{2}$, которые показаны в строке 17.

В строке 19 с помощью критерия χ^2 выполняется проверка гипотезы о соответствии эмпирических частот — теоретическим для нормального закона распределения $\mathcal{N}(\overline{x}, s_x)$. Из данных в строке 22 видно, что достигаемый для исследуемой реализации выборки «x» уровень значимости «p-value = 0.6849» значительно превосходит заданное значение $\alpha = 0{,}05$, что позволяет сделать вывод о согласии исследуемых данных с нулевой гипотезой: $H_0 : F(x) = \Phi(\frac{x-\overline{x}}{s_x}) + \frac{1}{2}$.

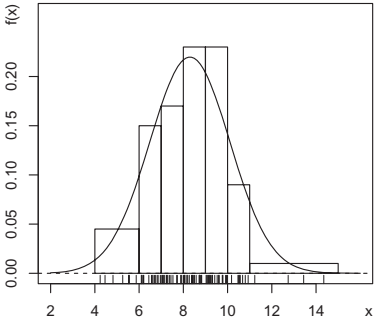


Рис. 3.5. Гистограмма $f_{n,h}(x)$ выборки значений и график плотности вероятности $f(x) = \frac{1}{1{,}82}\varphi(\frac{x-8{,}30}{1{,}82})$

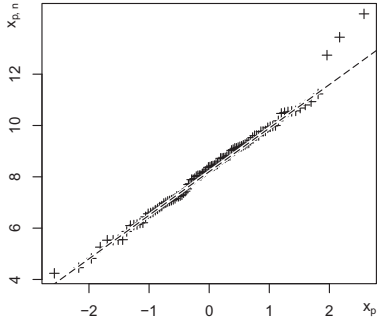


Рис. 3.6. Q–Q график для выборки значений и график функции распределения $F(x) = \Phi(\frac{x-8{,}30}{1{,}82}) + \frac{1}{2}$

В качестве иллюстрации в строках 23–24 выполняются построения гистограммы для вышеуказанной группировки выборочных данных и кривой плотности вероятности $f(x) = \frac{1}{1{,}82}\varphi(\frac{x-8{,}30}{1{,}82})$, 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 25 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 $X \sim \mathcal{N}(8{,}30,\ 1{,}82)$. Функция «`qqline()`» добавляет к выборочным данным график функции нормального распределения $F(x) = \Phi(\frac{x-8{,}30}{1{,}82}) + \frac{1}{2}$, выглядящий в указанной системе координат как прямая линия. При этом использованы два параметра: «`pch=3`» — отображение выборочных точек символами «+»; «`lty=2`» — отображение Q–Q графика $F(x)$ штриховой линией.

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

3.4.2. Критерии Колмогорова–Смирнова

Критерий согласия Колмогорова используется для проверки 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 $\boldsymbol{\theta}$.

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

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

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

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

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

Если объём выборки n достаточно велик, то α-квантиль распределения Колмогорова K_α можно приблизительно оценить 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 $H_1 : F(x) \neq \neq F_0(x)$. *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 $v_i = x_i - \overline{x}$, с помощью критерия согласия Колмогорова проанализировать зависимость достигаемого уровня значимости α_p от числа степеней свободы $m = 2, 3, \ldots, 20$ 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, dgts=NA)
3 > m <- 2:20; w <- sort(v <- x - mean(x))
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=".", xlab="v", ylab="F(v)")
8 > rug(v); lines(w, pt(w, m[which.max(alpha)]))
```

Назначение команд в строках 1–2 в целом аналогичны ранее указанным в примере 3.1. Отличие состоит в указании дополнительного параметра «`dgts=NA`» для функции «`samples()`», который отключает округление выборочных значений для генерируемой совокупности, поскольку использование критерия согласия Комогорова даёт корректные результаты лишь для непрерывных случайных величин. 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.

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 $F_0(v)$

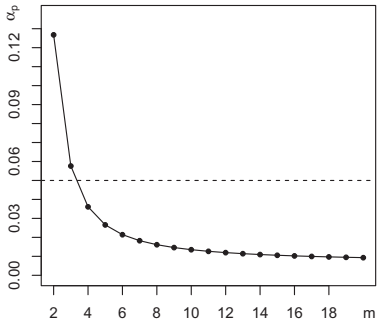


Рис. 3.7. Изменение уровня значимости α_p от числа степеней свободы m, достигаемого для нулевой гипотезы $H_0 : F(v) = F_0(v)$ при альтернативной $H_1 : F(v) \neq F_0(v)$

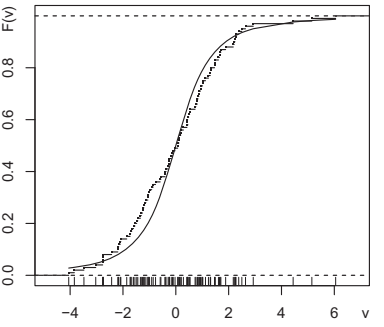


Рис. 3.8. Графики эмпирической $F_m(v)$ и наиболее близкой к ней теоретической функции распределения Стьюдента $F_0(v)$ с числом степеней свободы $m = 2$

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 объёмами 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 $F_m(x)$

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

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

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

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 чётных и нечётных элементов выборки, использованной в примере 3.4: $\{x_i\}_n = \{u_j\}_m \cup \{v_k\}_l$, 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 $H_1 : F_1(u) \neq F_2(v)$.

```
> u <- x[seq(1,99,2)]; v <- x[seq(2,100,2)]; ks.test(u,v)
Two-sample Kolmogorov-Smirnov test
data:  u and v
D = 0.14, p-value = 0.7166
alternative hypothesis: two-sided
> plot(ecdf(u), pch=25, xlim=range(x), xlab="u,v", ylab="F(u),F(v)")
> plot(ecdf(v), pch=24, add=TRUE); rug(u, side=1); rug(v, side=3)
```

В строке 9 вычисляются частичные выборочные векторы, содержащие чётные и нечётные элементы выборки $\{x_i\}_n = \{u_j\}_m \cup \{v_k\}_l$, 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 12 приведён достигаемый уровень значимости $\alpha_p \approx 0{,}72$, сравнение которого с заданным уровнем $\alpha = 0{,}05$ позволяет сделать вывод о хорошем согласовании нулевой гипотезы H_0 с выборочными данными $\{u_j\}_m$ и $\{u_k\}_l$.

В строках 14–15 с помощью композиции «plot(ecdf())» выполняется построение эмпирических функций распределения $F_m(u)$ и $F_l(v)$, приведённых на рис. 3.9, а с помощью команды «rug()» — отображение соответствующих этим функциям частей выборки вблизи нижней и верхней границ графика.

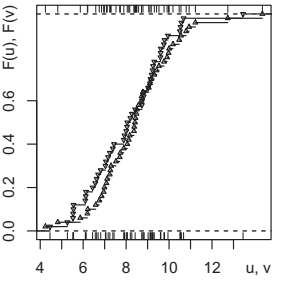


Рис. 3.9. Графики эмпирических функций распределения $F_m(u)$ и $F_l(v)$ для выборок $\{u_j\}_m$ и $\{v_k\}_l$

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 $H_0 : \mathbf{E}(X) = \mathbf{E}(Y)$ или гипотезы о равенстве дисперсий $H_0 : \mathbf{D}(X) = \mathbf{D}(Y)$.

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.

Одновыборочный t-критерий

Гипотезы: Проверяется *нулевая* гипотеза $H_0 : \mathbf{E}(X) = a$ против *альтернативных* H_1: а) $\mathbf{E}(X) \neq a$, б) $\mathbf{E}(X) < a$, в) $\mathbf{E}(X) > a$, где $\mathbf{E}(X)$ — математическое ожидание случайной величины X; a — заданное постоянное значение.

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

ней свободы $n-1$

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

где $\overline{x}$ — выборочное среднее значение; 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 $H_0 : \mathbf{E}(X) = a$ на уровне значимости α *отвергается* в пользу альтернативных H_1: а) $\mathbf{E}(X) \neq a$; б) $\mathbf{E}(X) < a$; в) $\mathbf{E}(X) > a$.
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 $a \in [\overline{x}-\frac{s_x}{2}, \overline{x}+\frac{s_x}{2}]$ для нулевой гипотезы $H_0 : \mathbf{E}(X) = a$ при альтернативных H_1: а) $\mathbf{E}(X) \neq a$, б) $\mathbf{E}(X) < a$, в) $\mathbf{E}(X) > a$.

```
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 > matplot(a, cbind(p, pl, pg), type="l", lty=c(1, 2, 4))
9 > abline(h=c(0, 0.05), lty=3)
```

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 различным альтернативным гипотезам H_1 (а–в).

В [8] выполняется построение кривых $\alpha_p(a)$ для основной гипотезы H_0: $\mathbf{E}(X) = a$ при альтернативных гипотезах H_1: $\mathbf{E}(X) \neq a$, $\mathbf{E}(X) < a$ и $\mathbf{E}(X) > a$. Указанные кривые изображаются на графике с параметром «`lty=c(1,2,4)`» с помощью сплошной, штриховой и штрихпунктирной линий.

В строке [9] с помощью функции «`abline()`» отмечается пятипроцентный уровень значимости, позволяющий приближённо оценить размеры доверительных интервалов для a к каждой из альтернативных гипотез H_1 (а–в).

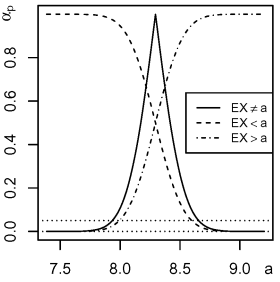


Рис. 3.10. Зависимости достигаемого уровня значимости $\alpha_p(a)$ для гипотезы H_0: $\mathbf{E}(X) = a$ при различных гипотезах H_1

Двухвыборочный t-критерий для независимых выборок

Для применения данного критерия помимо предположения о нормальности выборочных данных, также необходимо соблюдение условия равенства дисперсий: $\mathbf{D}(X) = \mathbf{D}(Y)$.

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 $H_0 : \mathbf{E}(X) = \mathbf{E}(Y)$ против *альтернативных* H_1: а) $\mathbf{E}(X) \neq \mathbf{E}(Y)$, б) $\mathbf{E}(X) < \mathbf{E}(Y)$, в) $\mathbf{E}(X) > \mathbf{E}(Y)$, где $\mathbf{E}(X)$, $\mathbf{E}(Y)$ — математическое ожидание случайных величин X и Y.

Статистика: При проверке нулевой гипотезы используется статистика, которая имеет распределение Стьюдента с числом степеней свободы $m + n - 2$:

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

где $\overline{x}$, $\overline{y}$ — выборочные средние значения; 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 $H_0 : \mathbf{E}(X) = \mathbf{E}(Y)$ на уровне значимости α *отвергается* в пользу альтернативных H_1: а) $\mathbf{E}(X) \neq \mathbf{E}(Y)$, б) $\mathbf{E}(X) < \mathbf{E}(Y)$, в) $\mathbf{E}(X) > \mathbf{E}(Y)$.

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

Пример 3.7. Для чётных и нечётных элементов выборки, использованной в примере 3.6: $\{x_i\}_n = \{u_j\}_m \cup \{v_k\}_l$, где $n = m+l$ — объёмы полной и частичных выборок, с помощью двухвыборочного t-критерия Стьюдента на уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу $H_0 : \mathbf{E}(U) = \mathbf{E}(V)$ при альтернативных H_1: а) $\mathbf{E}(U) \neq \mathbf{E}(V)$, б) $\mathbf{E}(U) < \mathbf{E}(V)$, в) $\mathbf{E}(U) > \mathbf{E}(V)$.

```
> u <- x[seq(1,99,2)]; v <- x[seq(2,100,2)]
> t.test(u,v, var.equal=TRUE, alter="two")
Two Sample t-test
data: u and v
t = -1.0833, df = 98, p-value = 0.2813
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1129101  0.3269101
sample estimates:
mean of x mean of y
    8.099     8.492
> t.test(u,v, var.equal=TRUE, alter="le")
Two Sample t-test
data:  u and v
t = -1.0833, df = 98, p-value = 0.1407
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.2094022
sample estimates:
mean of x mean of y
    8.099     8.492
> t.test(u,v, var.equal=TRUE, alter="gr")
Two Sample t-test
data:  u and v
t = -1.0833, df = 98, p-value = 0.8593
alternative hypothesis: true difference in means is greater than 0
```

```
95 percent confidence interval:
 -0.9954022        Inf
sample estimates:
mean of x mean of y
    8.099     8.492
```

В строке 9 вычисляются частичные выборочные векторы, содержащие чётные и нечётные элементы выборки $\{x_i\}_n = \{u_j\}_m \cup \{v_k\}_l$.

Функция «t.test()» в строках 11, 21 и 31 по двухвыборочному t-критерию Стьюдента вычисляет достигаемый уровень значимости для нулевой гипотезы $H_0 : \mathbf{E}(U) = \mathbf{E}(V)$ при альтернативных H_1: а) $\mathbf{E}(U) \neq \mathbf{E}(V)$, б) $\mathbf{E}(U) < \mathbf{E}(V)$, в) $\mathbf{E}(U) > \mathbf{E}(V)$.

В строках 14, 24 и 34 показаны достигаемые уровни значимости для соответствующих пар нулевой и альтернативных гипотез: а) $\alpha_p \approx$ $\approx 0{,}28$, б) $\alpha_p \approx 0{,}14$, в) $\alpha_p \approx 0{,}86$. Сравнение достигаемых уровней α_p с заданным $\alpha = 0{,}05$ позволяет для случаев (а, б) сделать выводы об удовлетворительном, а для случая (в) — о хорошем согласовании нулевой гипотезы H_0 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 $H_0 : \mathbf{D}(X) = \mathbf{D}(Y)$ против *альтернативных* H_1: а) $\mathbf{D}(X) \neq \mathbf{D}(Y)$, б) $\mathbf{D}(X) > \mathbf{D}(Y)$, где $\mathbf{D}(X)$, $\mathbf{D}(Y)$ — дисперсии случайных величин X и Y.

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 $H_0 : \mathbf{D}(X) = \mathbf{D}(Y)$ на уровне значимости α *отвергается* в пользу альтернативных H_1: а) $\mathbf{D}(X) \neq \mathbf{D}(Y)$, б) $\mathbf{D}(X) > \mathbf{D}(Y)$.

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

Пример 3.8. Для чётных и нечётных элементов выборки, использованной в примере 3.6: $\{x_i\}_n = \{u_j\}_m \cup \{v_k\}_l$, где $n = m + l$ — объёмы полной и частичных выборок, с помощью F-критерия Фишера на уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу $H_0 : \mathbf{D}(U) = \mathbf{D}(V)$ при альтернативных H_1: а) $\mathbf{D}(U) \neq \mathbf{D}(V)$, б) $\mathbf{D}(U) > \mathbf{D}(V)$; в) $\mathbf{D}(U) < \mathbf{D}(V)$.

```
> c(var(u), var(v))
[1] 3.144299 3.435894
> var.test(u, v, alter="two")
F test to compare two variances
data:  u and v
F = 0.9151, num df = 49, denom df = 49, p-value = 0.7575
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.519316 1.612636
sample estimates:
ratio of variances
         0.9151327
> var.test(u, v, alter="gr")
F test to compare two variances
data:  u and v
F = 0.9151, num df = 49, denom df = 49, p-value = 0.6213
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 0.569364      Inf
sample estimates:
ratio of variances
         0.9151327
> var.test(u, v, alter="le")
F test to compare two variances
data:  u and v
F = 0.9151, num df = 49, denom df = 49, p-value = 0.3787
```

