

Christian Heumann · Michael Schomaker  
Shalabh

# Introduction to Statistics and Data Analysis

With Exercises, Solutions and  
Applications in R

---

# Introduction to Statistics and Data Analysis

---

Christian Heumann · Michael Schomaker  
Shalabh

# Introduction to Statistics and Data Analysis

With Exercises, Solutions  
and Applications in R

Christian Heumann  
Department of Statistics  
Ludwig-Maximilians-Universität München  
München  
Germany

Shalabh  
Department of Mathematics and Statistics  
Indian Institute of Technology Kanpur  
Kanpur  
India

Michael Schomaker  
Centre for Infectious Disease Epidemiology  
and Research  
University of Cape Town  
Cape Town  
South Africa

ISBN 978-3-319-46160-1  
DOI 10.1007/978-3-319-46162-5

ISBN 978-3-319-46162-5 (eBook)

Library of Congress Control Number: 2016955516

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature  
The registered company is Springer International Publishing AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

---

## Preface

The success of the open-source statistical software “*R*” has made a significant impact on the teaching and research of statistics in the last decade. Analysing data is now easier and more affordable than ever, but choosing the most appropriate statistical methods remains a challenge for many users. To understand and interpret software output, it is necessary to engage with the fundamentals of statistics.

However, many readers do not feel comfortable with complicated mathematics. In this book, we attempt to find a healthy balance between explaining statistical concepts comprehensively and showing their application and interpretation using *R*.

This book will benefit beginners and self-learners from various backgrounds as we complement each chapter with various exercises and detailed and comprehensible solutions. The results involving mathematics and rigorous proofs are separated from the main text, where possible, and are kept in an appendix for interested readers. Our textbook covers material that is generally taught in introductory-level statistics courses to students from various backgrounds, including sociology, biology, economics, psychology, medicine, and others. Most often, we introduce the statistical concepts using examples and illustrate the calculations both manually and using *R*.

However, while we provide a gentle introduction to *R* (in the appendix), this is not a software book. Our emphasis lies on explaining statistical concepts correctly and comprehensively, using exercises and software to delve deeper into the subject matter and learn about the conceptual challenges that the methods present.

This book’s homepage, <http://chris.userweb.mwn.de/book/>, contains additional material, most notably the software codes needed to answer the software exercises, and data sets. In the remainder of this book, we will use grey boxes

*R-command()*



to introduce the relevant *R* commands. In many cases, the code can be directly pasted into *R* to reproduce the results and graphs presented in the book; in others, the code is abbreviated to improve readability and clarity, and the detailed code can be found online.

Many years of teaching experience, from undergraduate to postgraduate level, went into this book. The authors hope that the reader will enjoy reading it and find it a useful reference for learning. We welcome critical feedback to improve future editions of this book. Comments can be sent to [christian.heumann@stat.uni-muenchen.de](mailto:christian.heumann@stat.uni-muenchen.de), [shalab@iitk.ac.in](mailto:shalab@iitk.ac.in), and [michael.schomaker@uct.ac.za](mailto:michael.schomaker@uct.ac.za) who contributed equally to this book.

We thank Melanie Schomaker for producing some of the figures and giving graphical advice, Alice Blanck from Springer for her continuous help and support, and Lyn Imeson for her dedicated commitment which improved the earlier versions of this book. We are grateful to our families who have supported us during the preparation of this book.

München, Germany  
Cape Town, South Africa  
Kanpur, India  
November 2016

Christian Heumann  
Michael Schomaker  
Shalabh

---

# Contents

## Part I Descriptive Statistics

<b>1</b>	<b>Introduction and Framework</b>	3
1.1	Population, Sample, and Observations	3
1.2	Variables	4
1.2.1	Qualitative and Quantitative Variables	5
1.2.2	Discrete and Continuous Variables	6
1.2.3	Scales	6
1.2.4	Grouped Data	7
1.3	Data Collection	8
1.4	Creating a Data Set	9
1.4.1	Statistical Software	12
1.5	Key Points and Further Issues	13
1.6	Exercises	14
<b>2</b>	<b>Frequency Measures and Graphical Representation of Data</b>	17
2.1	Absolute and Relative Frequencies	17
2.2	Empirical Cumulative Distribution Function	19
2.2.1	ECDF for Ordinal Variables	20
2.2.2	ECDF for Continuous Variables	22
2.3	Graphical Representation of a Variable	24
2.3.1	Bar Chart	24
2.3.2	Pie Chart	26
2.3.3	Histogram	27
2.4	Kernel Density Plots	29
2.5	Key Points and Further Issues	32
2.6	Exercises	32
<b>3</b>	<b>Measures of Central Tendency and Dispersion</b>	37
3.1	Measures of Central Tendency	38
3.1.1	Arithmetic Mean	38
3.1.2	Median and Quantiles	40
3.1.3	Quantile–Quantile Plots (QQ-Plots)	44
3.1.4	Mode	45

3.1.5	Geometric Mean . . . . .	46
3.1.6	Harmonic Mean . . . . .	48
3.2	Measures of Dispersion. . . . .	48
3.2.1	Range and Interquartile Range. . . . .	49
3.2.2	Absolute Deviation, Variance, and Standard Deviation . . . . .	50
3.2.3	Coefficient of Variation . . . . .	55
3.3	Box Plots . . . . .	56
3.4	Measures of Concentration . . . . .	57
3.4.1	Lorenz Curve. . . . .	58
3.4.2	Gini Coefficient . . . . .	60
3.5	Key Points and Further Issues . . . . .	63
3.6	Exercises. . . . .	63
<b>4</b>	<b>Association of Two Variables . . . . .</b>	<b>67</b>
4.1	Summarizing the Distribution of Two Discrete Variables . . . . .	68
4.1.1	Contingency Tables for Discrete Data . . . . .	68
4.1.2	Joint, Marginal, and Conditional Frequency Distributions . . . . .	70
4.1.3	Graphical Representation of Two Nominal or Ordinal Variables. . . . .	72
4.2	Measures of Association for Two Discrete Variables . . . . .	74
4.2.1	Pearson's $\chi^2$ Statistic. . . . .	76
4.2.2	Cramer's $V$ Statistic. . . . .	77
4.2.3	Contingency Coefficient $C$ . . . . .	77
4.2.4	Relative Risks and Odds Ratios. . . . .	78
4.3	Association Between Ordinal and Continuous Variables. . . . .	79
4.3.1	Graphical Representation of Two Continuous Variables . . . . .	79
4.3.2	Correlation Coefficient. . . . .	82
4.3.3	Spearman's Rank Correlation Coefficient. . . . .	84
4.3.4	Measures Using Discordant and Concordant Pairs. . . . .	86
4.4	Visualization of Variables from Different Scales. . . . .	88
4.5	Key Points and Further Issues . . . . .	89
4.6	Exercises. . . . .	90
<b>Part II Probability Calculus</b>		
<b>5</b>	<b>Combinatorics . . . . .</b>	<b>97</b>
5.1	Introduction . . . . .	97
5.2	Permutations . . . . .	101
5.2.1	Permutations without Replacement . . . . .	101
5.2.2	Permutations with Replacement. . . . .	101
5.3	Combinations . . . . .	102

5.3.1	Combinations without Replacement and without Consideration of the Order . . . . .	102
5.3.2	Combinations without Replacement and with Consideration of the Order . . . . .	103
5.3.3	Combinations with Replacement and without Consideration of the Order . . . . .	103
5.3.4	Combinations with Replacement and with Consideration of the Order . . . . .	104
5.4	Key Points and Further Issues . . . . .	105
5.5	Exercises . . . . .	105
<b>6</b>	<b>Elements of Probability Theory</b> . . . . .	109
6.1	Basic Concepts and Set Theory . . . . .	109
6.2	Relative Frequency and Laplace Probability . . . . .	113
6.3	The Axiomatic Definition of Probability . . . . .	115
6.3.1	Corollaries Following from Kolomogorov's Axioms . . . . .	116
6.3.2	Calculation Rules for Probabilities . . . . .	117
6.4	Conditional Probability . . . . .	117
6.4.1	Bayes' Theorem . . . . .	120
6.5	Independence . . . . .	121
6.6	Key Points and Further Issues . . . . .	123
6.7	Exercises . . . . .	123
<b>7</b>	<b>Random Variables</b> . . . . .	127
7.1	Random Variables . . . . .	127
7.2	Cumulative Distribution Function (CDF) . . . . .	129
7.2.1	CDF of Continuous Random Variables . . . . .	129
7.2.2	CDF of Discrete Random Variables . . . . .	131
7.3	Expectation and Variance of a Random Variable . . . . .	134
7.3.1	Expectation . . . . .	134
7.3.2	Variance . . . . .	135
7.3.3	Quantiles of a Distribution . . . . .	137
7.3.4	Standardization . . . . .	138
7.4	Tschebyschev's Inequality . . . . .	139
7.5	Bivariate Random Variables . . . . .	140
7.6	Calculation Rules for Expectation and Variance . . . . .	144
7.6.1	Expectation and Variance of the Arithmetic Mean . . . . .	145
7.7	Covariance and Correlation . . . . .	146
7.7.1	Covariance . . . . .	147
7.7.2	Correlation Coefficient . . . . .	148
7.8	Key Points and Further Issues . . . . .	149
7.9	Exercises . . . . .	149

<b>8</b>	<b>Probability Distributions</b>	153
8.1	Standard Discrete Distributions	154
8.1.1	Discrete Uniform Distribution	154
8.1.2	Degenerate Distribution	156
8.1.3	Bernoulli Distribution	156
8.1.4	Binomial Distribution	157
8.1.5	Poisson Distribution	160
8.1.6	Multinomial Distribution	161
8.1.7	Geometric Distribution	163
8.1.8	Hypergeometric Distribution	163
8.2	Standard Continuous Distributions	165
8.2.1	Continuous Uniform Distribution	165
8.2.2	Normal Distribution	166
8.2.3	Exponential Distribution	170
8.3	Sampling Distributions	171
8.3.1	$\chi^2$ -Distribution	171
8.3.2	<i>t</i> -Distribution	172
8.3.3	<i>F</i> -Distribution	173
8.4	Key Points and Further Issues	174
8.5	Exercises	175

### Part III Inductive Statistics

<b>9</b>	<b>Inference</b>	181
9.1	Introduction	181
9.2	Properties of Point Estimators	183
9.2.1	Unbiasedness and Efficiency	183
9.2.2	Consistency of Estimators	189
9.2.3	Sufficiency of Estimators	190
9.3	Point Estimation	192
9.3.1	Maximum Likelihood Estimation	192
9.3.2	Method of Moments	195
9.4	Interval Estimation	195
9.4.1	Introduction	195
9.4.2	Confidence Interval for the Mean of a Normal Distribution	197
9.4.3	Confidence Interval for a Binomial Probability	199
9.4.4	Confidence Interval for the Odds Ratio	201
9.5	Sample Size Determinations	203
9.6	Key Points and Further Issues	205
9.7	Exercises	205
<b>10</b>	<b>Hypothesis Testing</b>	209
10.1	Introduction	209
10.2	Basic Definitions	210

10.2.1	One- and Two-Sample Problems . . . . .	210
10.2.2	Hypotheses . . . . .	210
10.2.3	One- and Two-Sided Tests . . . . .	211
10.2.4	Type I and Type II Error . . . . .	213
10.2.5	How to Conduct a Statistical Test . . . . .	214
10.2.6	Test Decisions Using the $p$ -Value . . . . .	215
10.2.7	Test Decisions Using Confidence Intervals . . . . .	216
10.3	Parametric Tests for Location Parameters . . . . .	216
10.3.1	Test for the Mean When the Variance is Known (One-Sample Gauss Test) . . . . .	216
10.3.2	Test for the Mean When the Variance is Unknown (One-Sample $t$ -Test) . . . . .	219
10.3.3	Comparing the Means of Two Independent Samples . . . . .	221
10.3.4	Test for Comparing the Means of Two Dependent Samples (Paired $t$ -Test) . . . . .	225
10.4	Parametric Tests for Probabilities . . . . .	227
10.4.1	One-Sample Binomial Test for the Probability $p$ . . . . .	227
10.4.2	Two-Sample Binomial Test . . . . .	230
10.5	Tests for Scale Parameters . . . . .	232
10.6	Wilcoxon–Mann–Whitney (WMW) U-Test . . . . .	232
10.7	$\chi^2$ -Goodness-of-Fit Test . . . . .	235
10.8	$\chi^2$ -Independence Test and Other $\chi^2$ -Tests . . . . .	238
10.9	Key Points and Further Issues . . . . .	242
10.10	Exercises . . . . .	242
11	<b>Linear Regression</b> . . . . .	249
11.1	The Linear Model . . . . .	250
11.2	Method of Least Squares . . . . .	252
11.2.1	Properties of the Linear Regression Line . . . . .	255
11.3	Goodness of Fit . . . . .	256
11.4	Linear Regression with a Binary Covariate . . . . .	259
11.5	Linear Regression with a Transformed Covariate . . . . .	261
11.6	Linear Regression with Multiple Covariates . . . . .	262
11.6.1	Matrix Notation . . . . .	263
11.6.2	Categorical Covariates . . . . .	265
11.6.3	Transformations . . . . .	267
11.7	The Inductive View of Linear Regression . . . . .	269
11.7.1	Properties of Least Squares and Maximum Likelihood Estimators . . . . .	273
11.7.2	The ANOVA Table . . . . .	274
11.7.3	Interactions . . . . .	276
11.8	Comparing Different Models . . . . .	280
11.9	Checking Model Assumptions . . . . .	285

11.10 Association Versus Causation . . . . .	288
11.11 Key Points and Further Issues . . . . .	289
11.12 Exercises. . . . .	290
<b>Appendix A: Introduction to <i>R</i> . . . . .</b>	<b>297</b>
<b>Appendix B: Solutions to Exercises . . . . .</b>	<b>321</b>
<b>Appendix C: Technical Appendix . . . . .</b>	<b>423</b>
<b>Appendix D: Visual Summaries . . . . .</b>	<b>443</b>
<b>References . . . . .</b>	<b>449</b>
<b>Index . . . . .</b>	<b>451</b>

---

## About the Authors

**Prof. Christian Heumann** is a professor at the Ludwig-Maximilians-Universität München, Germany, where he teaches students in Bachelor and Master programs offered by the Department of Statistics, as well as undergraduate students in the Bachelor of Science programs in business administration and economics. His research interests include statistical modeling, computational statistics and all aspects of missing data.

**Dr. Michael Schomaker** is a Senior Researcher and Biostatistician at the Centre for Infectious Disease Epidemiology & Research (CIDER), University of Cape Town, South Africa. He received his doctoral degree from the University of Munich. He has taught undergraduate students for many years and has written contributions for various introductory textbooks. His research focuses on missing data, causal inference, model averaging and HIV/AIDS.

**Prof. Shalabh** is a Professor at the Indian Institute of Technology Kanpur, India. He received his Ph.D. from the University of Lucknow (India) and completed his post-doctoral work at the University of Pittsburgh (USA) and University of Munich (Germany). He has over twenty years of experience in teaching and research. His main research areas are linear models, regression analysis, econometrics, measurement error models, missing data models and sampling theory.

---

**Part I**

**Descriptive Statistics**

Statistics is a collection of methods which help us to describe, summarize, interpret, and analyse data. Drawing conclusions from data is vital in research, administration, and business. Researchers are interested in understanding whether a medical intervention helps in reducing the burden of a disease, how personality relates to decision-making, whether a new fertilizer increases the yield of crops, how a political system affects trade policy, who is going to vote for a political party in the next election, what are the long-term changes in the population of a fish species, and many more questions. Governments and organizations may be interested in the life expectancy of a population, the risk factors for infant mortality, geographical differences in energy usage, migration patterns, or reasons for unemployment. In business, identifying people who may be interested in a certain product, optimizing prices, and evaluating the satisfaction of customers are possible areas of interest.

No matter what the question of interest is, it is important to collect data in a way which allows its analysis. The representation of collected data in a **data set** or **data matrix** allows the application of a variety of statistical methods. In the first part of the book, we are going to introduce methods which help us in *describing* data, and the second and third parts of the book focus on inferential statistics, which means *drawing conclusions* from data. In this chapter, we are going to introduce the framework of statistics which is needed to properly collect, administer, evaluate, and analyse data.

---

## 1.1 Population, Sample, and Observations

Let us first introduce some terminology and related notations used in this book. The **units** on which we measure data—such as persons, cars, animals, or plants—are called **observations**. These units/observations are represented by the Greek

symbol  $\omega$ . The collection of all units is called **population** and is represented by  $\Omega$ . When we refer to  $\omega \in \Omega$ , we mean a single unit out of all units, e.g. one person out of all persons of interest. If we consider a selection of observations  $\omega_1, \omega_2, \dots, \omega_n$ , then these observations are called **sample**. A sample is always a subset of the population,  $\{\omega_1, \omega_2, \dots, \omega_n\} \subseteq \Omega$ .