```
67 alternative hypothesis: true ratio of variances is less than 1
68 95 percent confidence interval:
69  0.000000 1.470883
70 sample estimates:
71 ratio of variances
72           0.9151327
```

В строке 41 с помощью функции «var()» вычисляются соответствующие ранее использованным частичным выборочным векторам исправленные выборочные дисперсии: $s_u^2 \approx 3{,}14$ и $s_v^2 \approx 3{,}44$.

Функция «var.test()» в строках 43, 53 и 63 по F-критерию Фишера вычисляет достигаемый уровень значимости для нулевой гипотезы $H_0 : \mathbf{D}(U) = \mathbf{D}(V)$ при альтернативных H_1: а) $\mathbf{D}(U) \neq \mathbf{D}(V)$, б) $\mathbf{D}(U) > \mathbf{D}(V)$; в) $\mathbf{D}(U) < \mathbf{D}(V)$.

В строках 46, 56 и 66 показаны достигаемые уровни значимости для соответствующих пар нулевой и альтернативных гипотез: а) $\alpha_p \approx$ $\approx 0{,}76$, б) $\alpha_p \approx 0{,}62$, в) $\alpha_p \approx 0{,}38$. Сравнение достигаемых уровней значимости с заданным $\alpha = 0{,}05$ позволяет в случаях (а–б) сделать выводы о хорошем, а в случае (в) — об удовлетворительном согласовании нулевой гипотезы H_0 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, то 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 не будут статистически значимо отличаться друг от друга.

Представим результаты наблюдений в виде таблицы:

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

где i — уровни фактора; $j = 1, 2, \ldots, n_i$ — 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 $\operatorname{cov}(x_{ij}, x_{lm}) = 0$, где $i \neq l$ и/или $j \neq m$;
2. Совокупности наблюдаемых значений $\{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ на каждом уровне i нормально распределены: $\mathcal{N}(a_i, \sigma_i)$, где a_i, σ_i^2 — математическое ожидание и дисперсия i-го уровня;
3. Дисперсии распределений на всех уровнях $i = 1, 2, \ldots, k$ одинаковы: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \text{const}$.

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

Статистика: Рассмотрим следующие величины:

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

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

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

где Q_t — сумма квадратов отклонений отдельных наблюдений x_{ij} от общего среднего $\overline{x}$; Q_d — сумма квадратов отклонений средних значений уровней $\overline{x}_i$ от общей средней $\overline{x}$, 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 $\overline{x}_i$, 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 $H_0 : a_1 = a_2 = \ldots = a_k$ статистика F_s будет иметь распределение Фишера с числами степеней свободы числителя $(k-1)$, и знаменателя $(n-k)$:

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

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 $F_s \geqslant F_{\alpha,\frac{k-1}{n-k}}$ 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 зафиксированы 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))
5  > adhf[c(1:2, 10:11, 19:20),]
6     values ind
7  1     4.0   D
8  2     4.5   D
9  10    4.5   B
10 11    4.9   B
11 19    5.4   S
12 20    4.9   S
13 > anova(lm(values ~ ind, data=adhf))
14 Analysis of Variance Table
15 Response: values
16           Df Sum Sq Mean Sq F value   Pr(>F)
17 ind        2 4.9119 2.45593  7.7578 0.002519 **
18 Residuals 24 7.5978 0.31657
19 ---
20 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 первом, вектор «B» — на втором, а вектор «S» — на третьем уровне качественного признака.

Далее в строке 4 с помощью композиции «stack(data.frame())» формируется таблица исходных данных, некоторые строки которой показаны в 6–12, а в 13 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 17–18 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 и выборочной совокупностью, объёмом и статистикой выборки?
2. Запишите определения эмпирических функций распределения и плотности вероятности.
3. Напишите формулы для вычисления основных выборочных характеристик: среднего, дисперсии, ковариации, коэффициента корреляции.
4. Какую оценку называют точечной. Поясните свойства состоятельности, несмещённости и эффективность оценок.
5. Какая точечная оценка является состоятельной, несмещённой и эффективной для математического ожидания?
6. Какие точечные оценки для дисперсии являются смещёнными и несмещёнными? Являются ли эти оценки состоятельными?
7. Напишите формулы точечных оценок ковариации и коэффициента корреляции.
8. Что называют доверительной вероятностью и доверительным интервалом для неизвестного параметра θ?
9. Что такое статистическая гипотеза? Какие статистические гипотезы называют: основными или альтернативными, сложными или простыми?
10. Что называют статистическим критерием и его уровнем значимости при проверке статистической гипотезы?
11. В чем заключаются статистические ошибки первого и второго рода? Что такое достигаемый уровень значимости (p-уровень)?
12. В чем состоит принцип двойственности при построении доверительных интервалов и проверке гипотез о значениях параметров распределения?
13. Какие статистические критерии называют критериями согласия и критериями значимости различий?
14. Какие задачи являются объектом исследования в дисперсионном анализе? Каковы предпосылки однофакторного дисперсионного анализа?
15. Как формулируются основная и альтернативная гипотезы однофакторного дисперсионного анализа?

Глава 4

Метод главных компонент

В задачах анализа многомерных наблюдений типичными являются ситуации, когда общее число признаков, регистрируемых на каждом из наблюдаемых объектов, очень велико. В этом случае вполне естественным выглядит стремление представить каждое из наблюдений в виде вектора Z с существенно меньшим числом компонент. Именно, предполагается, что непосредственно наблюдаемые *признаки* (индикаторы) являются функциями гораздо меньшего числа неявных (скрытых), но объективно существующих признаков, называемых обычно *факторами*. Эта фундаментальная идея является основой целого класса статистических методов, называемых факторным анализом, к которому относят и метод главных компонент.

4.1. Постановка задачи

Формально задача перехода к новому набору признаков может быть описана следующим образом. Пусть имеется p-мерная величина $X = (x_1, x_2, \ldots, x_p)$ с вектором средних значений $a = (a_1, a_2, \ldots, a_p)$ и ковариационной матрицей $\Sigma = (\sigma_{ij})$, где $i = 1, 2, \ldots, p$ и $j = 1, 2, \ldots, p$. Определим на множестве признаков в качестве класса допустимых преобразований всевозможные линейные ортогональные нормированные комбинации, то есть будем полагать, что

$$z_j = \sum_{i=1}^{p} l_{ij}(x_i - a_i),$$

где $\sum_{i=1}^{p} l_{ij}^2 = 1$ и $\sum_{i=1}^{p} l_{ji} l_{ki} = 0$ для $j = 1, 2, \ldots, p$ и $k = 1, 2, \ldots, p$, но $k \neq p$. При этом потребуем, чтобы эти преобразования удовлетворяли условиям монотонности дисперсии $\mathbf{D}(z_1) \geqslant \mathbf{D}(z_2) \geqslant \ldots \geqslant \mathbf{D}(z_p)$. Полученные таким образом переменные $z_1(X)$, $z_2(X)$, $\ldots$, $z_p(X)$ и называются *главными компонентами*. В результате можно сформулировать следующие определения главных компонент [8].

Первой главной компонентой $z_1(X)$ системы признаков X называется такая нормированно-центрированная линейная комбинация

этих показателей, которая среди всех прочих линейных комбинаций такого рода обладает наибольшей дисперсией.

k-ой главной компонентой $z_k(X)$ системы показателей X при $k = 2, 3, \ldots, p$ называется такая линейная комбинация этих показателей, которая не коррелированна с предыдущими $(k-1)$ главными компонентами и среди всех прочих некоррелированных с предыдущими $(k-1)$ главными компонентами линейных комбинаций переменных x_1, x_2, …, x_p обладает наибольшей дисперсией.

Отметим, что поскольку все главные компоненты ранжированы по величине, то это позволяет, в конечном счёте, отбросить часть компонент, сохранив для анализа только те $z_1(X)$, $z_2(X)$, …, $z_m(X)$, где $m \ll p$, которые воспроизводят большую часть дисперсии:

$$\sum_{i=1}^{p} \mathbf{D}(x_i) \approx \sum_{j=1}^{m} \mathbf{D}(z_j).$$

Поскольку в реальных статистических задачах имеются, как правило, лишь оценки средних значений и ковариационной матрицы, то в дальнейшем не будем делать различий между этими статистическими характеристиками. Кроме этого без ограничения общности будем полагать, что все данные предварительно стандартизированы, и, следовательно, ковариационная матрица $\Sigma = (\sigma_{ij})$ является одновременно корреляционной матрицей исходных признаков.

4.2. Вычисление главных компонент

Покажем, что для того, чтобы величина $\mathbf{D}(z_1)$ достигала максимума при условии $\sum_{i=1}^{p} l_{1i}^2 = 1$, необходимо, чтобы вектор l_1 был собственным вектором, соответствующим максимальному собственному значению ковариационной матрицы $\Sigma = (\sigma_{ij})$, где $i = 1, 2, \ldots, p$ и $j = 1, 2, \ldots, p$.

Пусть дана матрица исходных стандартизированных данных

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

Тогда произведение $R = X^{\mathsf{T}} X$ будет соответствовать корреляционной матрице.

Предположим, что матрица $L = (l_{ij})$, где $i = 1, 2, \ldots, p$ и $j = 1, 2, \ldots, p$, позволяющая вычислить координаты объектов в пространстве компонент, известна. Тогда можно определить матрицу Z

$$Z = XL = \begin{pmatrix} z_{11} & z_{12} & \ldots & z_{1p} \\ z_{21} & z_{22} & \ldots & z_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ z_{n1} & z_{n2} & \ldots & z_{np} \end{pmatrix}. \tag{4.1}$$

По определению первой главной компоненты вектор $z_1 = (z_{11}, z_{21}, \ldots, z_{n1})^{\mathsf{T}}$ должен иметь максимальную дисперсию. В результате получаем задачу условной оптимизации:

$$\mathbf{D}(z_1) = \frac{1}{n}\sum_{j=1}^{n} z_{j1}^2 \to \max \quad \text{при} \quad \sum_{i=1}^{p} l_{i1}^2 = 1, \tag{4.2}$$

для решения которой составим функцию Лагранжа

$$\varphi(l_1, \lambda_1) = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{k=1}^{p} l_{j1}x_{k1}\right)^2 - \lambda_1\left(\sum_{j=1}^{p} l_{j1}^2\right). \tag{4.3}$$

Продифференцируем (4.3) по компонентам l_1 и приравняем полученные выражения к нулю:

$$\frac{\partial\varphi}{\partial l_{j1}} = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{k=1}^{p} l_{k1}x_{jk}\right)x_{j1} - \lambda_1 l_{k1} = 0, \quad j = 1, 2, \ldots, p. \tag{4.4}$$

Изменяя порядок суммирования в (4.4) и внося постоянный множитель $\frac{1}{n}$ под знак суммы, получим

$$\sum_{k=1}^{p}\left(\left(\frac{1}{n}\sum_{j=1}^{n} x_{jl}x_{jk}\right)l_{k1} - \lambda_1 l_{k1}\right) = 0. \tag{4.5}$$

Для стандартизированных данных X внутреннюю сумму можно обозначить как $r_{mk} = \frac{1}{n}\sum_{j=1}^{n} x_{jm}x_{jk}$. Тогда уравнение (4.5) примет вид

$$\sum_{k=1}^{p}(r_{mk}l_{k1} - \lambda_1 l_{k1}) = 0, \quad m = 1, 2, \ldots, p. \tag{4.6}$$

В матричной форме уравнение (4.6) представляет собой характеристическое уравнение для корреляционной матрицы R

$$R\,l_1 - \lambda_1 l_1 = 0. \tag{4.7}$$

Из характеристического уравнения (4.7) следует, что вектор l_1 является собственным вектором матрицы R, а множитель λ_1 — соответствующим собственным значением. С учётом (4.1) дисперсия $\mathbf{D}(Z) = Z^\mathsf{T}Z = (XL)^\mathsf{T}XL = L^\mathsf{T}X^\mathsf{T}XL = L^\mathsf{T}RL$. Для первой главной компоненты будем иметь $\mathbf{D}(z_1) = l_1^\mathsf{T}R\,l_1$. Умножив уравнение (4.7) слева на l_1^T и принимая во внимание условие нормировки (4.2), получим

$$l_1^\mathsf{T}R\,l_1 = l_1^\mathsf{T}l_1\lambda_1 = \lambda_1\mathbf{D}(z_1). \tag{4.8}$$

Таким образом, для выполнения условий (4.2) для первой главной компоненты z_1, необходимо взять максимальное собственное значение матрицы R и соответствующий собственный вектор.

Аналогичным образом находятся остальные главные компоненты. Существование главных компонент гарантируется теоремой, утверждающей, что симметричные и неотрицательно определённые матрицы ранга p, каковой и является матрица R, имеют p вещественных неотрицательных собственных значений: $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$.

4.3. Основные свойства главных компонент

Обычно применение главных компонент даёт хороший результат в том случае, когда исходные переменные $X = (x_1, x_2, \ldots, x_p)$ сильно коррелированны. В этом случае их ковариационная матрица $\Sigma = (\sigma_{ij})$ плохо обусловлена, поскольку её определитель близок к нулю. С другой стороны, величина этого определителя соответствует произведению собственных значений

$$\det\Sigma = \prod_{i=1}^{p}\lambda_i.$$

Поэтому следует ожидать, что последние собственные значения окажутся достаточно малы, и соответствующие им главные компоненты можно будет исключить из рассмотрения. В результате осуществляется переход к меньшему числу новых переменных $z_1(X)$, $z_2(X)$, …, $z_m(X)$, где $m \ll p$. Это переход можно рассматривать как проекцию исходных данных $X \in \mathbf{R}^p$ в некоторое подпространство меньшей размерности $Z \in \mathbf{R}^m$.

Можно доказать, что среди всех подпространств заданной размерности m, в подпространстве $\mathbf{R}^m$, натянутом на первые главные компоненты, расстояния от рассматриваемых точек до общего «центра

тяжести» $\overline{X}$, а также углы между прямыми, соединяющими всевозможные пары точек наблюдений с «центром тяжести» $\overline{X}$, искажаются наименьшим образом. Следовательно, по значениям первой и второй главных компонент можно получить на соответствующей плоскости наилучшее представление о форме и структуре исходных многомерных данных. Именно это свойство главных компонент используется для оптимальной визуализации данных.

Иногда проводят различие между методом главных компонент, — когда критерием оптимальности является приближение ковариационной матрицы, и методом главных факторов, — когда речь идёт о приближении корреляционной матрицы. Однако, такие различия не всегда поддерживаются в литературе по прикладному статистическому анализу, поскольку в практических задачах использование стандартизированных данных является необходимым условием сопоставимости признаков различной природы.

Пример 4.1. Используя выборочные данные по ирисам Фишера $\{(x_1, x_2, x_3, x_4)_i\}$ для $i = 1, 2, \ldots, 150$ выбрать главные компоненты z_1 и z_2, обеспечивающие оптимальную визуализацию данных.

```
> data(iris)
> iris[c(1:2,75:76,149:150),]
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
75           6.4         2.9          4.3         1.3 versicolor
76           6.6         3.0          4.4         1.4 versicolor
149          6.2         3.4          5.4         2.3  virginica
150          5.9         3.0          5.1         1.8  virginica
> pairs(x <- iris[-5], pch=dots <- as.numeric(iris[,5]))
> summary(pca <- prcomp(x, center=TRUE, scale=TRUE))
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
> spec <- levels(iris[,5])
> windows()
> plot(pca[[5]], pch=dots)
> legend("topright", pch=seq(3), legend=paste("i.",spec))
> Dz <- pca[[1]]^2/sum(pca[[1]]^2)
> windows()
> barplot(cumsum(Dz), col="gray67", names.arg=paste("z",seq(4)))
> barplot(Dz, col="gray50", axes=FALSE, add=TRUE); box()
> windows()
```