### *Example 1.1.1*

- If we are interested in the social conditions under which Indian people live, then we would define all inhabitants of India as  $\Omega$  and each of its inhabitants as  $\omega$ . If we want to collect data from a few inhabitants, then those would represent a sample from the total population.
- Investigating the economic power of Africa's platinum industry would require to treat each platinum-related company as  $\omega$ , whereas all platinum-related companies would be collected in  $\Omega$ . A few companies  $\omega_1, \omega_2, \dots, \omega_n$  comprise a sample of all companies.
- We may be interested in collecting information about those participating in a statistics course. All participants in the course constitute the population  $\Omega$ , and each participant refers to a unit or observation  $\omega$ .

*Remark 1.1.1* Sometimes, the concept of a population is not applicable or difficult to imagine. As an example, imagine that we measure the temperature in New Delhi every hour. A sample would then be the time series of temperatures in a specific time window, for example from January to March 2016. A population in the sense of observational units does not exist here. But now assume that we measure temperatures in several different cities; then, all the cities form the population, and a sample is any subset of the cities.

## 1.2 Variables

If we have specified the population of interest for a specific research question, we can think of what is of interest about our observations. A particular feature of these observations can be collected in a statistical **variable**  $X$ . Any information we are interested in may be captured in such a variable. For example, if our observations refer to human beings,  $X$  may describe marital status, gender, age, or anything else which may relate to a person. Of course, we can be interested in many different features, each of them collected in a different variable  $X_i, i = 1, 2, \dots, p$ . Each observation  $\omega$  takes a particular value for  $X$ . If  $X$  refers to gender, each observation, i.e. each person, has a particular value  $x$  which refers to either "male" or "female".

The formal definition of a variable is

$$\begin{aligned} X : \Omega &\rightarrow S \\ \omega &\mapsto x \end{aligned} \tag{1.1}$$

This definition states that a variable  $X$  takes a value  $x$  for each observation  $\omega \in \Omega$ , whereby the number of possible values is contained in the set  $S$ .

*Example 1.2.1*

- If  $X$  refers to gender, possible  $x$ -values are contained in  $S = \{\text{male, female}\}$ . Each observation  $\omega$  is either male or female, and this information is summarized in  $X$ .
- Let  $X$  be the country of origin for a car. Possible values to be taken by an observation  $\omega$  (i.e. a car) are  $S = \{\text{Italy, South Korea, Germany, France, India, China, Japan, USA, ...}\}$ .
- A variable  $X$  which refers to age may take any value between 1 and 125. Each person  $\omega$  is assigned a value  $x$  which represents the age of this person.

### 1.2.1 Qualitative and Quantitative Variables

Qualitative variables are the variables which take values  $x$  that cannot be ordered in a logical or natural way. For example,

- the colour of the eye,
- the name of a political party, and
- the type of transport used to travel to work

are all qualitative variables. Neither is there any reason to list blue eyes before brown eyes (or vice versa) nor does it make sense to list buses before trains (or vice versa).

Quantitative variables represent measurable quantities. The values which these variables can take can be ordered in a logical and natural way. Examples of quantitative variables are

- size of shoes,
- price for houses,
- number of semesters studied, and
- weight of a person.

*Remark 1.2.1* It is common to assign numbers to qualitative variables for practical purposes in data analyses (see Sect. 1.4 for more detail). For instance, if we consider the variable “gender”, then each observation can take either the “value” male or female. We may decide to assign 1 to female and 0 to male and use these numbers instead of the original categories. However, this is arbitrary, and we could have also chosen “1” for male and “0” for female, or “2” for male and “10” for female. There is no logical and natural order on how to arrange male and female, and thus, the variable gender remains a qualitative variable, even after using numbers for coding the values that  $X$  can take.

### 1.2.2 Discrete and Continuous Variables

**Discrete variables** are variables which can only take a finite number of values. All qualitative variables are discrete, such as the colour of the eye or the region of a country. But also quantitative variables can be discrete: the size of shoes or the number of semesters studied would be discrete because the number of values these variables can take is limited.

Variables which can take an infinite number of values are called **continuous variables**. Examples are the time it takes to travel to university, the length of an antelope, and the distance between two planets. Sometimes, it is said that continuous variables are variables which are “measured rather than counted”. This is a rather informal definition which helps to understand the difference between discrete and continuous variables. The crucial point is that continuous variables can, in theory, take an infinite number of values; for instance, the height of a person may be recorded as 172 cm. However, the actual height on the measuring tape might be 172.3 cm which was rounded off to 172 cm. If one had a better measuring instrument, we may have obtained 172.342 cm. But the real height of this person is a number with indefinitely many decimal places such as 172.342975328... cm. No matter what we eventually report or obtain, a variable which can take an infinite amount of values is defined to be a continuous variable.

### 1.2.3 Scales

The thoughts and considerations from above indicate that different variables contain different amounts of information. A useful classification of these considerations is given by the concept of the **scale** of a variable. This concept will help us in the remainder of this book to identify which methods are the appropriate ones to use in a particular setting.

**Nominal scale.** The values of a *nominal variable* cannot be ordered. Examples are the gender of a person (male–female) or the status of an application (pending–not pending).

**Ordinal scale.** The values of an *ordinal variable* can be ordered. However, the differences between these values cannot be interpreted in a meaningful way. For example, the possible values of education level (none–primary education–secondary education–university degree) can be ordered meaningfully, but the differences between these values cannot be interpreted. Likewise, the satisfaction with a product (unsatisfied–satisfied–very satisfied) is an ordinal variable because the values this variable can take can be ordered, but the differences between “unsatisfied–satisfied” and “satisfied–very satisfied” cannot be compared in a numerical way.

**Continuous scale.** The values of a *continuous variable* can be ordered. Furthermore, the differences between these values can be interpreted in a meaningful way. For instance, the height of a person refers to a continuous variable because the values can be ordered (170 cm, 171 cm, 172 cm, ...), and differences between these

values can be compared (the difference between 170 and 171 cm is the same as the difference between 171 and 172 cm). Sometimes, the continuous scale is divided further into subscales. While in the remainder of the book we typically do not need these classifications, it is still useful to reflect on them:

*Interval scale.* Only differences between values, but not ratios, can be interpreted.

An example for this scale would be temperature (measured in °C): the difference between  $-2\text{ }^{\circ}\text{C}$  and  $4\text{ }^{\circ}\text{C}$  is  $6\text{ }^{\circ}\text{C}$ , but the ratio of  $4 / -2 = -2$  does not mean that  $-4\text{ }^{\circ}\text{C}$  is twice as cold as  $2\text{ }^{\circ}\text{C}$ .

*Ratio scale.* Both differences and ratios can be interpreted. An example is speed:  $60\text{ km/h}$  is  $40\text{ km/h}$  more than  $20\text{ km/h}$ . Moreover,  $60\text{ km/h}$  is three times faster than  $20\text{ km/h}$  because the ratio between them is 3.

*Absolute scale.* The absolute scale is the same as the ratio scale, with the exception that the values are measured in “natural” units. An example is “number of semesters studied” where no artificial unit such as km/h or °C is needed: the values are simply 1, 2, 3, ....

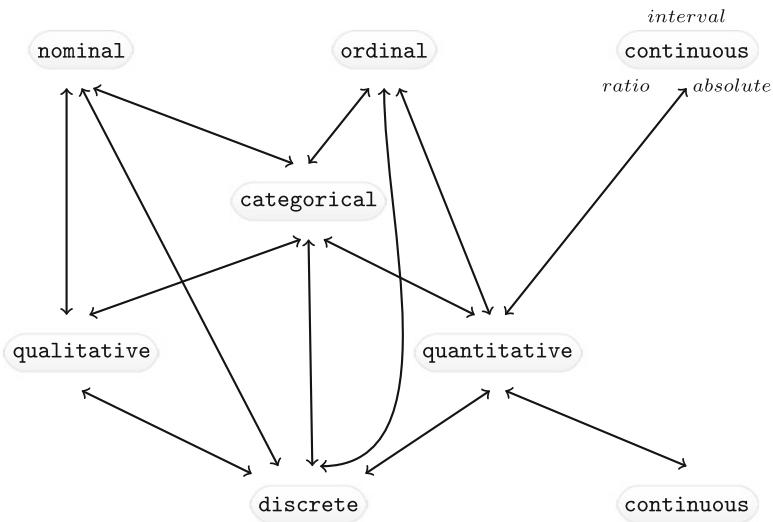
#### 1.2.4 Grouped Data

Sometimes, data may be available only in a summarized form: instead of the original value, one may only know the category or group the value belongs to. For example,

- it is often convenient in a survey to ask for the income (per year) by means of groups: [ $\text{€}0$ – $\text{€}20,000$ ], [ $\text{€}20,000$ – $\text{€}30,000$ ], ...,  $> \text{€}100,000$ ];
- if there are many political parties in an election, those with a low number of voters are often summarized in a new category “Other Parties”;
- instead of capturing the number of claims made by an insurance company customer, the variable “claimed” may denote whether or not the customer claimed at all (yes–no).