```
26 > pairs(pca[[5]], pch=dots)
```

Используемая для анализа таблица выборочных данных загружается под именем «iris» в строке 1. Указанный набор данных известен под названием «ирисов Фишера», впервые использованного в 1936 году статье Р. Фишера по дискриминантному анализу [1] и включающего результаты измерений для 150 экземпляров ирисов трёх видов: iris setosa (50 экз.), iris versicolor (50 экз.) и iris virginica (50 экз.).

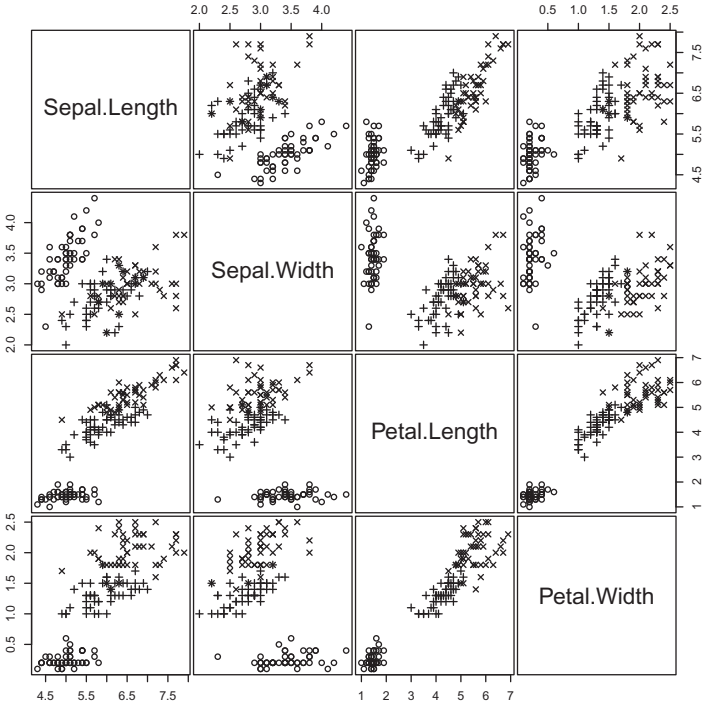


Рис. 4.1. Диаграммы рассеяния для каждой пары выборочных признаков. Символы "○" соответствуют выборочным данным по экземплярам вида iris setosa, "+" — по iris versicolor, а "×" — по iris virginica

[1] Fisher R.A. The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics. — 1936. — Vol.7. — P.179–188.

Для каждого экземпляра ириса измерялись в сантиметрах четыре характеристики: длина и ширина чашелистика (sepal length and width), длина и ширина лепестка (petal length and width). Команда «`iris[c(1:2,75:76,149:150),]`» отображает по две строки от начала, середины и конца этой таблицы 2–9.

Ирисы Фишера являются классическим набором данных, который часто используется в статистической литературе для иллюстрации работы различных алгоритмов. Диаграммы рассеяния для каждой пары выборочных признаков приведены на рис. 4.1. Для построения диаграмм в 10 используется функция «`pairs()`» с первым аргументом «`x<-iris[-5]`», что соответствует визуализации первых четырёх столбцов данных, и «`pch=dots<-as.numeric(iris[,5])`», что обеспечивает использование различных символов для отображения выборочных данных по ирисам каждого вида. На приведённом рисунке хорошо видно, что выборочные данные группируются в три кластера, форма и расположение которых сильно зависят от пары признаков, использованных для визуализации.

Для анализа главных компонент в 11 используются функции: «`prcomp(iris[-5],center=TRUE,scale=TRUE)`» — для стандартизации выборочных данных и выделения главных компонент; «`summary()`» — для печати распределения средних квадратических отклонений и относительных долей дисперсии по выделенным главным компонентам. В строках 12–16 хорошо заметно, что на две первых компоненты приходится почти 96% от общей дисперсии данных, в то время как на две последних компоненты — чуть больше 4%.

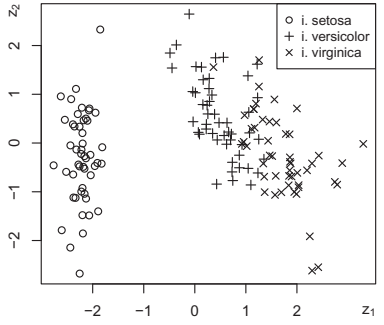


Рис. 4.2. Диаграмма рассеяния по двум главным компонентам z_1 и z_2

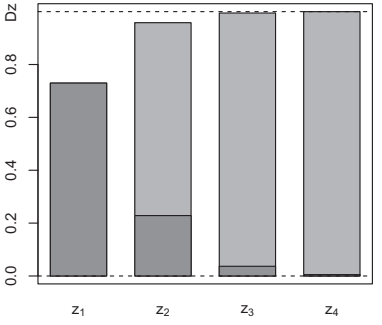


Рис. 4.3. Относительные доли дисперсии главных компонент $\mathbf{D}(z_i)$

Диаграмма рассеяния выборочных данных по первым двум главным компонентам z_1 и z_2 приведена на рис. 4.2, а по всем парам главных компонент — на рис. 4.4. Так же как и на рис. 4.1 символы "$\circ$" соответствуют выборочным данным по экземплярам вида iris setosa, символы "$+$" — по экземплярам iris versicolor, а символы "$\times$" — по экземплярам iris virginica. На рис. 4.3 показаны диаграммы распределения суммарной и индивидуальных относительных долей дисперсии по главным компонентам $\mathbf{D}(z_i)$.

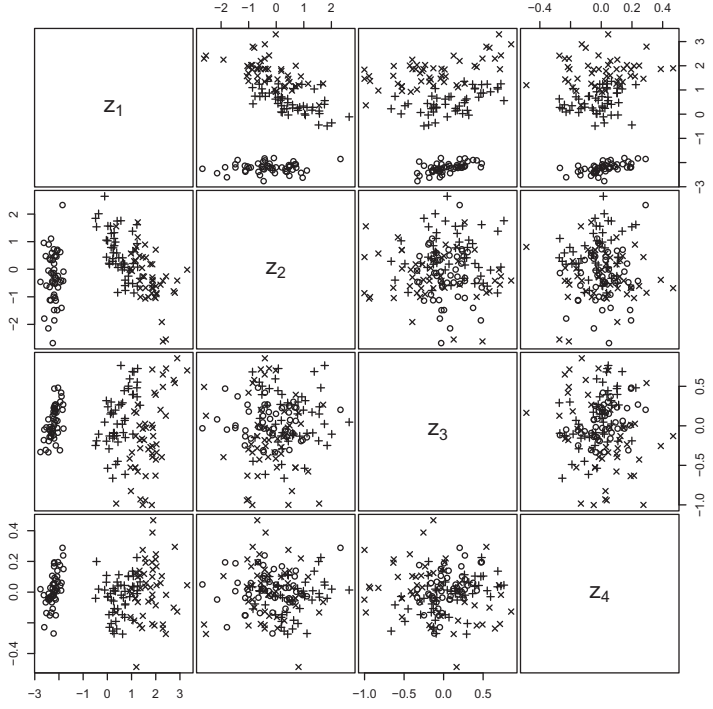


Рис. 4.4. Диаграммы рассеяния для каждой пары главных компонент. Символы "$\circ$" соответствуют выборочным данным по экземплярам вида iris setosa, "$+$" — по iris versicolor, а "$\times$" — по iris virginica

Для построения приведённых на рис. 4.2, 4.3 и 4.4 диаграмм в [17–24] были использованы следующие функции: «`windows()`» — для создания нового графического окна; «`plot(pca[[5]], pch=dots)`» —

для построения диаграммы рассеяния по первым двум главным компонентам, используя те же символы, что и на рис. 4.1; «`Dz <- ...`» — для расчёта индивидуальных относительных долей дисперсии главных компонент; «`barplot(cumsum(Dz),...)`» — для построения столбчатой диаграммы суммарной относительной доли дисперсии главных компонент $\mathbf{D}(z_i)$; «`barplot(Dz,...add=TRUE)`» — для добавления к столбчатой диаграмме индивидуальных относительных долей дисперсии главных компонент $\mathbf{D}(z_i)$; «`pairs(pca[[5]],pch=dots)`» — для построения диаграмм рассеяния выборочных данных для каждой пары главных компонент z_1, z_2, z_3, z_4.

Сопоставление диаграмм рассеяния, приведённых на рис. 4.1 и 4.4, наглядно демонстрирует, что непосредственное использование выборочных признаков в общем случае не позволяет выделить наилучшую пару для оптимальной визуализации, в то время как применение метода главных компонент приводит к вполне однозначному решению поставленного вопроса.

Контрольные вопросы

1. В чем разница между признаками (индикаторами) и факторами в постановке задачи факторного анализа? Какова основная цель метода главных компонент?
2. Сформулируйте в общем виде постановку задачи факторного анализа. Каковы основные предпосылки использования метода главных компонент?
3. Запишите формальные соотношения для вычисления главных компонент и перечислите их основные свойства.
4. Приведите геометрическую интерпретацию метода главных компонент. Какие свойства главных компонент используются при визуализации многомерных данных?
5. Дайте интерпретацию факторных нагрузок.
6. Как оценивается информативность первых k главных компонент?
7. Сформулируйте отличительные особенности метода главных компонент от общих подходов факторного анализа.

Глава 5

Начала регрессионного анализа

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, ..., 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, ..., x_k и регрессия называется *множественной*.

5.1. Парная линейная регрессия

Рассмотрим построение модели парной регрессии:

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

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 β_0, β_1, …, β_k обычно применяют *метод наименьших квадратов* (МНК). Однако существуют и другие методы оценки: метод максимального правдоподобия, метод наименьших модулей и тому подобное.

5.1.1. Оценка параметров уравнения регрессии

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

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

Будем рассматривать случайное слагаемое ε как последовательность n случайных величин: e_1, e_2, …, 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 $\hat{y}_i = b_0 + b_1 x_i$:

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

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

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

После преобразований получим *систему нормальных уравнений МНК*:

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

Решая систему нормальных уравнений, находим b_0, b_1:

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

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

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.

5.1.2. 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 $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$ — *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:

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

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*:

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

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.

5.1.3. 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*: $\mathbf{E}(\varepsilon_i) = 0$, где $i = 1, 2, \ldots, n$. 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]: $\mathbf{D}(\varepsilon_i) = \sigma^2$, где $i = 1, 2, \ldots, n$. 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: $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$, где $i, j = 1, 2, \ldots, n$. 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 $\operatorname{cov}(x_i, \varepsilon_j) = 0$.
4. Случайные составляющие ε_i распределены по *нормальному закону*: $\varepsilon_i \sim \mathcal{N}(0, \sigma)$, где $i = 1, 2, \ldots, n$. 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

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

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

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

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

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

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

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

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.

5.1.4. 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: $y \sim \mathcal{N}\left(\hat{y}, k_{\hat{y}}\sigma_{\hat{y}}\right)$, где

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

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:

$$\mathbf{I}_\gamma\left(y\right) = \left(\hat{y} - t_{\gamma,n-2}k_{\hat{y}}s_{\hat{y}},\ \hat{y} + t_{\gamma,n-2}k_{\hat{y}}s_{\hat{y}}\right).$$

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:

$$\mathbf{I}_\gamma\left(\dot{y}\right) = \left(\hat{y} - t_{\gamma,n-2}k_{\dot{y}}s_{\hat{y}},\ \hat{y} + t_{\gamma,n-2}k_{\dot{y}}s_{\hat{y}}\right),$$

где

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

Пример 5.1. Для заданных векторов выборочных данных $\{(x,y)_i\}$ построить линейную модель парной регрессии и проверить её качество.

```
1  > x <- c(1.89, 2.21, 2.37, 2.91, 2.72, 3.55, 3.84, 4.13, 4.25, 4.88)
2  > y <- c(0.75, 0.59, 0.19, 0.02, 0.04,-0.72,-0.85,-1.35,-1.36,-1.83)
3  > summary(fit <- lm(y ~ x))
4  Call:
5  lm(formula = y ~ x)
6  Residuals:
7       Min       1Q   Median       3Q      Max
8  -0.16195 -0.03849 -0.01237  0.08388  0.14776
9  Coefficients:
10             Estimate Std. Error t value Pr(>|t|)
11 (Intercept)  2.45732    0.12252   20.06 3.98e-08 ***
12 x           -0.88834    0.03594  -24.72 7.67e-09 ***
13 ---
14 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
15 Residual standard error: 0.1076 on 8 degrees of freedom
16 Multiple R-squared:  0.9871,        Adjusted R-squared:  0.9855
17 F-statistic:   611 on 1 and 8 DF,  p-value: 7.668e-09
18 > xin <- seq(0.8*min(x), 1.2*max(x), length=100)
19 > pre <- predict(fit, data.frame(x=xin), interval="confidence")
20 > plot(dat, pch=16)
21 > matplot(xin, pre, type="l", lty=c(1,2,2), add=TRUE)
22 > windows(); par(mfrow=c(2,1))
23 > plot(x, fit[[2]], pch=4); abline(h=0)
24 > qqnorm(fit[[2]], pch=8); qqline(as.vector(fit[[2]]))
```

Используемые для построения регрессионной модели выборочные данные определены в строках 1–2 под именами: «x» и «y». Функция «lm()» в строке 3 рассчитывает параметры линейной модели вида «y ~ x», что соответствует уравнению парной регрессии: $y_i = b_0 + b_1 x_i + e_i$. Функция «summary()» выводит сводку основных результатов расчёта в строках 4–17. В строках 11–12 приведены оценки коэффициентов выборочного уравнения регрессии: $b_0 \approx 2{,}46$, $b_1 \approx -0{,}88$, а также соответствующие значения стандартных ошибок и вероятности отклонения гипотез о равенстве полученных оценок истинным значениям: $\mathbf{P}\{\beta_0 \neq b_0\} \approx 4{,}0 \cdot 10^{-8}$, $\mathbf{P}\{\beta_1 \neq b_1\} \approx 7{,}7 \cdot 10^{-9}$. С учётом приведённого в строке 16 значения выборочного коэффициента детерминации $R^2 \approx 0{,}99$ качество построенного уравнения регрессии можно охарактеризовать как высокое.

График к полученной линейной модели парной регрессии показан на рис. 5.1. Для построения этого рисунка в строках 30–33 используются функции: «predict(...interval="confidence")» — вычисление границ 0,95-доверительных интервалов для уравнения регрессии;

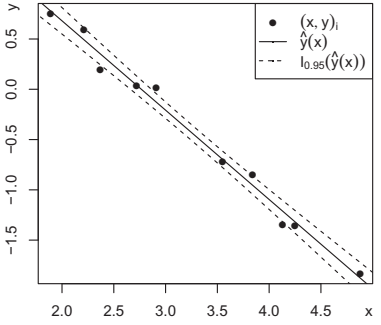


Рис. 5.1. Выборочные точки $(x, y)_i$ и $\mathbf{I}_{0.95}(\hat{y})$, где $\hat{y} = 2{,}46 - 0{,}88x$

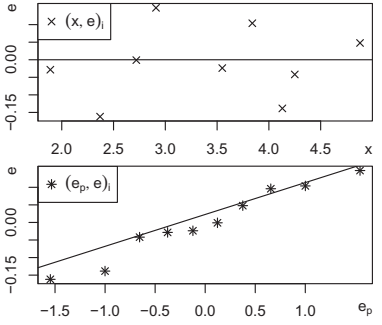


Рис. 5.2. Обычный и квантильный графики остатков $(x, e)_i$ и $(e_p, e)_i$

«`plot(...pch=16)`» — отображение выборочных значений (x_i, y_i), используя символы «•»; «`matplot()`» — отображение графика выборочного уравнения регрессии $\hat{y} = 2{,}46 - 0{,}88x$, а также верхней и нижней границ его доверительных интервалов $\mathbf{I}_{0,95}(\hat{y})$, 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 с помощью обычного и/или квантильного графиков, показанных на рис. 5.2. Для построения этих графиков в строках 22–24 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 «×»; «`abline(h=0)`» — отображение горизонтальной линии на нулевом уровне; «`qqnorm()`» — отображение на квантильном графике остатков e_i для исходных данных линейной модели; «`qqline()`» — отображение на квантильном графике функции нормального распределения $F_0(e) = \Phi(\frac{e}{0{,}1}) + \frac{1}{2}$, где значение $s_e \approx 0{,}1$ соответствует исправленному выборочному среднему квадратическому отклонению остатков.

5.2. Множественная линейная регрессия

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 всех факторов, кроме одного, на объект исследования можно пренебречь. Например, если коэффициент детерминации для построенного уравнения регрессии близок к единице: $R^2 \geqslant 0{,}8$. Однако в практических задачах такие ситуации являются скорее исключением, чем правилом. Поэтому модели множественной линейной регрессии имеют довольно широкое распространение.

5.2.1. Метод наименьших квадратов для множественной регрессии

Рассмотрим регрессионное уравнение, в котором определяется линейная связь зависимой переменой y от k независимых переменных x_1, x_2,..., x_k:

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

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

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

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

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

Рассмотрим следующие векторы и матрицы:

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

Столбцами матрицы $\boldsymbol{X}$ являются векторы $\boldsymbol{X}_s = (x_{1s}, x_{2s}, \ldots, x_{ns})$, где $s = 0, 1, \ldots, k$, соответствующие независимым переменным x_1, x_2,..., 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, $b_1,\ldots$, b_k в матричной форме примет вид

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e},$$

а подлежащая минимизации сумма квадратов отклонений

$$\sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^{\mathsf{T}}\boldsymbol{e} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^{\mathsf{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) \to \min.$$

Решение такой задачи базируется на простых геометрических соображениях. Рассмотрим в качестве примера модель линейной регрессии для двух наблюдений: $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, где $\boldsymbol{Y} = \binom{y_1}{y_2}$, $\boldsymbol{X} = \binom{x_1}{x_2}$ — векторы наблюдений, $\varepsilon = \binom{\varepsilon_1}{\varepsilon_2}$ — вектор случайной составляющей. Этой модели соответствуют построения, показанные на рис. 5.3.

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

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

$$\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = 0.$$

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

$$\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{Y}.$$

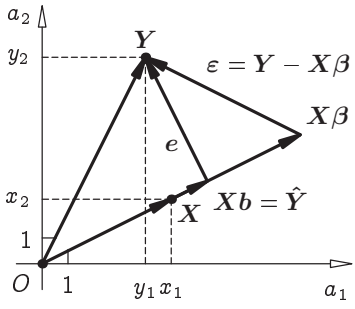


Рис. 5.3. Геометрическая интерпретация модели линейной регрессии на плоскости a_1a_2

Если матрица системы $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ невырожденная, то система нормальных уравнений будет иметь искомое решение

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Оценки $\boldsymbol{b}$ вектора $\boldsymbol{\beta}$, 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 $\boldsymbol{b}$ позволяет вычислять расчётные значения зависимой переменной $\hat{\boldsymbol{Y}}$

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

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

5.2.2. Статистические свойства МНК-оценок множественной регрессии

Теорема Гаусса–Маркова: Предположим, что:

1. Дана определённая ранее модель *множественной линейной регрессии*: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$;
2. Здесь $\boldsymbol{X}$ — детерминированная *матрица*, имеющая максимальный *ранг* $k + 1$; практически это означает линейную независимость векторов-столбцов матрицы $\boldsymbol{X}$, откуда следует *невырожденность* матрицы $\boldsymbol{X}^\top \boldsymbol{X}$;
3. $\mathbf{E}(\boldsymbol{\varepsilon}) = 0$, $\mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \boldsymbol{I}$, где $\boldsymbol{I}$ — 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 $\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ будут *несмещёнными* и *эффективными* в классе линейных несмещённых оценок.

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

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

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

5.2.3. Оценка качества уравнения множественной регрессии

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

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

Полученное разложение суммы квадратов можно непосредственно увидеть на рис. 5.3 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 $Q_t = \boldsymbol{Y}\boldsymbol{Y}^\mathsf{T} - n\overline{\boldsymbol{Y}}^2$ — общая сумма квадратов отклонений $\boldsymbol{Y}$ относительно среднего $\overline{\boldsymbol{Y}}$; $Q_e = \boldsymbol{Y}\boldsymbol{Y}^\mathsf{T} - \boldsymbol{b}^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}$ — сумма квадратов отклонений $\boldsymbol{Y}$, относительно расчётных значений по уравнению регрессии $\hat{\boldsymbol{Y}}$; $Q_r = \boldsymbol{b}^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y} - n\overline{\boldsymbol{Y}}^2$ — сумма квадратов отклонений расчётных значений $\hat{\boldsymbol{Y}}$ относительно среднего $\overline{\boldsymbol{Y}}$ или остаточная сумма квадратов. Все использованные обозначения соответствуют ранее введённым.

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

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

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

Если $R^2 = 0$, то $\|\boldsymbol{e}\|^2 = \|\boldsymbol{Y}\|^2$, то есть весь разброс величины $\boldsymbol{Y}$ соответствует случайным отклонениям, называемым ошибками. В этом

случае $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}}$ 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 $\|\boldsymbol{e}\|^2 = 0$ 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: $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k$ с зависимой переменной $\boldsymbol{Y}$, 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 $\boldsymbol{Y}$ и расчётными $\hat{\boldsymbol{Y}}$, то есть $|R| = |r_{\boldsymbol{Y}\hat{\boldsymbol{Y}}}|$.

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:

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

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.

5.2.4. 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: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

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 $\boldsymbol{Y}$ оказывает влияние хотя бы одна независимая переменная $\boldsymbol{X}_i$. И наоборот, уравнение регрессии считается незначимым, если все переменные $\boldsymbol{X}_i$ не связаны с $\boldsymbol{Y}$, то есть не оказывают на неё никакого влияния. В этом случае вся изменчивость величины $\boldsymbol{Y}$ 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 $\boldsymbol{X}_j$ равны нулю H_0: $\beta_j = 0$ при $j = 1, 2, \ldots, k$. Тогда альтернативная гипотеза будет состоять в том, что существует хотя бы одна переменная $\boldsymbol{X}_j$, 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 признается *значимым с надёжностью* γ и может быть использовано в практических расчётах.

Если же расчётное значение статистики F_s не превышает α-квантиль: $F_s \leqslant F_{\alpha,\frac{k}{n-k-1}}$, то в этом случае говорят, что имеющиеся *данные не позволяют отвергнуть нулевую гипотезу* на выбранном уровне значимости α, а *уравнение регрессии* признается *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.

5.2.5. 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 $\boldsymbol{Y}$ каждой переменной $\boldsymbol{X}_k$ 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, $i = 0, 1, \ldots, k$ будут иметь нормальные распределения, а их нормированные отклонения от истинных значений — распределе-

ния 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: $s_{b_i}^2 = s^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}$.

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 $\boldsymbol{Y}$, то есть точность прогноза.

Пусть вектор $\boldsymbol{X}_p = (1, x_{p1}, x_{p2}, \ldots, x_{pk})$ представляет значения независимых переменных, при которых требуется определить значение зависимой переменной $\boldsymbol{Y}$. Тогда $\hat{y}_p = b_0 + b_1x_{p1} + \ldots + b_kx_{pk}$ равняется условному математическому ожиданию (среднему значению) переменной $\boldsymbol{Y}$ при $\boldsymbol{X} = \boldsymbol{X}_p$. Рассуждая аналогично вышеизложенному, получим

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

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 $n-k-1$; $s^2_{\hat{y}_p} = s^2(\boldsymbol{X}_p^{\mathsf{T}}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_p)$. Аналогичный вид имеет и доверительный интервал для индивидуального значения переменной $\boldsymbol{Y}$, но его стандартная ошибка $s_{\hat{y}_p}$ будет вычисляться иначе: $s^2_{\hat{y}_p} = s^2(I + \boldsymbol{X}_p^{\mathsf{T}}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_p)$.

Пример 5.2. Для заданных векторов выборочных данных $\{(x_1, x_2, y)_i\}$ построить линейную модель множественной регрессии и проверить её качество.

```
> x1<- c(-9.08,-8.07,-7.75,-6.89,-7.01,-6.43,-6.02,-5.28,-4.71,-4.33)
> x2<- c( 2.37, 2.05, 3.45, 3.22, 3.97, 4.18, 5.74, 5.67, 6.00, 6.65)
> y <- c(-1.80,-2.61,-0.41,-1.61,-0.12,-0.04, 2.90, 1.56, 1.52, 2.55)
> summary(fit <- lm(y ~ x1 + x2))
Call:
lm(formula = y ~ x1 + x2)
Residuals:
     Min       1Q   Median       3Q      Max
-0.42254 -0.13359 -0.00026  0.13106  0.49084
  Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.5732     2.0670  -6.567 0.000314 ***
x1           -0.8487     0.1965  -4.320 0.003480 **
x2            1.8943     0.1858  10.197 1.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.2936 on 7 degrees of freedom
Multiple R-squared:  0.9812,    Adjusted R-squared:  0.9758
F-statistic: 182.7 on 2 and 7 DF,  p-value: 9.104e-07
> require(plot3D)
> scatter3D(x1, x2, y, pch=16, xlab="x1", ylab="x2", zlab="y")
> segments3D(x1, x2, rep(0, length(x1)), x1, x2, y, add=TRUE)
> x1m <- min(x1); x1x <- max(x1)
  x2m <- min(x2); x2x <- max(x2)
> polygon3D(c(x1m, x1m, x1x, x1x), (x2m, x2x, x2x, x2m),
+           rep(0, 4), border="black", facets=NA, add=TRUE)
> windows(); par(mfrow=c(2,1))
> plot(x1, y); segments(x1, 0, x1, y); abline(h=0)
> plot(x2, y); segments(x2, 0, x2, y); abline(h=0)
> windows(); par(mfrow=c(2,1))
> termplot(fit, se=TRUE, partial.resid=TRUE)
> windows(); par(mfrow=c(3,1))
> plot(x1, fit$res, pch=4); abline(h=0, lty=1)
> plot(x2, fit$res, pch=4); abline(h=0, lty=1)
> qqnorm(fit$res, pch=8); qqline(as.vector(fit$res))
```

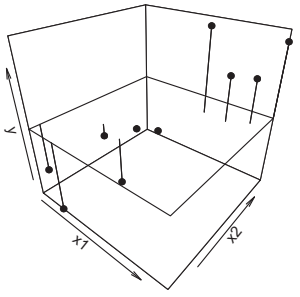


Рис. 5.4. Расположение аппликат выборочных значений $(x_1, x_2, y)_i$ относительно плоскости $y = 0$

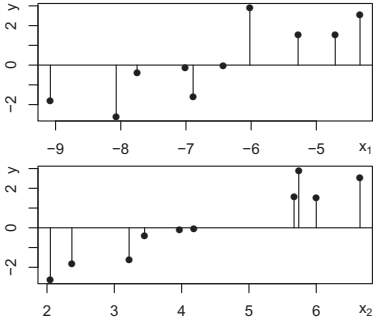


Рис. 5.5. Проекции аппликат выборочных значений $(x_1, y)_i$ и $(x_2, y)_i$ на плоскости $x_1 = 0$ и $x_2 = 0$

Используемые для построения регрессионной модели выборочные данные определены в строках 1–3 под именами: «x1», «x2» и «y». 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 4 на основе указанных выборочных данных рассчитывает параметры линейной модели «y ~ x1 + x2», которая соответствует регрессии: $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$.

Сводка основных результатов выводится в строках 5–19 с помощью функции «summary()». В строках 12–14 показаны оценки коэффициентов уравнения регрессии: $b_0 \approx -13{,}57$, $b_1 \approx -0{,}85$, $b_2 \approx -1{,}89$, соответствующие значения стандартных ошибок и вероятности отклонения гипотез о равенстве полученных оценок истинным значениям: $\mathbf{P}\{\beta_0 \neq b_0\} \approx 0{,}0003$, $\mathbf{P}\{\beta_1 \neq b_1\} \approx 0{,}0035$, $\mathbf{P}\{\beta_2 \neq b_2\} \approx 1{,}9 \cdot 10^{-5}$. С учётом значения скорректированного выборочного коэффициента детерминации $R_a^2 \approx 0{,}98$, показанного в строке 18, качество уравнения регрессии можно охарактеризовать как высокое.

Графики используемых для построения модели исходных данных показаны на рис. 5.4 и 5.5. Для их построения в строках 20–29 используются функции: «require(plot3D)» — загрузка библиотеки трёхмерной графики; «scatter3D()», «segments3D()», «polygon3D()» — отображение выборочных точек, отрезков аппликат и фрагмента плоскости в пространстве; «windows()» — создание нового графического окна; «par(mfrow=...)» — разделение графического окна на равные секции по вертикали; «segments()» — отображение отрезков ординат

на плоскости.

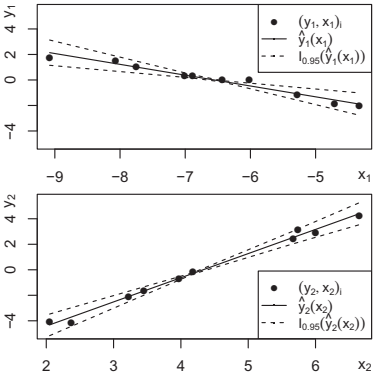


Рис. 5.6. Частные выборочные значения $(x_1, y_1)_i$, $(x_2, y_2)_i$ и доверительные интервалы $\mathbf{I}_{0,95}\,(\hat{y}_1(x_1))$, $\mathbf{I}_{0,95}\,(\hat{y}_2(x_2))$ для уравнения регрессии $\hat{y} = -13{,}57 - 0{,}85x_1 - 1{,}89x_2$

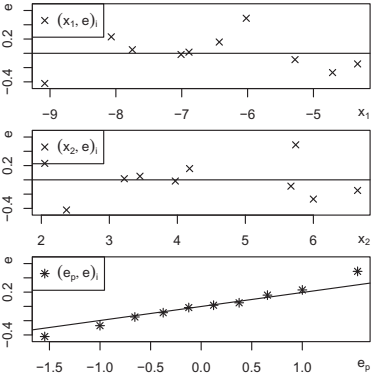


Рис. 5.7. Центрированные по $\hat{y}$ остатки значений $(x_1, e)_i$ и $(x_2, e)_i$. Квантильные графики остатков $(e, e_p)_i$ и функции распределения $F_0(e) = \Phi(\frac{e}{0{,}26}) + \frac{1}{2}$

Графики к полученной модели множественной регрессии показаны на рис. 5.6. Для их построения в строке 31 используется функция: «`termplot()`» — отображение частных графиков выборочного уравнения регрессии $\hat{y}(x_1, x_2) = \hat{y}_1(x_1) + \hat{y}_2(x_2)$, а также верхних и нижних границ доверительных интервалов $\mathbf{I}_{0,95}\,(\hat{y}_1)$ и $\mathbf{I}_{0,95}\,(\hat{y}_2)$.

После построения модели полезно провести анализ распределения её остатков e_i, показанных на рис. 5.7 вверху и в середине. Это можно сделать с помощью квантильного графика, показанного на рис. 5.7 внизу. Для построения этих графиков в строках 32–35 используются функции: «`windows()`» — создание нового графического окна; «`par(mfrow=...)`» — разбиение графического окна на равные секции по вертикали; «`plot(...pch=4)`» — отображение остатков $(x_1, e)_i$ и $(x_2, e)_i$, используя символ «×»; «`abline(h=0)`» — отображение горизонтальной линии на нулевом уровне; «`qqnorm()`» — отображение на квантильном графике остатков $(e_p, e)_i$ исходных данных линейной модели; «`qqline()`» — отображение на квантильном графике функции нормального распределения $F_0(e) = \Phi(\frac{e}{0{,}26}) + \frac{1}{2}$, где значение $s_e \approx 0{,}26$ соответствует исправленному выборочному среднему квадратическому отклонению остатков.

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?

Глава 6

Основы кластерного анализа

Задача классификации в самом общем виде заключается в том, чтобы в имеющейся совокупности выделить классы объектов, однородных в каком-то смысле, то есть разделить заданное множество таким образом, чтобы объекты, отнесённые к одному классу, были более схожи между собой, чем объекты, принадлежащие различным классам. При этом неявно предполагается, что классы однородных объектов составляют схожие (естественно связанные) объекты.

В такой постановке задачи классификации независимо друг от друга рассматривались в различных прикладных отраслях. Это нашло свое отражение, как в различных названиях самой задачи, так и в методах её решения. Например, в биологии эта задача называлась задачей численной таксономии, в технических приложениях — задачей автоматической классификации или распознавания образов без учителя, а в социально-экономических исследованиях — задачей многомерной классификации.

6.1. Содержательная постановка задачи

Для обозначения математических методов решения классификационной задачи в последние десятилетия наиболее широкое распространение получил термин *кластерного анализа.* В настоящее время этим термином обозначают раздел прикладного статистического анализа, в рамках которого с единых позиций рассматриваются подходы к решению этой задачи. Закреплению и распространению этого термина в значительной степени способствовало то, что он был широко использован в современной литературе и в прикладном программном обеспечении.

Содержательно задачу кластерного анализа можно сформулировать следующим образом [8]: заданную совокупность объектов, для каждой пары которых определена мера сходства, разбить на однородные в некотором смысле группы объектов. Полученные в результате группы объектов называются *кластерами* (классами, таксонами

или образами). Очевидно, что решение сформулированной задачи в значительной степени зависит от того, каким образом определяется *сходство объектов между собой*, а также как для полученного разбиения в целом оценивается его *однородность*. В отличие от других классификационных задач, кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных. Однако, необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

6.2. Формальная постановка задачи

Построение любой классификации начинается с отбора признаков объектов. Этот этап характеризуется рядом самостоятельных проблем, которых мы не будем касаться без особой необходимости. Необходимо только отметить, что классификация имеет практическую ценность тогда, когда выбранные признаки, на основе которых производится выделение кластеров, являются существенными для того, чтобы созданное разбиение сохраняло важные для исследователя характеристики, признаки и связи изучаемого феномена, то есть было пригодно для дальнейшей работы.

6.2.1. Измерение сходства объектов

Переход от содержательной постановки задачи к её математической формализации в большинстве случаев опирается на геометрическое представление данных о классифицируемых объектах. Суть этого подхода заключается в том, что каждый из n объектов заданной совокупности описывается одинаковым набором из p признаков. Значения этих признаков позволяют поставить в соответствие любому объекту точку p-мерного пространства, каждая из координат которой соответствует определённому признаку. Пространство, в котором рассматриваемые объекты представлены точками, называется *признаковым пространством*. Определяемая в признаковом пространстве метрика является основой для введения меры сходства объектов.

В основу геометрического подхода положено предположение о том, что признаки и метрика выбраны таким образом, что близким в содержательном смысле объектам соответствуют близкие в смысле выбранной меры сходства точки признакового пространства. Понятие

кластера как однородной совокупности точек признакового пространства отражает представление исследователя о тех свойствах, которыми должны обладать объекты, принадлежащие к одному классу.

Обычно результаты измерений анализируемого набора p признаков на каждом из n объектов представляют в виде матрицы данных $X = (x_{ij})$ размера $n \times p$, элементы которой соответствуют значению j-го признака на i-ом объекте. В качестве *меры взаимного сходства* объектов часто используются функции *расстояний* ρ. В результате задания в признаковом пространстве расстояния (как меры сходства) осуществляется переход к симметричной матрице $R = (\rho_{ij})$, элементы которой соответствуют расстоянию между объектами x_i и x_j. Выбор конкретного вида функции ρ_{ij} определяется способом измерения признаков.

Так, если все признаки измерены в интервальной шкале (или в шкале отношений), то расстояние между объектами удобно задать с помощью метрики Минковского:

$$\rho_M(x_i, x_j) = \left(\sum_k |x_{ik} - x_{jk}|^m \right)^{1/m}, \tag{6.1}$$

где $m \geqslant 1$. На практике обычно используется евклидово расстояние $\rho_E(x_i, x_j)$, являющееся частным случаем (6.1) при $m = 2$.

Если все признаки измерены в ранговой шкале, то есть являются некоторым упорядочением по предпочтению, то в качестве функции расстояния может быть использована функция Кемени–Снелла:

$$\rho_K(x_i, x_j) = \frac{1}{2} \sum_m \sum_n |\delta^i_{nm} - \delta^j_{nm}|, \quad \delta^i_{nm} = \begin{cases} -1, & x_{in} > x_{im}; \\ 0, & x_{in} = x_{im}; \\ 1, & x_{in} < x_{im}. \end{cases} \tag{6.2}$$

В случае, когда все признаки объектов измерены в номинальных шкалах, их рассмотрение заменяется дихотомическими переменными, фиксирующими наличие или отсутствие данной градации признака у объекта. В качестве расстояния между точками признакового пространства в этом случае берётся метрика Хемминга:

$$\rho_X(x_i, x_j) = \sum_{k=1}^{l} |x_{ik} - x_{jk}|, \tag{6.3}$$

где l — суммарное число градаций для всех признаков. Нетрудно видеть, что расстояние Хемминга равняется числу несовпадающих признаков у объектов x_i и x_j.

После того, как в признаковом пространстве будет введена функция расстояния, становится очевидным, что наиболее схожие в содержательном смысле объекты будут представлены наиболее близкими точками, и наоборот. Это будет означать, что качественная неоднородность совокупности эмпирических данных будет отражена в геометрическом расположении точек в признаковом пространстве, то есть в структуре многомерных данных. Представление о кластере как некотором подмножестве однородных объектов находит своё выражение в требованиях, предъявляемых к выделенным кластерам. Процедуры кластерного анализа обычно организованы таким образом, что получаемые в результате их выполнения классы разбиений гарантированно обладают требуемыми свойствами.

6.2.2. Измерение однородности объектов

Математически формализация понятия однородности обычно заключается в задании некоторого критерия (функционала) качества разбиения. В этом случае задача кластерного анализа может быть сформулирована как задача оптимизации с точки зрения выбранного критерия разбиения. Выбор функционала качества разбиения (классификационного критерия) отражает представления исследователя о формировании совокупности эмпирических данных и связан с некоторыми предположениями (гипотезами) относительно характера неоднородности расположения точек в признаковом пространстве.

В качестве основных структурных гипотез обычно используются гипотезы компактности и связности.

Гипотеза компактности. Согласно этой гипотезе структура многомерных данных характеризуется наличием компактных кластеров, которые можно заключить в сферические или эллипсоидальные гиперповерхности.

Компактность в данном случае понимается как близость объектов одного класса в признаковом пространстве к некоторому типичному представителю (центру), вокруг которого группируются все остальные объекты. Формирование такой структуры рассматривается как результат воздействия случайных факторов на реализацию эталонных объектов. В признаковом пространстве положение l-го кластера, состоящего из объектов, определяется центром

$$z_l = \frac{1}{n_l} \sum_{m=1}^{n_l} x_m, \tag{6.4}$$

а в качестве меры рассеяния (внутренней однородности) могут выступать величины

$$I_l = \sqrt{\frac{1}{n_l} \sum_{i=1}^{n_l} \rho^2(z_l, x_i)} \tag{6.5}$$

Тогда качество полученного разбиения оценивается функционалом $I = \sum_l I_l$.

Гипотеза связности. Ограничения, накладываемые на форму класса, в некоторых содержательных задачах бывают неоправданно строгими. Более естественными в таких классификационных задачах выглядят требования связности объектов одного класса при одновременной изолированности различных классов. В этом случае наличие отчётливого разрыва в системе признаков, отделяющего кластеры друг от друга, является основой для формулировки классификационного критерия. Оптимальное разбиение характеризуется удалённостью кластеров друг от друга на расстояние, превышающее значение заданного порога. В наиболее распространённом случае расстояние между кластерами P_L и P_M определяется

$$\rho(P_L, P_M) = \min \rho_E(x_l, x_m), \tag{6.6}$$

где $x_l \in P_L$, $x_m \in P_M$ — элементы кластеров P_L и P_M.

Расстояние между множествами можно так же определить с помощью расстояния Хаусдорфа. Пусть P_E и P_F — не пустые компактные подмножества $\mathbf{R}^n$. Тогда расстоянием Хаусдорфа между P_E и P_F будет величина

$$\rho_H(P_E, P_F) = \min(\varepsilon > 0 : P_E \subset P_F + \varepsilon \text{ и } P_F \subset P_E + \varepsilon). \tag{6.7}$$

6.3. Алгоритмы кластерного анализа

Области применения кластерного анализа чрезвычайно разнообразны. Это приводит к тому, что общее число алгоритмов, упоминающихся в литературе по автоматической классификации и кластерному анализу, варьируется от 200 до 500 и с каждым годом это число возрастает. По способу обработки данных *алгоритмы классификации* можно разделить на две большие группы: *иерархические* и *неиерархические*.

Иерархические алгоритмы характеризуются последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров. В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

В качестве примера рассмотрим один из самых простых иерархических алгоритмов кластерного анализа, работающий по принципу выделения единственной связи или «ближайшего соседа».

1. Для инициализации алгоритма в признаковом пространстве задаётся некоторая, например, евклидова метрика (6.1) и согласно (6.6) определяется расстояние между классами P_L и P_M.
2. Выполнение алгоритма начинается с рассмотрения каждого объекта x_i как отдельного класса $P_i = x_i$, где $i = 1, 2, \ldots, n$. Среди совокупности P_i согласно (6.6) определяется пара самых близких классов. Если будет обнаружено несколько таких пар, то выбирается любая из них.
3. Затем пара самых близких классов объединяется $P_i = P_i \cup P_j$, а число классов уменьшается на единицу.
4. К вновь получившейся совокупности классов применяют ту же самую процедуру до тех пор, пока не будет получен тривиальный результат, содержащий один класс.

Характерной особенностью иерархических алгоритмов является графическая форма представления результатов в виде *дендрограммы*, — древовидного графа, позволяющей составить представление о стратифицированной структуре данных.

Преимуществом иерархических методов кластеризации является их наглядность, но при большом количестве наблюдений они трудны в восприятии результатов. В таких случаях более удобными оказываются неиерархические итеративные процедуры.

Неиерархические алгоритмы основаны на дроблении исходной совокупности многомерных данных на определённое число классов. Одним из наиболее распространённых методов кластерного анализа такого типа является метод k-средних.

Входным параметром этого алгоритма является число классов разбиения k, а его выполнение заключается в последовательности следующих действий:

1. Произвольным образом выбираются k исходных центров классов z_{j0} для $j = 1, 2, \ldots, k$. Для этой выборки удобнее всего использовать множество подлежащих классификации объектов x_i при $i = 1, 2, \ldots, n$.
2. Все объекты x_i распределяются по k классам в соответствии с правилом $x_i \in P_n$, если

$$\rho(x_i, z_{n0}) = \min_{1 \leqslant j \leqslant k} \rho(x_i, z_{j0}). \tag{6.8}$$

   То есть объект x_i относится к классу P_n, если расстояние от него до центра класса z_{n0} является наименьшим среди всех возможных.
3. Центры классов пересчитываются

$$z_{p1} = \frac{1}{n_p} \sum_{i=1}^{n_p} x_i, \quad p = 1, 2, \ldots, k. \tag{6.9}$$

4. Выполнение равенств $z_{j1} = z_{j0}$ для $j = 1, 2, \ldots, k$ с заранее выбранной точностью является условием окончания работы алгоритма. При нарушении хотя бы одного из указанных равенств выполняется присваивание $z_{j0} \leftarrow z_{j1}$ и переход к пункту 2.

Задача классификации в общем случае не является математически корректной. Это связано как с неединственностью получаемых решений, так и с отсутствием, в общем случае, устойчивости результатов. Использование отдельных алгоритмов накладывает определённые ограничения, связанные с выбором значений параметров. Например, в некоторых методах надо задать число кластеров как, например, в рассмотренном алгоритме k-средних. В зависимости от значений этих параметров могут быть получены различные результаты разбиения.

Если рассматривать задачу классификации с этой точки зрения, то можно говорить о её некорректности. Это означает, что при использовании только одного метода удовлетворительного результата получить редко удаётся. Иногда приходится использовать несколько алгоритмов, прежде чем будет получен нужный результат, что вызывает определённую сложность. Проблема заключается в том, что

выбор алгоритма заранее определяет форму кластера, что может повлиять на получаемую информацию в ходе исследования, тем самым и на дальнейшее изучение структуры многомерных данных. Создание алгоритма, успешно работающего во всех ситуациях без исключения, представляется трудной задачей. Поэтому целесообразно использовать набор различных алгоритмов, имеющих различные критерии для выделения структур многомерных данных.

При решении классификационной задачи часто не имеет смысла ставить цель найти оптимальный алгоритм для выделения структуры. Более прагматичным выглядит подход, при котором предлагается использовать набор алгоритмов, имеющих различные классификационные критерии, чтобы составить представление о выделяемой структуре данных, количестве классов, их взаимном расположении и центрах классов. Чтобы получить хорошо интерпретированное разбиение, необходимо в каждом случае опираться на содержательные соображения. Описание результатов полученного разбиения делается на основе той теории, которая была использована при выборе признаков этого построения и постановки задачи классификации.

Пример 6.1. Используя частичные выборочные данные по ирисам Фишера $\{(x_1, x_2, x_3, x_4)_i\}$ для $i = 5, 6, \ldots, 149$ провести их автоматическую классификацию методами единственной связи и k-средних.