If data is available in grouped form, we call the respective variable capturing this information a **grouped variable**. Sometimes, these variables are also known as **categorical variables**. This is, however, not a complete definition because categorical variables refer to any type of variable which takes a finite, possibly small, number of values. Thus, any discrete and/or nominal and/or ordinal and/or qualitative variable may be regarded as a categorical variable. Any grouped or categorical variable which can only take two values is called a **binary variable**.

To gain a better understanding on how the definitions from the above sections relate to each other see Fig. 1.1. Qualitative data is always discrete, but quantitative data can be both discrete (e.g. size of shoes or a grouped variable) and continuous (e.g. temperature). Nominal variables are always qualitative and discrete (e.g. colour of the eye), whereas continuous variables are always quantitative (e.g. temperature). Categorical variables can be both qualitative (e.g. colour of the eye) and quantitative (satisfaction level on a scale from 1 to 5). Categorical variables are never continuous.



**Fig. 1.1** Summary of variable classifications

### 1.3 Data Collection

When collecting data, we may ask ourselves how to facilitate this in detail and how much data needs to be collected. The latter question will be partly answered in Sect. 9.5; but in general, we can think of collecting data either on all subjects of interest, such as in a national census, or on a representative sample of the population. Most commonly, we gather data on a sample (described in the Part I of this book) and then draw conclusions about the population of interest (discussed in the Part III of this book). A sample might either be chosen by us or obtained through third parties (hospitals, government agencies), or created during an experiment. This depends on the context as described below.

**Survey.** A survey typically (but not always) collects data by asking questions (in person or by phone) or providing questionnaires to study participants (as a printout or online). For example, an opinion poll before a national election provides evidence about the future government: potential voters are asked by phone which party they are going to vote for in the next election; on the day of the election, this information can be updated by asking the same question to a sample of voters who have just delivered their vote at the polling station (so-called exit poll). A behavioural research survey may ask members of a community about their knowledge and attitudes towards drug use. For this purpose, the study coordinators can send people with a questionnaire to this community and interview members of randomly selected households.

Ideally, a survey is conducted in a way which makes the chosen sample representative of the population of interest. If a marketing company interviews people in a pedestrian zone to find their views about a new chocolate bar, then these people

may not be representative of those who will potentially be interested in this product. Similarly, if students are asked to fill in an online survey to evaluate a lecture, it may turn out that those who participate are on average less satisfied than those who do not. Survey sampling is a complex topic on its own. The interested reader may consult Groves et al. (2009) or Kauermann and Küchenhoff (2011).

**Experiment.** Experimental data is obtained in “controlled” settings. This can mean many things, but essentially it is data which is generated by the researcher with full control over one or many variables of interest. For instance, suppose there are two competing toothpastes, both of which promise to reduce pain for people with sensitive teeth. If the researcher decided to randomly assign toothpaste *A* to half of the study participants, and toothpaste *B* to the other half, then this is an experiment because it is only the researcher who decides which toothpaste is to be used by any of the participants. It is not decided by the participant. The data of the variable toothpaste is controlled by the experimenter. Consider another example where the production process of a product can potentially be reduced by combining two processes. The management could decide to implement the new process in three production facilities, but leave it as it is in the other facilities. The production process for the different units (facilities) is therefore under control of the management. However, if each facility could decide for themselves if they wanted a change or not, it would not be an experiment because factors not directly controlled by the management, such as the leadership style of the facility manager, would determine which process is chosen.

**Observational Data.** Observational data is data which is collected routinely, without a researcher designing a survey or conducting an experiment. Suppose a blood sample is drawn from each patient with a particular acute infection when they arrive at a hospital. This data may be stored in the hospital’s folders and later accessed by a researcher who is interested in studying this infection. Or suppose a government institution monitors where people live and move to. This data can later be used to explore migration patterns.

**Primary and Secondary Data.** Primary data is data we collect ourselves, i.e. via a survey or experiment. Secondary data, in contrast, is collected by someone else. For example, data from a national census, publicly available databases, previous research studies, government reports, historical data, and data from the internet, among others, are secondary data.

---

## 1.4 Creating a Data Set

There is a unique way in which data is prepared and collected to utilize statistical analyses. The data is stored in a data matrix (=data set) with  $p$  columns and  $n$  rows (see Fig. 1.2). Each row corresponds to an observation/unit  $\omega$  and each column to a variable  $X$ . This means that, for example, the entry in the fourth row and second column ( $x_{42}$ ) describes the value of the fourth observation on the second variable. The examples below will illustrate the concept of a data set in more detail.

$$\begin{array}{c} \omega \\ \left( \begin{array}{ccccc} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ n & x_{n1} & x_{n2} & \cdots & x_{np} \end{array} \right) \end{array}$$

**Fig. 1.2** Data set or data matrix

$\omega$	Music	Mathematics	Biology	Geography
Student A	65	70	85	45
Student B	77	82	80	60
Student C	78	73	93	68
Student D	88	71	63	58
Student E	75	83	63	57

**Fig. 1.3** Data set of marks of five students

*Example 1.4.1* Suppose five students take examinations in music, mathematics, biology, and geography. Their marks, measured on a scale between 0 and 100 (where 100 is the best mark), can be written down as illustrated in Fig. 1.3. Note that each row refers to a student and each column to a variable. We consider a larger data set in the next example.

*Example 1.4.2* Consider the data set described in Appendix A.4. A pizza delivery service captures information related to each delivery, for example the delivery time, the temperature of the pizza, the name of the driver, the date of the delivery, the name of the branch, and many more. To capture the data of all deliveries during one month, we create a data matrix. Each row refers to a particular delivery, therefore representing the observations of the data. Each column refers to a variable. In Fig. 1.4, the variables  $X_1$  (delivery time in minutes),  $X_2$  (temperature in °C), and  $X_{12}$  (name of branch) are listed.

Delivery	Delivery Time	Temperature	...	Branch
1	35.1	68.3	...	East (1)
2	25.2	71.0	...	East (1)
$\vdots$	$\vdots$	$\vdots$		$\vdots$
1266	35.7	60.8	...	West (2)

**Fig. 1.4** Pizza data set

**Table 1.1** Coding list for branch

Variable	Values	Code
Branch	East	1
	West	2
	Centre	3
	Missing	4

The first row tells us about the features of the first pizza delivery: the delivery time was 35.1 min, the pizza arrived with a temperature of 68.3 °C, and the pizza was delivered from the branch in the East of the city. In total, there were  $n = 1266$  deliveries. For nominal variables, such as branch, we may decide to produce a coding list, as illustrated in Table 1.1: instead of referring to the branches as “East”, “West”, and “Centre”, we may simply call them 1, 2, and 3. As we will see in Chap. 11, this has benefits for some analysis methods, though this is not needed in general.

If some values are missing, for example because they were never captured or even lost, then this requires special attention. In Table 1.1, we assign missing values the number “4” and therefore treat them as a separate category. If we work with statistical software (see below), we may need other coding such as NA in the statistical software *R* or in Stata. More detail can be found in Appendix A.

Another consideration when collecting data is that of **transformations**: we may have captured the velocity of cars in kilometres per hour, but may need to present the data in miles per hour; we have captured the temperature in degrees Celsius, whereas we need to communicate results in degrees Fahrenheit, or we have created a satisfaction score which we want to range from  $-5$  to  $+5$ , while the score currently runs from 0 to 20. This is not a problem at all. We can simply create a new variable which reflects the required transformation. However, valid transformations depend on the scale of a variable. Variables on an interval scale can use transformations of the following kind:

$$g(x) = a + bx, \quad b > 0. \quad (1.2)$$

For ratio scales, only the following transformations are valid:

$$g(x) = bx, \quad b > 0. \quad (1.3)$$

In the above equation,  $a$  is set to 0 because ratios only stay the same if we respect a variable’s natural point of origin.

*Example 1.4.3* The temperature in °F relates to the temperature in °C as follows:

$$\begin{aligned} \text{Temperature in } ^\circ\text{F} &= 32 + 1.8 \text{ Temperature in } ^\circ\text{C} \\ g(x) &= a + b \quad x \end{aligned}$$

This means that  $25^\circ\text{C}$  relates to  $(32 + 1.8 \cdot 25)^\circ\text{F} = 77^\circ\text{F}$ . If  $X_1$  is a variable representing temperature by °C, we can simply create a new variable  $X_2$  which is temperature in °F. Since temperature is measured on an interval scale, this transformation is valid.

Changing currencies is also possible. If we would like to represent the price of a product not in South African Rand but in €, we simply apply the transformation

$$\text{Price in South African Rand} = b \cdot \text{Price in } \text{€}$$

whereby  $b$  is the currency exchange rate.