```
> data(iris)
> smp <- seq(5, 150, 4)
> dat <- scale(iris[smp,-5])
> sps <- as.numeric(iris[smp,5])
> dat[c(1:2, 18:19, 36:37),]
    Sepal.Length Sepal.Width Petal.Length Petal.Width
5     -1.2153779   1.0700351   -1.3824430  -1.2858243
9     -2.0507150  -0.4527071   -1.3824430  -1.2858243
73     0.5945192  -1.3228455    0.5805654   0.2526869
77     1.2906335  -0.6702417    0.5244795   0.1343398
145    1.1514106   0.4174313    1.0292531   1.4361569
149    0.4552964   0.6349659    0.8609952   1.1994629
> pca <- prcomp(dat)[[5]]
> windows(); plot(pca, pch=sps)
> legend("top", pch=unique(sps), legend=paste("i.",levels(iris[,5])))
> fit <- kmeans(dat, centers=3); mps <- fit[[1]]; fit
K-means clustering with 3 clusters of sizes 9, 16, 12
Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1   -0.2717563 -1.08114042    0.1443413  0.01599284
2    0.9686806 -0.05842568    0.9135758  0.93318216
3   -1.0877569  0.88875622   -1.3263570 -1.25623751
```

```
Clustering vector:
   5   9  13  17  21  25  29  33  37  41  45  49  53  57  61  65
   3   3   3   3   3   3   3   3   3   3   3   2   2   1   1
  69  73  77  81  85  89  93  97 101 105 109 113 117 121 125 129
   1   1   2   1   1   1   1   1   2   2   2   2   2   2   2   2
 133 137 141 145 149
   2   2   2   2   2
Within cluster sum of squares by cluster:
[1] 8.382990 9.607702 8.554262
(between_SS / total_SS =  81.6 %)
Available components:
[1] "cluster"    "centers"  "totss"  "withinss"  "tot.withinss"
[6] "betweenss"  "size"     "iter"   "ifault"
> windows(); plot(pca, pch=mps)
> legend("top", pch=unique(mps), legend=paste("Cluster", unique(mps)))
> dst <- dist(dat, method="euclidean")
> fit <- hclust(dst, method="single")
> gps <- cutree(fit, k=3); fit
Call:
hclust(d = dst, method = "single")
Cluster method   : single
Distance         : euclidean
Number of objects: 37
> windows(); plot(fit, cex=.8, ann=FALSE)
> rect.hclust(fit, k=3, which=c(1,3))
> windows(); plot(pca, pch=gps)
> legend("top", pch=unique(gps), legend=paste("Cluster", unique(gps)))
```

Используемая для анализа таблица выборочных данных загружается под именем «iris» в строке 1. Содержание данного набора данных было описано ранее в примере 4.1. Для облегчения визуализации их структуры из исходных выборочных данных в 2–3 формируется подвыборка строк «dat», номера которых «smp» указаны в задании. Функция «scale()» в 3 используется для стандартизации выборочных данных, а вектор «sps» в 4 содержит эталонную классификацию данных, соответствующую их видовой принадлежности. Команда «dat[c(1:2,18:19,36:37),]» отображает по две строки от начала, середины и конца полученной таблицы 5–12.

На рис. 6.1 показана оптимальная визуализация выборочных данных с помощью описанного в главе 4 метода главных компонент. Для её построения в 13–15 были использованы функции: «prcomp()» — для вычисления главных компонент классифицируемых выборочных данных; «windows()» — для создания графического окна; «plot(pca, pch=sps)» — для визуализации выборочных данных в пространстве двух первых главных компонент z_1 и z_2 с использованием эталонной классификации; «legend()» — для описания соответствия символов и

видовой принадлежности выборочных данных.

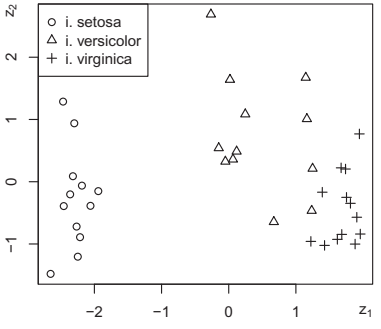


Рис. 6.1. Эталонная классификация выборочных данных по видовой принадлежности

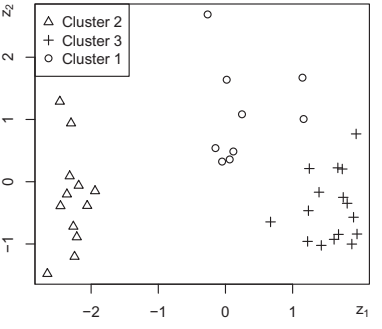


Рис. 6.2. Неиерархическая классификация выборочных данных методом k-средних

В 16 с помощью функции «kmeans()» строится неиерархическая классификация выборочных данных по методу k-средних. Использование параметра «centers=3» соответствует разбиению исходной выборки на три кластера, а вектор «mps» содержит индексы классифицированных данных. Описание выборочных значений признаков, усреднённых по найденным кластерам приведено в 18–22, а сам вектор классификации — в 23–29.

Визуализация кластеров выборочных данных, полученных с помощью метода k-средних, выполняется в 36–37, а результат демонстрируется на рис. 6.2. Сопоставление кластеров на рис. 6.1 и 6.2 показывает, что для используемых выборочных данных метод k-средних обеспечивает вполне адекватный результат, ошибочно классифицировав лишь три точки, лежащие на границе кластеров 1 и 3.

В 38–40 с помощью функции «dist()» с параметром «method="euclidean"» вычисляется матрица евклидовых расстояний между выборочными точками, далее с помощью «hclust()» с параметром «method="single"» строится их иерархическая классификация по методу единственной связи, а с помощью «cutree()» с параметром «k=3» в векторе «gps» сохраняются индексы данных, сгруппированных в три кластера.

Для визуализации иерархической структуры классифицируемых выборочных данных на рис. 6.3 с помощью функций «plot(fit,...)»

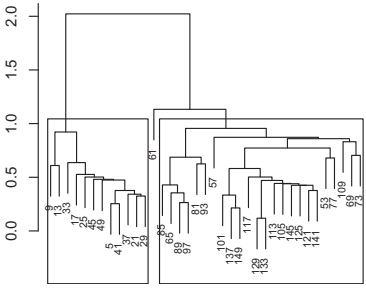


Рис. 6.3. Дендрограмма иерархической классификации методом единственной связи

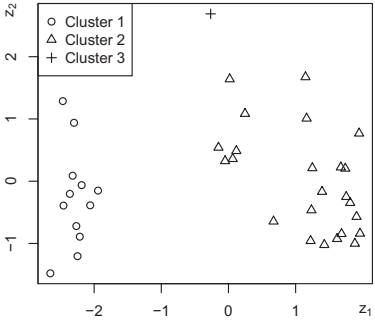


Рис. 6.4. Иерархическая классификация выборочных данных методом единственной связи

и «rect.hclust(fit,...)» в 46–47 построена дендрограмма, на которой выделены два наиболее крупных кластера.

Визуализация результатов иерархической классификации по методу единственной связи в пространстве главных компонент z_1 и z_2 выполняется в 48–49, а результат демонстрируется на рис. 6.4. Сопоставление кластеров на рис. 6.1 и 6.4 показывает, что для используемых выборочных данных метод единственной связи адекватного результата не обеспечивает, ошибочно присоединив к кластеру 2 одиннадцать точек, в действительности относящихся к кластеру 3.

Контрольные вопросы

1. Сформулируйте общую постановку задачи кластерного анализа. Приведите примеры содержательных постановок классификационных задач.
2. В чем суть геометрического подхода к задаче кластерного анализа?
3. Как формализуется понятие сходства между объектами в задачах кластерного анализа?
4. Какие типы алгоритмов кластерного анализа известны в настоящее время? В чем заключается их различие?
5. Укажите классификационные критерии для алгоритмов k-средних и иерархического алгоритма единственной связи. В чем осо-

бенности форм представления результатов выполнения этих алгоритмов?

6. В каких случаях необходима предварительная стандартизация исходных данных классификационной задачи?
7. Какие статистические методы могут быть использованы для подтверждения полученных результатов кластерного анализа?

Литература

1. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. — Vienna, Austria, 2014. — URL: `http://www.r-project.org/` (online; accessed: 24.02.2014).
2. CRAN: The Comprehensive R Archive Network, R Foundation for Statistical Computing. — Vienna, Austria, 2014. — URL: `http://www.cran.r-project.org/` (online; accessed: 24.02.2014).
3. Воеводин В. В., Воеводин Вл. В. Энциклопедия линейной алгебры. Электронная система ЛИНЕАЛ. — СПб.: БХВ–Петербург, 2006. — 544 с.
4. Бородин А. Н. Элементарный курс теории вероятностей и математической статистики. — СПб.: Лань, 2011. — 256 с.
5. Кибзун А. И., Горяинова Е. Р., Наумов А. В. Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами. — М.: ФИЗМАТЛИТ, 2007. — 232 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. Айвазян С. А. и др. Прикладная статистика. Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.
9. Крамер Г. Математические методы статистики. — М.: Мир, 1975. — 648 с.
10. 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 можно найти на сайте проек-

та [1]. Для установки **R** на компьютере с операционной системой семейства Microsoft Windows необходимо загрузить исполняемый файл вида «R-3.1.0-win.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-3.1.0», где имя «R-3.1.0» 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-3.1.0\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-3.1.0\bin» на папку, действительно содержащую текущие файлы пользователя.

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 $k = 2, 3, \ldots$ 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-

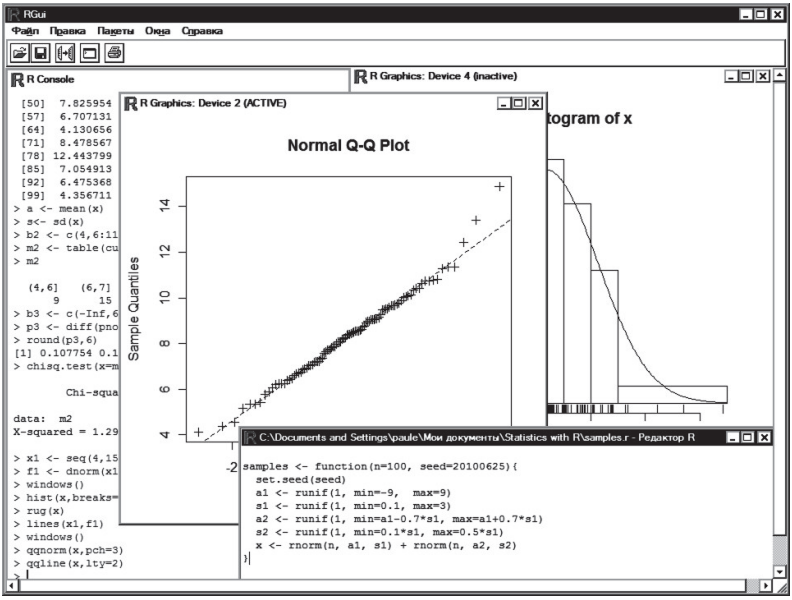


Рис. А.1. Типичное представление рабочей области **R** в операционных системах семейства Microsoft Windows

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** представ-

лена на английском языке. Исключение составляют тексты в строке меню и наиболее распространённые диагностические сообщения об ошибках при исполнении 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")`».

А.2. Основные возможности языка R

В данном разделе на примерах простых выражений приводятся краткие иллюстрации использования для основных команд и функций системы **R**, употребляемых в листингах примеров и приложений.

А.2.1. Скалярные выражения

Для построения выражений в **R** используется стандартный набор операторов «`+`», «`-`», «`*`», «`^`», «`/`», «`%%`» и функций «`sqrt()`», «`exp()`», «`log2()`», «`log()`», «`log10()`», «`sin()`», «`cos()`», «`tan()`», «`asin()`», «`acos()`», «`atan()`», «`abs()`», «`round()`», «`floor()`», «`ceil()`». Для разделения нескольких выражений в одной строке используется символ «`;`». Для отображения результатов вычисления неограниченных и неопределённых выражений в **R** используются служебные слова «`Inf`» и «`NaN`».

При присваивании результатов вычислений скалярных и векторных выражений пользовательским переменным используются двухсимвольные операторы лево- и правосторонних присваиваний «`<-`» и «`->`». Сразу оговоримся, что разделение выражений в **R** на скалярные и векторные абсолютно условно, так как во внутреннем представлении системы любой скаляр является одномерным вектором, а все перечисленные здесь функции и операторы работают с векторами, размерность которых ограничена лишь объёмом доступной оперативной памяти.

«`2-3; 2*3; 2/3`» вычисления арифметических выражений: $2-3=$ $=-1$; $2\cdot 3=6$; $\frac{2}{3}\approx 0.67$;

«`1/0; 0/0; 1/-Inf; Inf/Inf`» арифметические вычисления, приводящие к неограниченным и неопределённым выражениям: $\frac{1}{0}=$ $=\mathrm{Inf}\approx\infty$; $\frac{0}{0}=\mathrm{NaN}$; $\frac{1}{-\infty}=0$; $\frac{\infty}{\infty}=\mathrm{NaN}$;

«`sqrt(3); 2^3; exp(2)`» вычисления значений степенных и показательных функций: $\sqrt{3}\approx 1.73$; $2^3=8$; $e^2\approx 7.39$;

«`log(2); log2(3); log10(2)`» вычисления значений логарифмических функций: $\ln 2\approx 0.69$; $\log_2 3\approx 1.59$; $\lg 2\approx 0.30$;

«`sqrt(-2); log2(0); (-1)^(1/3)`» вычисления функций, приводящие к неопределённым и неограниченным выражениям: $\sqrt{-2}=$ $=\mathrm{NaN}$; $\lg 0=-\mathrm{Inf}\approx-\infty$; $(-1)^{\frac{1}{3}}=\mathrm{NaN}$;

«`sin(pi); cos(2*pi/3); tan(-pi/2)`» вычисления значений тригонометрических функций: $\sin\pi\approx 1.23\cdot 10^{-16}\approx 0$; $\cos(\frac{2\pi}{3})=-0.5$; $\mathrm{tg}(-\frac{\pi}{2})\approx -1.63\cdot 10^{16}\approx-\infty$;

«`asin(1); acos(-1/2); atan(-1)`» вычисления значений обратных тригонометрических функций: $\arcsin 1=\frac{\pi}{2}\approx 1.57$; $\arccos(-\frac{1}{2})=$ $=\frac{2\pi}{3}\approx 2.09$; $\mathrm{arctg}(-1)=-\frac{\pi}{4}\approx -0.79$.

«`abs(-2); 5%%2`» вычисления значений функций модуля и деления по модулю: $|-2|=2$; $5 \bmod 2=1$;

«`round(-1.5); floor(-1.5); ceiling(-1.5)`» округление действительных чисел до ближайшего, меньшего или большего целых: $\lceil -1.5\rceil=-2$; $\lfloor -1.5\rfloor=-2$; $\lceil -1.5\rceil=-1$;

«`5-2*pi->x; y<-sin(x)`» право- и левосторонние присваивания: $x=$ $=5-2\pi\approx -1.28$, $y=\sin(-1.28)\approx -0.96$.

А.2.2. Векторные и матричные выражения