### 1.4.1 Statistical Software

There are number of statistical software packages which allow data collection, management, and—most importantly—analysis. In this book, we focus on the statistical software *R* which is freely available at <http://cran.r-project.org/>. A gentle introduction to *R* is provided in Appendix A. A data matrix can be created manually using commands such as `matrix()`, `data.frame()`, and others. Any data can be edited using `edit()`. However, typically analysts have already typed their data into databases or spreadsheets, for example in Excel, Access, or MySQL. In most of these applications, it is possible to save the data as an ASCII file (`.dat`), as a tab-delimited file (`.txt`), or as a comma-separated values file (`.csv`). All of these formats allow easy switching between different software and database applications. Such data can easily be read into *R* by means of the following commands:

```
setwd('C:/directory')
read.table('pizza_delivery.dat')
read.table('pizza_delivery.txt')
read.csv('pizza_delivery.csv')
```



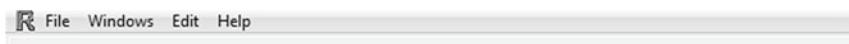
where `setwd` specifies the working directory. Alternatively, loading the library `foreign` allows the import of data from many different statistical software packages, notably Stata, SAS, Minitab, SPSS, among others. A detailed description of data import and export can be found in the respective *R* manual available at <http://cran.r-project.org/doc/manuals/r-release/R-data.pdf>. Once the data is read into *R*, it can be viewed with

```
fix()      # option 1
View()    # option 2
```



We can also get an overview of the data directly in the *R*-console by displaying only the top lines of the data with `head()`. Both approaches are visualized in Fig. 1.5 for the pizza data introduced in Example 1.4.2.

```
> pizza <- read.csv("pizza_delivery.csv")
> head(pizza)
  day      date    time operator branch   driver temperature bill
1 Thursday 01-May-14 35.12837 Laura   East    Bruno  68.28772 58.4
2 Thursday 01-May-14 25.20307 Melissa East Salvatore 70.99779 26.4
3 Thursday 01-May-14 45.64340 Melissa West Salvatore 53.39415 58.1
4 Thursday 01-May-14 29.37430 Melissa East Salvatore 70.30660 35.2
5 Thursday 01-May-14 29.99461 Melissa West Salvatore 71.50169 38.4
6 Thursday 01-May-14 40.25432 Melissa Centre Bruno  60.75950 61.8
> fix(pizza)
```



The screenshot shows the RStudio interface with the title bar "File Windows Edit Help". Below it is a grid of data representing the "pizza\_delivery.csv" file. The grid has columns labeled "day", "date", "time", "operator", "branch", "driver", "temperature", and "bill". The data consists of 11 rows, each representing a delivery record with values corresponding to the variables listed.

	day	date	time	operator	branch	driver	temperature	bill
1	Thursday	01-May-14	35.12837	Laura	East	Bruno	68.28772	58.4
2	Thursday	01-May-14	25.20307	Melissa	East	Salvatore	70.99779	26.4
3	Thursday	01-May-14	45.64340	Melissa	West	Salvatore	53.39415	58.1
4	Thursday	01-May-14	29.37430	Melissa	East	Salvatore	70.30660	35.2
5	Thursday	01-May-14	29.99461	Melissa	West	Salvatore	71.50169	38.4
6	Thursday	01-May-14	40.25432	Melissa	Centre	Bruno	60.75950	61.8
7	Thursday	01-May-14	48.72861	Laura	West	Bruno	58.2587	57.9
8	Thursday	01-May-14	34.02772	Melissa	West	Mario	68.12793	35.8
9	Thursday	01-May-14	28.20943	Laura	Centre	Mario	64.94661	36.6
10	Thursday	01-May-14	37.95479	Melissa	Centre	Bruno	60.00738	44.8
11	Thursday	01-May-14	42.07956	Melissa	Centre	Bruno	65.53872	49.5

**Fig. 1.5** Viewing data in *R*

## 1.5 Key Points and Further Issues

### Note:

- ✓ The scale of variables is not only a formalism but an essential framework for choosing the correct analysis methods. This is particularly relevant for association analysis (Chap. 4), statistical tests (Chap. 10), and linear regression (Chap. 11).
- ✓ Even if variables are measured on a nominal scale (i.e. if they are categorical/qualitative), we may choose to assign a number to each category of this variable. This eases the implementation of some analysis methods introduced later in this book.
- ✓ Data is usually stored in a data matrix where the rows represent the observations and the columns are variables. It can be analysed with statistical software. We use *R* (R Core Team 2016) in this book. A gentle introduction is provided in Appendix A and throughout the book. A more comprehensive introduction can be found in other books, for example in Albert and Rizzo (2012), Crawley (2013), or Ligges (2008). Even advanced books, e.g. Adler (2012) or Everitt and Hothorn (2011), can offer insights to beginners.

## 1.6 Exercises

*Exercise 1.1* Describe both the population and the observations for the following research questions:

- (a) Evaluation of the satisfaction of employees from an airline.
- (b) Description of the marks of students from an assignment.
- (c) Comparison of two drugs which deal with high blood pressure.

*Exercise 1.2* A national park conducts a study on the behaviour of their leopards. A few of the park's leopards are registered and receive a GPS device which allows measuring the position of the leopard. Use this example to describe the following concepts: population, sample, observation, value, and variable.

*Exercise 1.3* Which of the following variables are qualitative, and which are quantitative? Specify which of the quantitative variables are discrete and which are continuous:

Time to travel to work, shoe size, preferred political party, price for a canteen meal, eye colour, gender, wavelength of light, customer satisfaction on a scale from 1 to 10, delivery time for a parcel, blood type, number of goals in a hockey match, height of a child, subject line of an email.

*Exercise 1.4* Identify the scale of the following variables:

- (a) Political party voted for in an election
- (b) The difficulty of different levels in a computer game
- (c) Production time of a car
- (d) Age of turtles
- (e) Calender year
- (f) Price of a chocolate bar
- (g) Identification number of a student
- (h) Final ranking at a beauty contest
- (i) Intelligence quotient.

*Exercise 1.5* Make yourself familiar with the pizza data set from Appendix A.4.

- (a) First, browse through the introduction to *R* in Appendix A. Then, read in the data.
- (b) View the data both in the *R* data editor and in the *R* console.
- (c) Create a new data matrix which consists of the first 5 rows and first 5 variables of the data. Print this data set on the *R* console. Now, save this data set in your preferred format.
- (d) Add a new variable "NewTemperature" to the data set which converts the temperature from °C to °F.

- (e) Attach the data and list the values from the variable “NewTemperature”.
- (f) Use “?” to make yourself familiar with the following commands: `str`, `dim`, `colnames`, `names`, `nrow`, `ncol`, `head`, and `tail`. Apply these commands to the data to get more information about it.

*Exercise 1.6* Consider the research questions of describing parents’ attitudes towards immunization, what proportion of them wants immunization against chicken pox for their last-born child, and whether this proportion differs by gender and age.

- (a) Which data collection method is the most suitable one to answer the above questions: survey or experiment?
- (b) How would you capture the attitudes towards immunization in a single variable?
- (c) Which variables are needed to answer all the above questions? Describe the scale of each of them.
- (d) Reflect on what an appropriate data set would look like. Now, given this data set, try to write down the above research questions as precisely as possible.

→ Solutions to all exercises in this chapter can be found on p. [321](#)

---

# Frequency Measures and Graphical Representation of Data

# 2

In Chap. 1, we highlighted that different variables contain different levels of information. When summarizing or visualizing one or more variable(s), it is this information which determines the appropriate statistical methods to use.

Suppose we are interested in studying the employment opportunities and starting salaries of university graduates with a master's degree. Let the variable  $X$  denote the starting salaries measured in €/year. Now suppose 100 graduate students provide their initial salaries. Let us write down the salary of the first student as  $x_1$ , the salary of the second student as  $x_2$ , and so on. We therefore have 100 observations  $x_1, x_2, \dots, x_{100}$ . How can we summarize those 100 values best to extract meaningful information from them? The answer to this question depends upon several aspects like the nature of the recorded data, e.g. how many observations have been obtained (either small in number or large in number) or how the data was recorded (either exact values were obtained or the values were obtained in intervals). For example, the starting salaries may be obtained as exact values, say 51,500 €/year, 32,350 €/year, etc. Alternatively, these values could have been summarized in categories such as low income ( $<30,000$  €/year), medium income (30,000–50,000 €/year), high income (50,000–70,000 €/year), and very high income ( $>70,000$  €/year). Another approach is to ask whether the students were employed or not after graduating and record the data in terms of “yes” or “no”. It is evident that the latter classification is less detailed than the grouped income data which is less detailed than the exact data. Depending on which conceptualization of “starting salary” we use, we need to choose the approach to summarize the data, that is the 100 values relating to the 100 graduated students.