Для формирования векторов используются оператор «`:`» и функции «`c()`», «`seq()`», а для матриц — функции «`matrix()`», «`diag()`», «`cbind()`» и «`rbind()`». Оператор «`%*%`», используемый для матриц, обеспечивает выполнение матричного умножения, а для векторов — скалярного произведения. Функция «`t()`» обеспечивает транспонирование переданной матрицы или вектора. Функция «`det()`» вычисляет определитель квадратной матрицы. Функция «`solve()`», используемая для одиночной квадратной матрицы, выполняет её обращение, а для квадратной матрицы и совместимого с ней вектора — выполняет

решение соответствующей системы линейных алгебраических уравнений.

«`a<-1:3; b<-c(3.4,2,0.7)`» инициализация регулярного и нерегулярного действительных векторов:

$$a = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}; \quad b = \begin{pmatrix} 3.4 \\ 2 \\ 0.7 \end{pmatrix};$$

«`d<-seq(1,2,0.3); e<-seq(9,8,len=3)`» инициализация двух регулярных действительных векторов:

$$d = \begin{pmatrix} 1 \\ 1.3 \\ 1.6 \\ 1.9 \end{pmatrix}; \quad e = \begin{pmatrix} 9 \\ 8.5 \\ 8 \end{pmatrix};$$

«`f<-cos(e-b*a/2)^(-1/3)`» построение вектора «`f`» в виде функции векторов «`a`», «`b`» и «`e`»:

$$f = \cos^{-1/3}\begin{pmatrix} 9 - 3.4 \cdot \frac{1}{2} \cdot a \\ 8.5 - 2 \cdot \frac{1}{2} \cdot b \\ 8 - 0.7 \cdot \frac{1}{2} \cdot e \end{pmatrix} \approx \begin{pmatrix} 1.24 \\ 1.01 \\ 1.08 \end{pmatrix};$$

«`G<-matrix(c(a,-sqrt(b),e/3),nrow=3)`» построение матрицы «`g`» как функции векторов «`a`», «`b`» и «`e`»:

$$G = \begin{pmatrix} 1 & -\sqrt{3.4} & 9/3 \\ 2 & -\sqrt{2} & 8.5/3 \\ 3 & -\sqrt{0.7} & 8/3 \end{pmatrix} \approx \begin{pmatrix} 1 & -1.84 & 3 \\ 2 & -1.41 & 2.83 \\ 3 & -0.84 & 2.67 \end{pmatrix};$$

«`G+f`» вычисление суммы элементов столбцов матрицы «`G`» и вектора «`f`»:

$$\begin{pmatrix} 1+1.24 & -1.84+1.24 & 3+1.24 \\ 2+1.01 & -1.41+1.01 & 2.83+1.01 \\ 3+1.08 & -0.84+1.08 & 2.67+1.08 \end{pmatrix} \approx \begin{pmatrix} 2.24 & -0.61 & 4.24 \\ 3.01 & -0.41 & 3.84 \\ 4.08 & 0.25 & 3.75 \end{pmatrix};$$

«`G*f`» вычисление произведения элементов столбцов матрицы «`G`» и вектора «`f`»:

$$\begin{pmatrix} 1\cdot 1.24 & -1.84\cdot 1.24 & 3\cdot 1.24 \\ 2\cdot 1.01 & -1.41\cdot 1.01 & 2.83\cdot 1.01 \\ 3\cdot 1.08 & -0.84\cdot 1.08 & 2.67\cdot 1.08 \end{pmatrix} \approx \begin{pmatrix} 1.24 & -2.28 & 3.71 \\ 2.01 & -1.43 & 2.86 \\ 3.25 & -0.91 & 2.89 \end{pmatrix};$$

«`G%*%f`» вычисление произведения матрицы «`G`» на вектор «`f`»:

$$\begin{pmatrix} 1 & -1.84 & 3 \\ 2 & -1.41 & 2.83 \\ 3 & -0.84 & 2.67 \end{pmatrix} \cdot \begin{pmatrix} 1.24 \\ 1.01 \\ 1.08 \end{pmatrix} \approx \begin{pmatrix} 2.63 \\ 4.12 \\ 5.76 \end{pmatrix};$$

«`det(G)`» вычисление определителя матрицы «`G`»:

$$\det G \approx \begin{vmatrix} 1 & -1.86 & 3 \\ 2 & -1.41 & 2.83 \\ 3 & -0.84 & 2.67 \end{vmatrix} \approx 0.47;$$

«`solve(G)`» вычисление матрицы, обратной к матрице «`G`»:

$$G^{-1} \approx \begin{pmatrix} 1 & -1.84 & 3 \\ 2 & -1.41 & 2.83 \\ 3 & -0.84 & 2.67 \end{pmatrix}^{-1} \approx \begin{pmatrix} -2.99 & 5.14 & -2.1 \\ 6.76 & -13.53 & 6.76 \\ 5.49 & -10.03 & 4.86 \end{pmatrix};$$

«`solve(G,f)`» решение системы линейных уравнений, определяемой элементами матрицы «`G`» и вектора «`f`»:

$$\begin{pmatrix} 1 & -1.84 & 3 \\ 2 & -1.41 & 2.83 \\ 3 & -0.84 & 2.67 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.24 \\ 1.01 \\ 1.08 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} \approx \begin{pmatrix} -0.8 \\ 2.07 \\ 1.95 \end{pmatrix};$$

«`diag(3); diag(1:3)`» построение единичной матрицы третьего порядка и диагональной матрицы, след которой соответствует заданному вектору $(1, 2, 3)$:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix};$$

«`cbind(G,f); t(rbind(G,f))`» объединение матрицы «`G`» и вектора «`f`» вдоль строк и вдоль столбцов с транспонированием последнего варианта:

$$\begin{pmatrix} 1 & -1.84 & 3 & 1.24 \\ 2 & -1.41 & 2.83 & 1.01 \\ 3 & -0.84 & 2.67 & 1.08 \end{pmatrix}; \quad \begin{pmatrix} 1 & 2 & 3 & 1.24 \\ -1.84 & -1.41 & -0.84 & 1.01 \\ 3 & 2.83 & 2.67 & 1.08 \end{pmatrix};$$

А.2.3. Индексы и выборочные выражения

Язык **R** предоставляет широкие возможности по выборочной обработке элементов векторов и матриц. При этом пользователь может указывать как конкретные номера элементов, строк и/или столбцов, так и их диапазоны, используя оператор «:». Отрицательные индексы используются для исключения определённых элементов, строк и/или столбцов, а логические выражения — для выборки всех элементов, отвечающих заданному условию.

«`G[1,]; G[,3]; G[1:2,2:3]`» выборки матрицы «`G`» с включением: элементов первой строки, третьего столбца, а также элементов строк с первой по вторую и столбцов со второго по третий:

$$\begin{pmatrix} 1 \\ -1.84 \\ 3 \end{pmatrix}; \quad \begin{pmatrix} 3 \\ 2.83 \\ 2.67 \end{pmatrix}; \quad \begin{pmatrix} -1.84 & 3 \\ -1.41 & 2.83 \end{pmatrix};$$

«`f[-3]; G[-3,-1]`» выборки вектора «`f`» и матрицы «`G`» с исключением: третьего элемента вектора «`f`», а также третьей строки и первого столбца матрицы «`G`»:

$$\begin{pmatrix} 0.55 \\ 0.77 \end{pmatrix}; \qquad \begin{pmatrix} -1.84 & 3 \\ -1.41 & 2.83 \end{pmatrix};$$

«`f[f>0.1]; G[G<0]`» выборка всех элементов вектора «`f`», значение которых превышает 0.1, а также всех отрицательных элементов матрицы «`G`»:

$$\begin{pmatrix} 0.55 \\ 0.77 \end{pmatrix}; \qquad \begin{pmatrix} -1.84 \\ -1.41 \\ -0.84 \end{pmatrix}.$$

А.2.4. Управляющие структуры

Управляющие структуры в **R** включают в себя операторы и функции, используемые для выборочного изменения порядка выполнения в последовательности команд в зависимости от выполнения или невыполнения заданных условий.

Оператор «if(){}else{}» используется для выбора между альтернативными группами команд в зависимости от логического скаляра. Фигурные скобки «{}» здесь используются для объединения нескольких выражений в составное. Для получения вектора значений с помощью логического вектора используется функция «ifelse()».

Для повторения группы команд некоторое, заранее известное число раз, используется оператор «for(){}», а для повторения группы команд заранее неизвестное число раз — оператор «while(){}». В первом случае аргументом оператора будет переменная, последовательно принимающая значения компонент некоторого вектора, а во втором — условие продолжения итераций. Исчерпание вектора или нарушение условия продолжения означают выход из цикла. С помощью команды «repeat{}» можно задать цикл с потенциально неограниченным числом итераций, которые могут быть прекращены только с помощью команды «if() break». Та же команда используется для получения досрочного выхода из цикла.

В качестве альтернативы циклам можно рассматривать семейство функций «apply()», используемых для последовательного применения определённой функции к выбранным компонентам вектора или матрицы.

«x<-if(b[i]<1) 1 else 0» результат условного выражения равен 1, если условие «b[i]<1» истинно, или 0 в противном случае:

$$\begin{array}{lll} \text{для} \quad i=1: & b_1 = 3.4 \not< 1 & \Rightarrow \quad x = 0; \\ \text{для} \quad i=2: & b_2 = 2 \not< 1 & \Rightarrow \quad x = 0; \\ \text{для} \quad i=3: & b_3 = 0.7 < 1 & \Rightarrow \quad x = 1; \end{array}$$

«x<-ifelse(b<1,1,0)» элементы вектора «x» принимают значения 1, если соответствующие элементы логического вектора «b<1» истинны, или 0, если соответствующие элементы «b<1» ложны:

$$b = \begin{pmatrix} 3.4 \\ 2 \\ 0.7 \end{pmatrix}; \quad b < 1 = \begin{pmatrix} \text{FALSE} \\ \text{FALSE} \\ \text{TRUE} \end{pmatrix}; \quad x = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix};$$

«for(i in seq(along=b)) print(sum(b[1:i]))» в цикле по элементам вектора «b» на экран выводятся значения сумм всех элементов «b» от первого до текущего элемента включительно: 3.4; 5.4; 6.1;

«`g<-0; while(g<1) print(g<-rnorm(1))`» текущие значения в цикле с предусловием выводятся на экран до тех пор, пока не встретится значение, большее или равное единицы; например: 0.45; 0.34; -0.78; 1.19;

«`repeat{print(g<-rnorm(1)); if(g>1)break}`» текущие значения в цикле с постусловием выводятся на экран до тех пор, пока не встретится значение, строго большее единицы; например: 0.5; 0.26; -0.25; 1.26;

«`apply(G,1,mean); apply(G,2,mean)`» применение этих функций к матрице G формирует векторы значений, усреднённых по строкам $\langle G\rangle_r$ или по столбцам $\langle G\rangle_c$:

$$\langle G\rangle_r \approx \begin{pmatrix} 0.72 \\ 1.14 \\ 1.61 \end{pmatrix}; \quad \langle G\rangle_c^{\mathsf{T}} \approx \begin{pmatrix} 2 \\ -1.36 \\ 2.83 \end{pmatrix};$$

А.2.5. Функции ввода и вывода

Функции ввода и вывода обеспечивают пользователю взаимодействие с текстовой и графической консолью, а также с накопителями данных. Функции «`cat()`» и «`print()`» позволяют осуществлять вывод информации о работе программы на текстовую, а «`plot()`» — на текущую графическую консоль. Чтение и запись векторных, матричных или табличных данных из/в текстовые файлы, расположенные в текущей рабочей папке, осуществляются с помощью функций «`write.table()`» и «`read.table()`».

«`cat("Номер итерации: i<-i+1,"n")`» выводит в текущую позицию текстовой консоли комментарий, значение переменной «`i`» и завершает строку; при повторном вызове этой команды значение «`i`» будет увеличиваться на единицу;

«`print(G,digits=2); print(G,digits=1)`» начиная с новой строки, выводит на текстовую консоль значение матрицы «`G`» с двумя или с одной значащими цифрами;

«`x<-seq(0,1,0.01); plot(x=x,y=exp(-x*x/2),type="l")`» строит с помощью ломаной график функции $e^{-x^2/2}$, используя текущую графическую консоль; интервалы по оси абсцисс и по оси ординат определяются по размахам значений «`x`» и «`y`»;

«`write.table(tbl, file="data.tbl")`» записывает таблицу данных «`tbl`» в текстовый файл «`data.tbl`», используя в качестве разделителя столбцов символ «` `» и в качестве десятичного разделителя — символ «`.`»;

«`tbl<-read.table(file="data.tbl", sep=" ")`» считывает таблицу данных «`tbl`» из текстового файла «`data.tbl`», используя в качестве разделителя столбцов символ «` `» и в качестве десятичного разделителя — символ «`.`»;

А.2.6. Пользовательские функции

Система **R** содержит реализации многих тысяч базовых функций. Для расширения возможностей системы можно воспользоваться дополнительными библиотеками или написать собственные функции. Определение собственной функции включает имя функции, оператор присваивания и служебное слово «`function`»: «`f<-function(){}`», где «`f`» — имя новой функции; в круглых скобках «`()`» указаны её аргументы, а в фигурных «`{}`» — определяющая функцию последовательность выражений. Если определение функции состоит из одного выражения, то фигурные скобки могут не использоваться.

«`x<-function(h,a=0,b=1) seq(a,b,h)`» пользовательская функция «`x(h,a,b)`» для заданной тройки аргументов h, a, b возвращает вектор вида $(a, a+h, \ldots, b)$;

«`f<-function(x) exp(-x^2/2)`» пользовательская функция «`f(x)`» для вектора аргумента $x = (x_1, x_2, \ldots, x_n)$ возвращает вектор значений $f = (e^{-x_1^2/2}, e^{-x_2^2/2}, \ldots, e^{-x_n^2/2})$;

Приложение Б

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, а затем 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. Путь к текущей рабочей папке можно увидеть с помощью команды «`getwd()`», а проверить её содержимое — командой «`dir()`». Используемая в некоторых листингах функция «`windows()`» является платформозависимой. Пользователям операционных систем семейств GNU/Linux и Apple Mac OS вместо функции «`windows()`» следует использовать «`x11()`».

Б.1. Визуализация законов распределения

Для построения графиков различных законов распределения дискретных и непрерывных случайных величин используются функции из файла «`plot2graph.r`», листинг которого показан ниже. Поэтому, при построении всех графиков из главы 2 файл «`plot2graph.r`» должен находиться в текущей рабочей папке.

Листинг файла «`plot2graph.r`»