---

## 2.1 Absolute and Relative Frequencies

**Discrete Data.** Let us first consider a simple example to illustrate our notation.

*Example 2.1.1* Suppose there are ten people in a supermarket queue. Each of them is either coded as “F” (if the person is female) or “M” (if the person is male). The collected data may look like

$$\text{M, F, M, F, M, M, M, F, M, M.}$$

There are now two categories in the data: male (M) and female (F). We use  $a_1$  to refer to the male category and  $a_2$  to refer to the female category. Since there are seven male and three female students, we have 7 values in category  $a_1$ , denoted as  $n_1 = 7$ , and 3 values in category  $a_2$ , denoted as  $n_2 = 3$ . The number of observations in a particular category is called the **absolute frequency**. It follows that  $n_1 = 7$  and  $n_2 = 3$  are the absolute frequencies of  $a_1$  and  $a_2$ , respectively. Note that  $n_1 + n_2 = n = 10$ , which is the same as the total number of collected observations. We can also calculate the **relative frequencies** of  $a_1$  and  $a_2$  as  $f_1 = f(a_1) = \frac{n_1}{n} = \frac{7}{10} = 0.7 = 70\%$  and  $f_2 = f(a_2) = \frac{n_2}{n} = \frac{3}{10} = 0.3 = 30\%$ , respectively. This gives us information about the proportions of male and female customers in the queue.

We now extend these concepts to a general framework for the summary of **data on discrete variables**. Suppose there are  $k$  categories denoted as  $a_1, a_2, \dots, a_k$  with  $n_j (j = 1, 2, \dots, k)$  observations in category  $a_j$ . The **absolute frequency**  $n_j$  is defined as the number of units in the  $j$ th category  $a_j$ . The sum of absolute frequencies equals the total number of units in the data:  $\sum_{j=1}^k n_j = n$ . The **relative frequencies** of the  $j$ th class are defined as

$$f_j = f(a_j) = \frac{n_j}{n}, \quad j = 1, 2, \dots, k. \quad (2.1)$$

The relative frequencies always lie between 0 and 1 and  $\sum_{j=1}^k f_j = 1$ .

**Grouped Continuous Data.** Data on continuous variables usually has a large number ( $k$ ) of different values. Sometimes  $k$  may even be the same as  $n$  and in such a case the relative frequencies become  $f_j = \frac{1}{n}$  for all  $j$ . However, it is possible to define intervals in which the observed values are contained.

*Example 2.1.2* Consider the following  $n = 20$  results of the written part of a driving licence examination (a maximum of 100 points could be achieved):

$$28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.$$

We can summarize the results in class intervals such as 0–20, 21–40, 41–60, 61–80, and 81–100, and the data can be presented as follows:

Class intervals	0–20	21–40	41–60	61–80	81–100
Absolute frequencies	$n_1 = 1$	$n_2 = 3$	$n_3 = 3$	$n_4 = 9$	$n_5 = 4$
Relative frequencies	$f_1 = \frac{1}{20}$	$f_2 = \frac{3}{20}$	$f_3 = \frac{3}{20}$	$f_4 = \frac{9}{20}$	$f_5 = \frac{5}{20}$

We have  $\sum_{j=1}^5 n_j = 20 = n$  and  $\sum_{j=1}^5 f_j = 1$ .

**Table 2.1** Frequency distribution for discrete data

Class intervals ( $a_j$ )	$a_1$	$a_2$	...	$a_k$
Absolute frequencies ( $n_j$ )	$n_1$	$n_2$	...	$n_k$
Relative frequencies ( $f_j$ )	$f_1$	$f_2$	...	$f_k$

Now, suppose the  $n$  observations can be classified into  $k$  class intervals  $a_1, a_2, \dots, a_k$ , where  $a_j (j = 1, 2, \dots, k)$  contains  $n_j$  observations with  $\sum_{j=1}^k n_j = n$ . The relative frequency of the  $j$ th class is  $f_j = n_j/n$  and  $\sum_{j=1}^k f_j = 1$ . Table 2.1 displays the **frequency distribution** of a discrete variable  $X$ .

*Example 2.1.3* Consider the pizza delivery service data (Example 1.4.2, Appendix A.4). We are interested in the pizza deliveries by branch and generate the respective frequency table, showing the distribution of the data, using the `table` command in R (after reading in and attaching the data) as

```
table(branch)          # absolute frequencies
table(branch)/length(branch) # relative frequencies
```



$a_j$	Centre	East	West
$n_j$	421	410	435
$f_j$	$\frac{421}{1266} \approx 0.333$	$\frac{410}{1266} \approx 0.323$	$\frac{435}{1266} \approx 0.344$

We have  $n = \sum_j n_j = 1266$  deliveries and  $\sum_j f_j = 1$ . We can see from this table that each branch has a similar absolute number of pizza deliveries and each branch contributes to about one-third of the total number of deliveries.

## 2.2 Empirical Cumulative Distribution Function

Another approach to summarize and visualize the (frequency) distribution of variables is the **empirical cumulative distribution function**, often abbreviated as “ECDF”. As the name itself suggests, it gives us an idea about the cumulative relative frequencies up to a certain point. For example, say we want to know how many people scored up to 60 points in Example 2.1.2. Then, this can be calculated by adding the number of people in the class intervals 0–20, 21–40, and 41–60, which corresponds to  $n_1 + n_2 + n_3 = 1 + 3 + 3 = 7$  and is the **cumulative frequency**. If we want to know the relative frequency of people obtaining up to 60 points, we have to add the relative frequencies of the people in the class intervals 0–20, 21–40, and 41–60 as  $f_1 + f_2 + f_3 = \frac{1}{20} + \frac{3}{20} + \frac{3}{20} = \frac{7}{20}$ .

Before discussing the empirical cumulative distribution function in a more general framework, let us first understand the concept of ordered values. Suppose the values of height of four people are observed as  $x_1 = 180$  cm,  $x_2 = 160$  cm,  $x_3 = 175$  cm, and  $x_4 = 170$  cm. We arrange these values in an order, say ascending order, i.e. first the smallest value (denoted as  $x_{(1)}$ ) and lastly the largest value (denoted as  $x_{(4)}$ ). We obtain

$$\begin{aligned}x_{(1)} &= x_2 = 160 \text{ cm}, & x_{(2)} &= x_4 = 170 \text{ cm}, \\x_{(3)} &= x_3 = 175 \text{ cm}, & x_{(4)} &= x_1 = 180 \text{ cm}.\end{aligned}$$

The values  $x_{(1)}$ ,  $x_{(2)}$ ,  $x_{(3)}$ , and  $x_{(4)}$  are called **ordered values** for which  $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)}$  holds. Note that  $x_1$  is not necessarily the smallest value but  $x_{(1)}$  is necessarily the smallest value. In general, if we have  $n$  observations  $x_1, x_2, \dots, x_n$ , then the ordered data is  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ .

Consider  $n$  observations  $x_1, x_2, \dots, x_n$  of a variable  $X$ , which are arranged in ascending order as  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  (and are thus on an at least ordinal scale). The **empirical cumulative distribution function**  $F(x)$  is defined as the cumulative relative frequencies of all values  $a_j$ , which are smaller than, or equal to,  $x$ :

$$F(x) = \sum_{a_j \leq x} f(a_j). \quad (2.2)$$

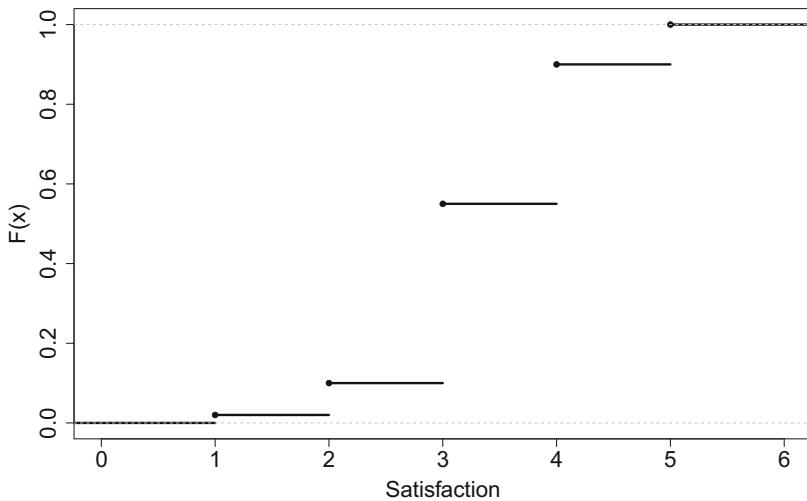
This definition implies that  $F(x)$  is a monotonically non-decreasing function,  $0 \leq F(x) \leq 1$ ,  $\lim_{x \rightarrow -\infty} F(x) = 0$  (the lower limit of  $F$  is 0),  $\lim_{x \rightarrow +\infty} F(x) = 1$  (the upper limit of  $F$  is 1), and  $F(x)$  is right continuous.