```
# Аргументы функций:
# x, y - значения абсцисс и ординат графиков;
# xl, yl - подписи осей абсцисс и ординат графиков;
# lt, lc - подписи и цвета графиков.
# # # # # # # # # # # # # # # # # # # #
# Графики вероятностей дискретных с.в.
# # # # # # # # # # # # # # # # # # # # #
pmfPlot <- function(x, y, xl="k", yl=expression(p[k]),
                    lt="", lc=rainbow(ncol(y))) {
  matplot(x, y, type="o", col=lc, pch=16, lty=1,
          lwd=2, xlab=xl, ylab=yl)
```

```
  abline(h=0, lty=2)
  legend("topright", col=lc, pch=16, lty=1, lwd=2, legend=lt)
}
# Графики функций распределения дискретных с.в.
# # # # # # # # # # # # # # # # # # # # # # # #
cmfPlot <- function(x, y, xl="x", yl="F(x)",
                    lt="", lc=rainbow(ncol(y))) {
  plot(x, y[,1], type="n", xlab=xl, ylab=yl,
       xlim=range(x), ylim=c(0,1))
  for(k in seq(ncol(y)))
    lines(stepfun(x, c(0, y[,k])), col=lc[k], pch=16, lwd=2)
  abline(h=c(0,1), lty=2)
  legend("bottomright", col=lc, pch=16, lty=1, lwd=2, legend=lt)
}
# Графики плотностей вероятностей непрерывных с.в.
# # # # # # # # # # # # # # # # # # # # # # # #
pdfPlot <- function(x, y, xl="x", yl="f(x)",
                    lt="", lc=rainbow(ncol(y))) {
  matplot(x, y, col=lc, type="l", lty=1,
          lwd=2, xlab=xl, ylab=yl)
  abline(h=0, lty=2)
  legend("topright", col=lc, lty=1, lwd=2, legend=lt)
}
# Графики функций распределения непрерывных с.в.
# # # # # # # # # # # # # # # # # # # # # # # #
cdfPlot <- function(x, y, xl="x", yl="F(x)",
                    lt="", lc=rainbow(ncol(y))) {
  matplot(x, y, col=lc, type="l", lty=1,
          lwd=2, xlab=xl, ylab=yl)
  abline(h=c(0,1), lty=2)
  legend("bottomright", col=lc, lty=1, lwd=2, legend=lt)
}
```

Б.2. Методы оценивания и проверки гипотез

При решении большинства задач из главы 3 используется одноимённая функция «samples()», листинг которой показан ниже. Поэтому, при запуске программ из данного раздела файл «samples.r» должен находиться в текущей рабочей папке.

Листинг файла «samples.r»

```
# Генерирование реализации случайной выборки
# с законом распределения, близким к нормальному
# # # # # # # # # # # # # # # # # # # # # # # #
```

```
# Аргументы и переменные функции:
# n, x - объём и реализация случайной выборки;
# seed - начальное состояние генератора п.с.ч.;
# dgts - число десятичных знаков при округлении x;
# a1, a2 - математическое ожидание с.в.;
# s1, s2 - среднеквадратическое отклонение с.в.
# # # # # # # # # # # # # # # # # # # # # # # # #
samples <- function(n=100, seed=NA, dgts=2) {
  if (!is.na(seed)) set.seed(seed)
  a1 <- runif(1, min=-9, max=9)
  s1 <- runif(1, min=.1, max=3)
  a2 <- runif(1, min=a1-.3*s1, max=a1+.3*s1)
  s2 <- runif(1, min=.1*s1, max=.3*s1)
  if (is.na(dgts))
    return(rnorm(n, a1, s1) + rnorm(n, a2, s2))
  else
    return(round(rnorm(n, a1, s1) +
                 rnorm(n, a2, s2), digits=dgts))
}
```

Основные выборочные характеристики

```
# Вычисление и визуализация основных характеристик реализации
# случайной выборки с законом распределения, близким к нормальному
# # # # # # # # # # # # # # # # # # # # # # # # # # #
# Используемые переменные:
# n, x - объём и реализация случайной выборки;
# a1, s1 - выборочные среднее и среднеквадратическое отклонения;
# x1 - регулярный вектор абсцисс в интервале [a1-4*s1, a1+4*s1];
# f1, F1 - плотность вероятности и функция распределения с.в.
# # # # # # # # # # # # #
source("samples.r")
n <- 100; print(x <- samples(n))
print(c(a1 <- mean(x), s1 <- sd(x))); print(quantile(x))
x1 <- seq(a1-4*s1, a1+4*s1, len=n)
f1 <- dnorm(x1, a1, s1); F1 <- pnorm(x1, a1, s1)
lt <- sprintf("X ~ N(%.2f, %.2f)", a1, s1)
windows(); hist(x, breaks="Scott", freq=FALSE, xlab="x", ylab="f(x)")
rug(x); lines(x1, f1); abline(h=0, lty=2)
legend("topleft", lty=1, legend=paste("f(x):",lt))
windows(); plot(ecdf(x), pch=".", xlab="x", ylab="F(x)")
rug(x); lines(x1, F1)
legend("topleft", lty=1, legend=paste("F(x):",lt))
```

Интервальные оценки параметров распределения

```
# Построение реализаций доверительных интервалов для
# параметров нормально распределённой случайной величины
```

```
# # # # # # # # # # # # # # # # # # # # #
# Используемые переменные:
# n, x - объём и реализация случайной выборки; g - доверительная
# вероятность; ci{E,D}{n,g} - границы доверительных интервалов для EX
# или DX при постоянной доверительной вероятности или объёме выборки.
# # # # # # # # # # # # # # # # # # # # #
set.seed(20100625)
n <- seq(100, 1000, 20); g <- seq(0.95, 0.995,,length(n))
ciE <- function(x,n,g) mean(x)-sd(x)/sqrt(n)*qt((1+c(g,0,-g))/2,n-1)
ciD <- function(x,n,g) sd(x)^2*(n-1)/qchisq((1+c(g,0,-g))/2,n-1)
ciEn <- sapply(n, function(nn) ciE(rnorm(nn), nn, g[1]))
ciDn <- sapply(n, function(nn) ciD(rnorm(nn), nn, g[1]))
ciEg <- sapply(g, function(gg) ciE(rnorm(n[1]), n[1], gg))
ciDg <- sapply(g, function(gg) ciD(rnorm(n[1]), n[1], gg))
textEn <- parse(text=sprintf("EX*(list(n,gamma==%.3g))",g[1]))
textDn <- parse(text=sprintf("DX*(list(n,gamma==%.3g))",g[1]))
textEg <- parse(text=sprintf("EX*(list(gamma,n==%.0f))",n[1]))
textDg <- parse(text=sprintf("DX*(list(gamma,n==%.0f))",n[1]))
ci_graph <- function(x, y, point, text, xlb, ylb) { windows()
  plot(range(x), range(y), type="n", xlab=xlb, ylab=ylb)
  for(j in seq(length(x))) {
    lines(rep(x[j],3), y[,j])
    points(x[j], y[2,j], pch=20) }
  legend("topright", lty=1, pch=20, legend=text, bg="white")
  abline(h=point, lty=2) }
ci_graph(n, ciEn, 0, textEn, "n", "EX")
ci_graph(n, ciDn, 1, textDn, "n", "DX")
ci_graph(g, ciEg, 0, textEg, expression(gamma), "EX")
ci_graph(g, ciDg, 1, textDg, expression(gamma), "DX")
```

Пирсона χ^2-критерий согласия

```
# Проверка гипотезы о нормальности реализации выборочной
# совокупности по критерию согласия Пирсона хи-квадрат
# # # # # # # # # # # # # # # # # # # # #
# Используемые переменные:
# x - реализация случайной выборки; a, s - выборочные среднее и
# среднеквадратическое отклонения; b1, b2, b3 - векторы граничных
# точек для интервалов группировки; m1, m2 - распределение частот
# выборки по интервалам; p3 - распределение вероятностей значений
# случайной величины по интервалам группировки.
# # # # # # # # # # # # # # # # # # # # #
source("samples.r")
x <- samples(seed=20100625); print(c(a <- mean(x), s <- sd(x)))
b1 <- c(4:15); print(m1 <- table(cut(x, breaks=b1)))
b2 <- c(4, 6:11, 15); print(m2 <- table(cut(x, breaks=b2)))
b3 <- c(-Inf, 6:11, Inf); p3 <- diff(pnorm(b3, a, s))
print(chisq.test(x=m2, p=p3))
x1 <- seq(4, 15, length=300); f1 <- dnorm(x1, a, s)
```

```
hist(x, breaks=b2); rug(x); lines(x1, f1)
windows(); qqnorm(x, pch=3); qqline(x, lty=2)
```

Критерий согласия Колмогорова

```
# Зависимость достигаемого уровня значимости от числа степеней
# свободы для центрированной реализации выборки по критерию согласия
# Колмогорова для нулевой гипотезы о соответствии выборочного
# и t-распределения с тем же числом степеней свободы.
# # # # # # # # # # # # # # # # # # # # #
# Используемые переменные:
# x, v - исходная и центрированная реализации случайной выборки;
# w - центрированный вариационный ряд; m - число степеней свободы
# t-распределения; alpha - достигаемый уровень значимости при
# проверке нулевой гипотезы по критерию согласия Колмогорова.
# # # # # # # # # # # # # # # # # # # # #
source("samples.r")
x <- samples(seed=20100625, dgts=NA)
m <- 2:20; w <- sort(v <- x - mean(x))
alpha <- sapply(m, function(k) ks.test(v, "pt", k)[[2]])
plot(m, alpha, type="b", pch=20, ylim=c(0, max(alpha, 0.05)))
abline(h=0.05, lty=2)
windows(); plot(ecdf(v), pch=".", xlab="v", ylab="F(v)")
rug(v); lines(w, pt(w, m[which.max(alpha)]))
```

Критерий однородности Смирнова

```
# Проверка гипотезы об однородности распределений для двух частей
# реализации случайной выборки по критерию Смирнова
# # # # # # # # # # # # # # # # # # # #
# Используемые переменные:
# x, u, v - исходная реализация случайной выборки и две её части.
# # # # # # # # # # # # # # # # # # # #
source("samples.r")
x <- samples(seed=20100625, dgts=NA)
u <- x[seq(1,99,2)]; v <- x[seq(2,100,2)]; print(ks.test(u,v))
plot(ecdf(u), pch=25, xlim=range(x), xlab="u,v", ylab="F(u),F(v)")
plot(ecdf(v), pch=24, add=TRUE); rug(u, side=1); rug(v, side=3)
```

Одновыборочный t-критерий значимости различий

```
# Зависимость p-уровня от параметра a для реализации случайной выборки
# по одновыборочному t-критерию для гипотезы H_0: EX = a:
# # # # # # # # # # # # # # # # # # # # #
# Используемые перерменные:
# x - реализация случайной выборки; a, s - выборочные среднее и
```

```
# среднеквадратическое отклонения; p, pl, pg - p-уровни
# при гипотезах H_1: а) EX != a; б) EX < a; в) EX > a.
# # # # # # # # # # # # # # # # # # # # # # # # # # # # #
source("samples.r"); x <- samples(seed=20100625)
a <- mean(x); s <- sd(x); a <- seq(a-s/2, a+s/2, length=99)
p <- sapply(a, function(aa) t.test(x, mu=aa, alter="two")[[3]])
pl <- sapply(a, function(aa) t.test(x, mu=aa, alter="le")[[3]])
pg <- sapply(a, function(aa) t.test(x, mu=aa, alter="gr")[[3]])
matplot(a, cbind(p, pl, pg), type="l", lty=c(1, 2, 4))
abline(h=c(0, 0.05), lty=3)
```

Двухвыборочный t-критерий значимости различий

```
# Проверка гипотезы H_0: EU = EV по двухвыборочному t-критерию при
# гипотезах H_1: а) EU != EV; б) EU < EV; в) EU > EV.
# # # # # # # # # # # # # # # # # # #
# Используемые переменные:
# x, u, v - исходная реализация случайной выборки и две её части.
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
source("samples.r"); x <- samples(seed=20100625)
u <- x[seq(1,99,2)]; v <- x[seq(2,100,2)]
print(t.test(u,v, var.equal=TRUE, alter="two"))
print(t.test(u,v, var.equal=TRUE, alter="le"))
print(t.test(u,v, var.equal=TRUE, alter="gr"))
```

Фишера F-критерий значимости различий

```
# Проверка гипотезы H_0: DU = DV по двухвыборочному F-критерию при
# гипотезах H_1: а) DU != DV; б) DU < DV; в) DU > DV.
# # # # # # # # # # # # # # # # # # # # #
# Используемые переменные:
# x, u, v - исходная реализация случайной выборки и две её части.
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
source("samples.r"); x <- samples(seed=20100625)
u <- x[seq(1,99,2)]; v <- x[seq(2,100,2)]
print(var.test(u,v, alter="two"))
print(var.test(u,v, alter="le"))
print(var.test(u,v, alter="gr"))
```

Однофакторный дисперсионный анализ

```
# Проверка значимости влияния изменений качественного признака
# на величину количественного признака с помощью методики
# однофакторного дисперсионного анализа.
# # # # # # # # # # # # # # # # # #
# Используемые переменные:
```

```
# D, B, S - значения количественного признака, наблюдаемые
# на каждом из уровней качественного признака;
# adhf - вспомогательная таблица.
# # # # # # # # # # # # # # # # # # # # # #
D = c(4.0, 4.5, 4.3, 5.6, 4.9, 5.4, 3.8, 3.7, 4.0)
B = c(4.5, 4.9, 5.0, 5.7, 5.5, 5.6, 4.7, 4.5, 4.7)
S = c(5.4, 4.9, 5.6, 5.8, 6.1, 6.3, 5.5, 5.0, 5.0)
adhf = stack(data.frame(D, B, S))
print(anova(lm(values ~ ind, data=adhf)))
```

Б.3. Метод главных компонент

Визуализация главных компонент

```
# Выбор и визуализация на плоскости двух первых главных
# компонент для "ирисов Фишера".
# # # # # # # # # # # # # # #
# Используемые переменные:
# iris - таблица выборочных данных для "ирисов Фишера"; x - матрица
# количественных признаков для "ирисов Фишера"; dots, spec - векторы
# значений и индексов качественного признака для "ирисов Фишера";
# pca - главные компоненты для матрицы количественных признаков x;
# Dz - относительные доли дисперсии главных компонент.
# # # # # # # # # # # # # # #
data(iris)
pairs(x <- iris[-5], pch=dots <- as.numeric(iris[,5]))
print(summary(pca <- prcomp(x, center=TRUE, scale=TRUE)))
spec <- levels(iris[,5])
windows(); plot(pca[[5]], pch=dots)
legend("topright", pch=seq(3), legend=paste("i.",spec))
Dz <- pca[[1]]^2/sum(pca[[1]]^2)
windows();
barplot(cumsum(Dz), col="gray67", names.arg=paste("z",seq(4)))
barplot(Dz, col="gray50", axes=FALSE, add=TRUE); box()
windows(); pairs(pca[[5]], pch=dots)
```

Б.4. Начала регрессионного анализа

Парная линейная регрессия

```
# Построение линейной модели парной регрессии
# для заданных выборочных векторов x, y и
# проверка качества этой модели.
# # # # # # # # # # # # # # #
# Используемые переменные:
```