### 2.2.1 ECDF for Ordinal Variables

The empirical cumulative distribution function of ordinal variables is a **step function**.

*Example 2.2.1* Consider a customer satisfaction survey from a car service company. The 200 customers who had a car service done within the last 30 days were asked to respond regarding their overall level of satisfaction with the quality of the car service on a scale from 1 to 5 based on the following options: 1 = not satisfied at all, 2 = unsatisfied, 3 = satisfied, 4 = very satisfied, and 5 = perfectly satisfied. Based on the frequency of each option, we can calculate the relative frequencies and then plot the empirical cumulative distribution function, either manually (takes longer) or by using R (quick):

Satisfaction level ( $a_j$ )	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$n_j$	4	16	90	70	20
$f_j$	4/200	16/200	90/200	70/200	20/200
$F_j$	4/200	20/200	110/200	180/200	200/200



**Fig. 2.1** ECDF for the satisfaction survey

The  $F_j$ 's are calculated as follows:

$$\begin{aligned} F_1 &= f_1, & F_3 &= f_1 + f_2 + f_3, \\ F_2 &= f_1 + f_2, & F_4 &= f_1 + f_2 + f_3 + f_4. \end{aligned}$$

The ECDF for this data can be obtained by summarizing the data in a vector and using the `plot.ecdf()` function in R, see Fig. 2.1:

```
sv <- c(rep(1,4),rep(2,16),rep(3,90),rep(4,70),rep(5,20))
plot.ecdf(sv)
```

R

The ECDF can be used to obtain the relative frequencies for values contained in certain intervals as

$$H(c \leq x \leq d) = \text{relative frequency of values } x \text{ with } c \leq x \leq d.$$

It further follows that:

$$H(x \leq a_j) = F(a_j) \quad (2.3)$$

$$H(x < a_j) = H(x \leq a_j) - f(a_j) = F(a_j) - f(a_j) \quad (2.4)$$

$$H(x > a_j) = 1 - H(x \leq a_j) = 1 - F(a_j) \quad (2.5)$$

$$H(x \geq a_j) = 1 - H(X < a_j) = 1 - F(a_j) + f(a_j) \quad (2.6)$$

$$H(a_{j_1} \leq x \leq a_{j_2}) = F(a_{j_2}) - F(a_{j_1}) + f(a_{j_1}) \quad (2.7)$$

$$H(a_{j_1} < x \leq a_{j_2}) = F(a_{j_2}) - F(a_{j_1}) \quad (2.8)$$

$$H(a_{j_1} < x < a_{j_2}) = F(a_{j_2}) - F(a_{j_1}) - f(a_{j_2}) \quad (2.9)$$

$$H(a_{j_1} \leq x < a_{j_2}) = F(a_{j_2}) - F(a_{j_1}) - f(a_{j_2}) + f(a_{j_1}) \quad (2.10)$$

*Example 2.2.2* Suppose, in Example 2.2.1, we want to know how many customers are not satisfied with their car service. Then, using the data relating to the responses “1” and “2”, we observe from the ECDF that  $(16 + 4)/200\% = 10\%$  of the customers were not satisfied with the car service. This relates to using rule (2.3):  $H(X \leq 2) = F(2) = 0.1$  or 10 %. Similarly, the proportion of customers who are more than satisfied can be obtained using (2.5) as  $H(X > 3) = 1 - H(x \leq 3) = 1 - 110/200 = 0.45\%$  or 45 %.

## 2.2.2 ECDF for Continuous Variables

In general, we can apply formulae (2.2)–(2.10) to continuous data as well. However, before demonstrating their use, let us consider a somewhat different setting. Let us assume that a continuous variable of interest is only available in the form of grouped data. We may assume that the observations within each group, i.e. each category or each interval, are distributed uniformly over the entire interval. The ECDF then consists of straight lines connecting the lower and upper values of the ECDF in each of the intervals. To understand this concept in more detail, we introduce the following notation:

$k$	number of groups (or intervals),
$e_{j-1}$	lower limit of $j$ th interval,
$e_j$	upper limit of $j$ th interval,
$d_j = e_j - e_{j-1}$	width of the $j$ th interval,
$n_j$	number of observations in the $j$ th interval.

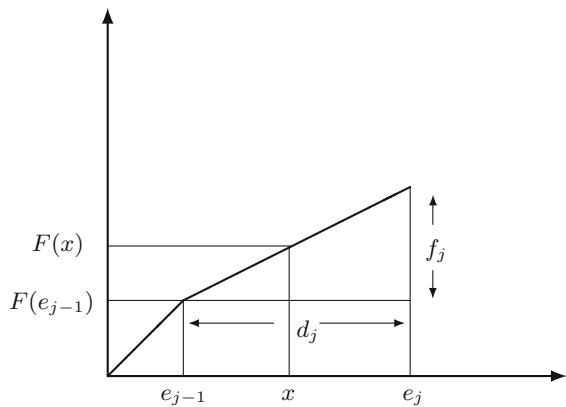
Under the assumption that all values in a particular interval are distributed uniformly within this interval, the empirical cumulative distribution function relates to a **polygonal chain** connecting the points  $(0, 0)$ ,  $(e_1, F(e_1))$ ,  $(e_2, F(e_2))$ , ...,  $(e_k, 1)$ . The ECDF can then be defined as

$$F(x) = \begin{cases} 0, & x < e_0 \\ F(e_{j-1}) + \frac{f_j}{d_j}(x - e_{j-1}), & x \in [e_{j-1}, e_j] \\ 1, & x \geq e_k \end{cases} \quad (2.11)$$

with  $F(e_0) = 0$ . The idea behind (2.11) is presented in Fig. 2.2. For any interval  $[e_{j-1}, e_j]$ , the respective lower and upper limits of the ECDF are  $F(e_j)$  and  $F(e_{j-1})$ . If we assume values to be distributed uniformly over this interval, we can connect  $F(e_j)$  and  $F(e_{j-1})$  with a straight line. To obtain  $F(x)$  with  $x > e_{j-1}$  and  $x < e_j$ , we simply add the height of the ECDF between  $F(e_{j-1})$  and  $F(x)$  to  $F(e_{j-1})$ .

*Example 2.2.3* Consider Example 2.1.3 of the pizza delivery service. Suppose we are interested in determining the distribution of the pizza delivery times. Using the function `plot.ecdf()` in R, we obtain the ECDF of the continuous data, see Fig. 2.3a. Note that the structure of the curve is a step function but now almost looks like a continuous curve. The reason for this is that when the number of observations is large, then the lengths of class intervals become small. When these small lengths are

**Fig. 2.2** Illustration of the ECDF for continuous data available in groups/intervals\*



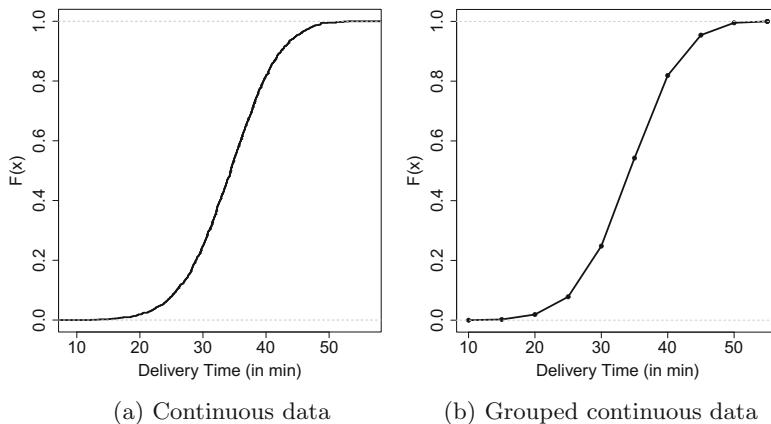
joined together, they appear like a continuous curve. As the number of observations increases, the smoothness of the curve increases too. If the number of observations is not large, e.g. suppose the data is reported as a summary from the drivers, i.e. whether the delivery took <15 min, between 15 and 20 min, between 20 and 25 min, and so on, then we can construct the ECDF by creating a table summarizing the data features as in Table 2.2.

Figure 2.3b shows the ECDF based on the grouped data evaluated in Table 2.2. It is interesting to see that the graphs emerging from the use of the grouped data and ungrouped data are similar in this specific example.

Suppose we are interested in calculating how many deliveries were completed within the desired time limit of 30 min, with a tolerance of maximum 10 % deviation, i.e. a deviation of 3 min. We can evaluate the ECDF at  $x = 33$  min.