```
# x, y - векторы выборочных данных;
# fit - линейная модель парной регрессии;
# xin - регулярный вектор абсцисс;
# pre - матрица ординат для центров и границ для
# 0,95-доверительных интервалов уравнения регрессии.
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
x <- c(1.89, 2.21, 2.37, 2.91, 2.72, 3.55, 3.84, 4.13, 4.25, 4.88)
y <- c(0.75, 0.59, 0.19, 0.02, 0.04,-0.72,-0.85,-1.35,-1.36,-1.83)
print(summary(fit <- lm(y ~ x)))
xin <- seq(0.8*min(x), 1.2*max(x), length=100)
pre <- predict(fit, data.frame(x=xin), interval="confidence")
plot(dat, pch=16)
matplot(xin, pre, type="l", lty=c(1,2,2), add=TRUE)
windows(); par(mfrow=c(2,1))
plot(x, fit[[2]], pch=4); abline(h=0)
qqnorm(fit[[2]], pch=8); qqline(as.vector(fit[[2]]))
```

Множественная линейная регрессия

```
# Построение линейной модели множественной регрессии
# для заданных выборочных векторов x1, x2, y и
# проверка качества этой модели.
# # # # # # # # # # # # # # #
# Используемые переменные:
# x1, x2, y - векторы выборочных данных;
# fit - линейная модель множественной регрессии;
# x1m, x1x, x2m, x2x - наименьшие и наибольшие
# выборочные значения абсцисс и ординат.
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
x1<- c(-9.08,-8.07,-7.75,-6.89,-7.01,-6.43,-6.02,-5.28,-4.71,-4.33)
x2<- c( 2.37, 2.05, 3.45, 3.22, 3.97, 4.18, 5.74, 5.67, 6.00, 6.65)
y <- c(-1.80,-2.61,-0.41,-1.61,-0.12,-0.04, 2.90, 1.56, 1.52, 2.55)
print(summary(fit <- lm(y ~ x1 + x2)))
require(plot3D)
scatter3D(x1, x2, y, pch=16, xlab="x1", ylab="x2", zlab="y")
segments3D(x1, x2, rep(0, length(x1)), x1, x2, y, add=TRUE)
x1m <- min(x1); x1x <- max(x1)
x2m <- min(x2); x2x <- max(x2)
polygon3D(c(x1m, x1m, x1x, x1x), c(x2m, x2x, x2x, x2m),
          rep(0, 4), border="black", facets=NA, add=TRUE)
windows(); par(mfrow=c(2,1))
plot(x1, y); segments(x1, 0, x1, y); abline(h=0)
plot(x2, y); segments(x2, 0, x2, y); abline(h=0)
windows(); par(mfrow=c(2,1))
termplot(fit, se=TRUE, partial.resid=TRUE)
windows(); par(mfrow=c(3,1))
plot(x1, fit$res, pch=4); abline(h=0, lty=1)
plot(x2, fit$res, pch=4); abline(h=0, lty=1)
qqnorm(fit$res, pch=8); qqline(as.vector(fit$res))
```

Б.5. Основы кластерного анализа

Автоматическая классификация данных

```
# Автоматическая классификация выборочных
# данных построенных на основе "ирисов Фишера"
# методами единственной связи и k-средних.
# # # # # # # # # # # # # # # # # #
# Используемые переменные:
# iris - таблица выборочных данных для "ирисов Фишера";
# smp - индексы строк для частичных выборочных данных;
# dat - матрица количественных признаков для "ирисов Фишера";
# sps - вектор индексов качественного признака для "ирисов Фишера";
# pca - главные компоненты для матрицы количественных признаков dat;
# fit - модели классификации методами единственной связи и k-средних;
# mps - вектор индексов для классификации данных методом k-средних;
# dst - вектор расстояний между выборочными точками; gps - вектор
# индексов для классификации данных методом единственной связи.
# # # # # # # # # # # # # # # # #
data(iris); smp <- seq(5,150,4)
dat <- scale(iris[smp,-5])
sps <- as.numeric(iris[smp,5])
plot(pca <- prcomp(dat)[[5]], pch=sps)
legend("top", pch=unique(sps),
       legend=paste("i.",levels(iris[,5])))
print(fit <- kmeans(dat, centers=3)); mps <- fit[[1]]
windows(); plot(pca, pch=mps)
legend("top", pch=unique(mps),
       legend=paste("Cluster", unique(mps)))
dst <- dist(dat, method="euclidean")
print(fit <- hclust(dst, method="single"))
gps <- cutree(fit, k=3)
windows(); plot(fit, cex=.8, ann=FALSE)
rect.hclust(fit, k=3, which=c(1,3))
windows(); plot(pca, pch=gps)
legend("top", pch=unique(gps),
       legend=paste("Cluster", unique(gps)))
```

Алексей Георгиевич БУХОВЕЦ,
Павел Валентинович МОСКАЛЕВ

АЛГОРИТМЫ ВЫЧИСЛИТЕЛЬНОЙ СТАТИСТИКИ В СИСТЕМЕ R

Учебное пособие

Издание второе,
переработанное и дополненное

Зав. редакцией физико-математической
литературы *Н. Р. Нигмадзянова*
Выпускающий *О. В. Шилкова*

ЛР № 065466 от 21.10.97
Гигиенический сертификат 78.01.07.953.П.007216.04.10
от 21.04.2010 г., выдан ЦГСЭН в СПб

Издательство «ЛАНЬ»
lan@lanbook.ru; www.lanbook.com
192029, Санкт-Петербург, Общественный пер., 5.
Тел./факс: (812) 412-29-35, 412-05-97, 412-92-72.
Бесплатный звонок по России: 8-800-700-40-71

Подписано в печать 15.08.14.
Бумага офсетная. Гарнитура Школьная. Формат 84×108 $^{1}/_{32}$.
Печать офсетная. Усл. п. л. 8,40. Тираж 700 экз.

Заказ № .

Отпечатано в полном соответствии
с качеством предоставленных материалов
в ГУП ЧР «ИПК “Чувашия”».
428019, г. Чебоксары, пр. И. Яковлева, д. 13.
Тел.: (8352) 56-00-23

ГДЕ КУПИТЬ

ДЛЯ ОРГАНИЗАЦИЙ:

Для того, чтобы заказать необходимые Вам книги, достаточно обратиться в любую из торговых компаний Издательского Дома «ЛАНЬ»:

по России и зарубежью

«ЛАНЬ-ТРЕЙД»
192029, Санкт-Петербург, ул. Крупской, 13
тел.: (812) 412-85-78, 412-14-45, 412-85-82
тел./факс: (812) 412-54-93
e-mail: trade@lanbook.ru
ICQ: 446-869-967
www.lanpbl.spb.ru/price.htm

в Москве и в Московской области

«ЛАНЬ-ПРЕСС»
109263, Москва, 7-ая ул. Текстильщиков, д. 6/19
тел.: (499) 178-65-85
e-mail: lanpress@lanbook.ru

в Краснодаре и в Краснодарском крае

«ЛАНЬ-ЮГ»
350901, Краснодар, ул. Жлобы, д. 1/1
тел.: (861) 274-10-35
e-mail:lankrd98@mail.ru

ДЛЯ РОЗНИЧНЫХ ПОКУПАТЕЛЕЙ:

интернет-магазины:

Издательство «Лань»: http://www.lanbook.com

«Сова»: http://www.symplex.ru
«Ozon.ru»: http://www.ozon.ru
«Библион»: http://www.biblion.ru

Издательство «ЛАНЬ»

Мы будем благодарны Вам
за пожелания по издаваемой нами литературе,
а также за предложения по изданию книг
новых авторов или переизданию
уже существующих трудов.

Мы заинтересованы в сотрудничестве
с высшими учебными заведениями
и открыты для Ваших предложений
по улучшению нашего взаимодействия.

Теперь Вы можете звонить нам бесплатно
из любых городов России по телефону

8-800-700-40-71

Дополнительную информацию
и ответы на вопросы Вы также можете получить,
обратившись по электронной почте:

market@lanbook.ru

Издательство
«ЛАНЬ»
ЛАНЬ®

ПРЕДСТАВЛЯЕМ
НОВЫЕ УЧЕБНИКИ И УЧЕБНЫЕ ПОСОБИЯ

А. Ю. ВДОВИН, Н. Л. ВОРОНЦОВА, Л. А. ЗОЛКИНА И ДР.

СПРАВОЧНИК ПО МАТЕМАТИКЕ ДЛЯ БАКАЛАВРОВ

УЧЕБНОЕ ПОСОБИЕ

Справочное пособие предназначается для студентов всех направлений и специальностей, имеющих в учебном плане математические дисциплины. Содержит материал по основным разделам математики, предусмотренным программами бакалавриата: линейная алгебра, векторная алгебра, аналитическая геометрия, математический анализ, комплексный анализ, дифференциальные уравнения, ряды, теория вероятностей, математическая статистика, дискретная математика.

Рекомендуется студентам дневной и, особенно, заочной и дистанционной форм обучения. По мнению авторов, его использование существенно повышает эффективность проведения практических и лабораторных занятий, а также результативность самостоятельной работы при подготовке к контрольным мероприятиям, зачетам и экзаменам, проводимым как в традиционной форме, так и в форме тестирования.

Пособие обладает оптимальным сочетанием небольшого объема с полнотой и доступностью изложения материала.

Издательство
«ЛАНЬ»
ЛАНЬ®

ПРЕДСТАВЛЯЕМ
НОВЫЕ УЧЕБНИКИ И УЧЕБНЫЕ ПОСОБИЯ

А. А. ЮРЬЕВА

МАТЕМАТИЧЕСКОЕ ПРОГРАММИРОВАНИЕ

УЧЕБНОЕ ПОСОБИЕ

Учебное пособие состоит из семи разделов. Три раздела посвящены математическому программированию, два — теории игр, два — теории графов и сетей. По объему информации оно соответствует курсу математического программирования, читаемому во всех технических и экономических вузах страны. Книга написана преемственно с линейной алгеброй и теорией вероятностей. Основное внимание уделено прикладному аспекту. Все методы решения иллюстрируются типовыми примерами, а в конце каждой главы приведены упражнения (25–30 вариантов) для самостоятельной работы студентов. Задачи данных упражнений, в основном, оригинальны и лишь некоторые взяты из источников, указанных в списке литературы.

За первые три раздела книги автор имеет гриф УМО, а за следующие два — Диплом участника программы «300 лучших учебников для высшей школы в честь 300-летия Санкт-Петербурга» (г. Санкт-Петербург, 2006 г).

Издательство
«ЛАНЬ»
ЛАНЬ®

ПРЕДСТАВЛЯЕМ
НОВЫЕ УЧЕБНИКИ И УЧЕБНЫЕ ПОСОБИЯ

А. В. ОСИПОВ

ЛЕКЦИИ ПО ВЫСШЕЙ МАТЕМАТИКЕ

УЧЕБНОЕ ПОСОБИЕ

Книга является конспектом лекций, которые читаются автором студентам нематематических специальностей. Программа курса реализуется в течение трех семестров. Однако содержание конспекта несколько шире, поскольку часть материала не входит в эту программу. Книга снабжена задачами для практических занятий. Как правило, к задачам приводятся ответы или указания к решению (в конце каждой главы). В таких случаях над номером задачи ставится небольшой кружочек (например, 1°). Если задача потруднее (что встречается не слишком часто), то над номером ставится звездочка (например, 7*). Если звездочка стоит перед заголовком раздела, то этот раздел можно пропустить без ущерба для понимания дальнейшего материала.

Учебное пособие предназначено для студентов вузов, обучающихся по направлениям подготовки «Геология», «География», «Социология», «Психология», «Экономика» и «Менеджмент».

Издательство
«ЛАНЬ» ЛАНЬ®

ПРЕДСТАВЛЯЕМ
НОВЫЕ УЧЕБНИКИ И УЧЕБНЫЕ ПОСОБИЯ

А. С. ГЕРАСИМОВ

КУРС МАТЕМАТИЧЕСКОЙ ЛОГИКИ И ТЕОРИИ ВЫЧИСЛИМОСТИ

УЧЕБНОЕ ПОСОБИЕ

Пособие предназначено для изучения математической логики и теории алгоритмов. В нём описаны язык логики высказываний и язык логики предикатов первого порядка, семантика этих языков. Изложены исчисления гильбертовского типа, секвенциальные исчисления и метод резолюций как способы формального математического доказательства. Рассмотрены основные формальные аксиоматические теории. Теория алгоритмов представлена теорией вычислимости, в рамках которой дано несколько точных определений понятия алгоритма и доказана неразрешимость ряда проблем. Рассмотрены теоремы Гёделя о неполноте. Изложено исчисление Хоара для формального доказательства корректности программ некоторого императивного языка программирования. В книге имеется более 200 упражнений. Учебное пособие адресовано студентам, обучающимся по направлениям подготовки укрупнённых групп «Компьютерные и информационные науки», «Информатика и вычислительная техника», «Математика и механика».

Издательство
«ЛАНЬ»
ЛАНЬ®

ПРЕДСТАВЛЯЕМ
НОВЫЕ УЧЕБНИКИ И УЧЕБНЫЕ ПОСОБИЯ

А. И. БЛАГОДАТСКИХ, Н. Н. ПЕТРОВ

СБОРНИК ЗАДАЧ И УПРАЖНЕНИЙ ПО ТЕОРИИ ИГР

УЧЕБНОЕ ПОСОБИЕ

Задачник предназначен как для первоначального, так и для углубленного изучения теории игр. Представлены задачи и упражнения по всем основным классам игр: матричным, антагонистическим, позиционным, кооперативным, дифференциальным играм, играм n лиц в нормальной форме. Приведены индивидуальные задания для студентов. Каждый параграф начинается со сводки основных фактов.

Для студентов, аспирантов и научных работников, изучающих теорию игр.

Издательство
«ЛАНЬ»
ЛАНЬ®

ПРЕДСТАВЛЯЕМ
НОВЫЕ УЧЕБНИКИ И УЧЕБНЫЕ ПОСОБИЯ

ПОД РЕД. Г. И. КУРБАТОВОЙ

ПРАКТИЧЕСКИЕ ЗАНЯТИЯ ПО АЛГЕБРЕ. ЭЛЕМЕНТЫ ТЕОРИИ МНОЖЕСТВ, ТЕОРИИ ЧИСЕЛ, КОМБИНАТОРИКИ. АЛГЕБРАИЧЕСКИЕ СТРУКТУРЫ

УЧЕБНОЕ ПОСОБИЕ

Книга охватывает материал первых лекций курса алгебры. В пособии рассмотрены задачи из элементарной теории множеств и отображений, простейшие задачи по алгебраическим структурам, задачи по элементарной теории чисел, комбинаторные задачи.

Настоящее учебное пособие является первым в запланированной серии, состоящей из четырех частей и охватывающей весь обязательный практический материал по курсу алгебры для обучающихся по образовательным программам подготовки бакалавров университетов и технических вузов по направлениям «Прикладные математика и физика», «Прикладные математика и информатика» и «Фундаментальная информатика и информационные технологии».

В. А. ОХОРЗИН, К. В. САФОНОВ

ТЕОРИЯ УПРАВЛЕНИЯ

УЧЕБНИК

В учебнике рассматриваются модели и методы автоматического и оптимального управления системами, описываемыми обыкновенными дифференциальными уравнениями на основе передаточных функций, частотных методов, методов вариационного исчисления, принципа максимума, динамического программирования и метода моментов. Теория сопровождается многочисленными примерами и программами в системе MathCAD. Алгоритмы управления иллюстрированы примерами задач управления орбитами геостационарных спутников.

Предназначен студентам высших учебных заведений, обучающимся по направлениям подготовки «Информатика и вычислительная техника», «Прикладная математика», а также всем, кто интересуется применением методов классического и оптимального управления.

Издательство
«ЛАНЬ»
ЛАНЬ®

ПРЕДСТАВЛЯЕМ
НОВЫЕ УЧЕБНИКИ И УЧЕБНЫЕ ПОСОБИЯ

А. Н. ВАСИЛЬЕВ

ЧИСЛОВЫЕ РАСЧЕТЫ В EXCEL

УЧЕБНОЕ ПОСОБИЕ

Книга посвящена методам решения вычислительных задач с помощью приложения Excel. Тематика книги охватывает алгебраические уравнения и системы, интерполирование и аппроксимацию функциональных зависимостей, дифференцирование и интегрирование, решение дифференциальных и интегральных уравнений, а также некоторые другие темы из области вычислительных методов. Помимо этого, в книге описываются основные приемы работы с приложением Excel, обсуждаются способы организации рабочих документов, анализируются методы ввода и редактирования данных в рабочих документах, изучаются возможности применения форматов и стилей, иллюстрируются принципы использования встроенных вычислительных утилит, а также даются основы программирования в VBA. Книга может использоваться в качестве учебного пособия при изучении курсов вычислительной математики и математического программирования, в качестве самоучителя или справочного пособия при решении вычислительных задач средствами приложения Excel.

Издательство
«ЛАНЬ» ЛАНЬ®

ПРЕДСТАВЛЯЕМ
НОВЫЕ УЧЕБНИКИ И УЧЕБНЫЕ ПОСОБИЯ

Л. И. ВЫСОЦКИЙ, Г. Р. КОПЕРНИК, И. С. ВЫСОЦКИЙ

МАТЕМАТИЧЕСКОЕ И ФИЗИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПОТЕНЦИАЛЬНЫХ ТЕЧЕНИЙ ЖИДКОСТИ

УЧЕБНОЕ ПОСОБИЕ

Учебное пособие будет полезно студентам при изучении раздела курса «Механика жидкости и газа», посвященного рассмотрению модельного течения жидкости, называемого потенциальным или безвихревым. Предлагается способ наглядного воспроизводства и построения очертаний линий тока и эквипотенциалей с помощью ПЭВМ и вывода соответствующей информации в виде гидродинамической сетки на экран дисплея.

Предназначено для студентов технических направлений подготовки и специальностей, может быть полезно аспирантам и преподавателям.