**Table 2.2** The values needed to calculate the ECDF for the grouped pizza delivery time data in Example 2.2.3

Delivery time	$j$	$e_{j-1}$	$e_j$	$n_j$	$f_j$	$F(e_j)$
[0; 10]	1	0	10	0	0.0000	0.0000
(10; 15]	2	10	15	3	0.0024	0.0024
(15; 20]	3	15	20	21	0.0166	0.0190
(20; 25]	4	20	25	75	0.0592	0.0782
(25; 30]	5	25	30	215	0.1698	0.2480
(30; 35]	6	30	35	373	0.2946	0.5426
(35; 40]	7	35	40	350	0.2765	0.8191
(40; 45]	8	40	45	171	0.1351	0.9542
(45; 50]	9	45	50	52	0.0411	0.9953
(50; 55]	10	50	55	6	0.0047	1.0000



**Fig. 2.3** Empirical cumulative distribution function for pizza delivery time

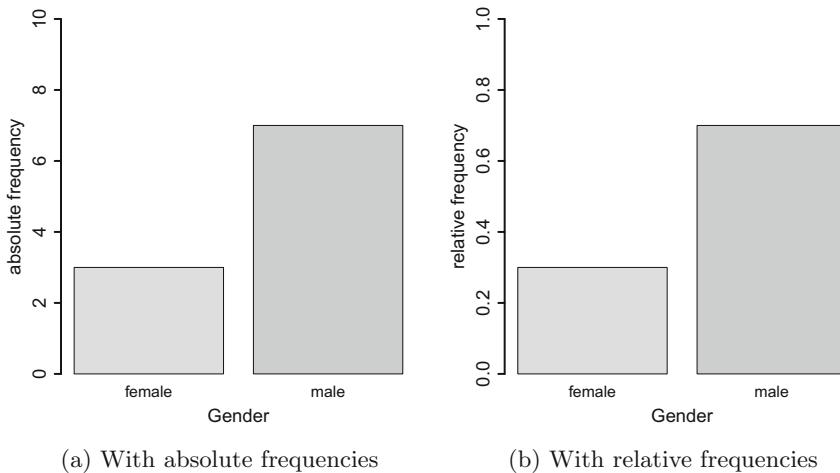
Based on (2.11), we calculate  $H(X \leq 33) = F(33) = F(30) + f(6)/5(33 - 30) = 0.2480 + 0.2946/5 \cdot 3 = 0.42476$ . Thus, we conclude, based on the grouped data, that only about 42 % of the deliveries were completed in the desired time frame.

## 2.3 Graphical Representation of a Variable

Frequency tables and empirical cumulative distribution functions are useful in providing a numerical summary of a variable. Graphs are an alternative way to summarize a variable's information. In many situations, they have the advantage of conveying the information hidden in the data more compactly. Similarly, someone's mood can be more easily understood when looking at a smiley ☺ than by reading an essay about one's mood in a long paragraph.

### 2.3.1 Bar Chart

A simple tool to visualize the relative or absolute frequencies of observed values of a variable is a **bar chart**. A bar chart can be used for nominal and ordinal variables, as long as the number of categories is not very large. It consists of one bar for each category. The height of each bar is determined by either the absolute frequency or the relative frequency of the respective category and is shown on the y-axis. If the variable is measured on an ordinal level, then it is recommended to arrange the bars on the x-axis according to their ranks or values. If the number of categories is large, then the number of bars will be large too and the bar chart, in turn, may not remain informative.

**Fig. 2.4** Bar charts

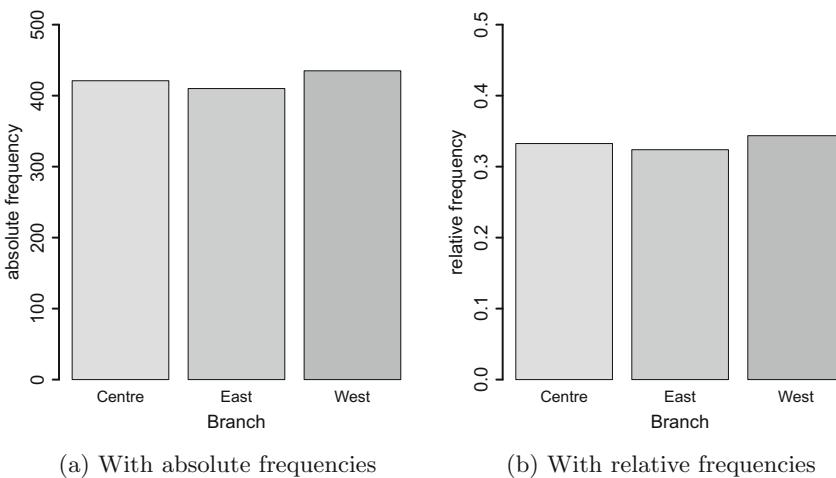
*Example 2.3.1* Consider Example 2.1.1 in which ten people, queueing in a supermarket, were classified as being either male (M) or female (F). The absolute frequencies for males and females are  $n_1 = 7$  and  $n_2 = 3$ , respectively. Since there are two categories, M and F, two bars are needed to construct the chart—one for the male category and another for the female category. The heights of the bars are determined as either  $n_1 = 7$  and  $n_2 = 3$  or  $f_1 = 0.7$  and  $f_2 = 0.3$ . These graphs are shown in Fig. 2.4.

*Example 2.3.2* Consider the data in Example 2.1.3, where the pizza delivery times for each branch are recorded over a period of 1 month. The frequency table forms the basis for the bar chart, either using the absolute or relative frequencies on the y-axis. Figure 2.5 shows the bar charts for the number and proportion of pizza deliveries per branch. The graphs can be produced in R by applying the `barplot` command to a frequency table:

```
barplot(table(branch))
barplot(table(branch)/length(branch))
```



*Remark 2.3.1* Instead of vertical bars, horizontal bars can be drawn using the optional argument `horiz=TRUE` in the `barplot` command.



**Fig. 2.5** Bar charts for the pizza deliveries per branch

### 2.3.2 Pie Chart

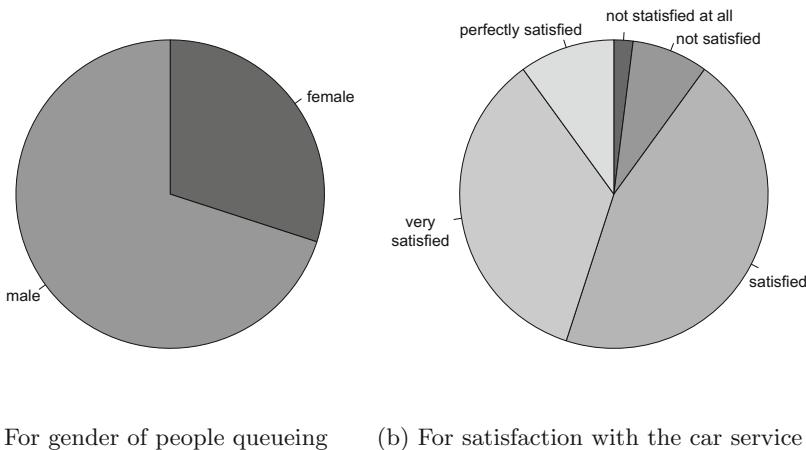
Pie charts are another option to visualize the absolute and relative frequencies of nominal and ordinal variables. A pie chart is a circle partitioned into segments, where each of the segments represents a category. The size of each segment depends upon the relative frequency and is determined by the angle  $f_j \cdot 360^\circ$ .

*Example 2.3.3* To illustrate the construction of a pie chart, let us consider again Example 2.1.1 in which ten people in a supermarket queue were classified as being either male (M) or female (F): M, F, M, F, M, M, M, F, M, M. The pie chart for this data will have two segments: one for males and another one for females. The relative frequencies are  $f_1 = 7/10$  and  $f_2 = 3/10$ , respectively. The size of the segment for the first category (M) is  $f_1 \cdot 360^\circ = (7/10) \cdot 360^\circ = 252^\circ$ , and the size of the segment for the second category (F) is  $f_2 \cdot 360^\circ = (3/10) \cdot 360^\circ = 108^\circ$ . The pie chart is shown in Fig. 2.6a.

*Example 2.3.4* Consider again Example 2.2.1, where 200 customers were asked about their level of satisfaction (5 categories) with their car service. The pie chart for this example consists of five segments representing the categories 1, 2, 3, 4, and 5. The size of the  $j$ th segment is  $f_j \cdot 360^\circ$ ,  $j = 1, 2, 3, 4, 5$ . For example, for category 1, there are 4 out of 200 customers, who are not satisfied at all. The angle of the segment “not satisfied at all” therefore is  $f_1 \cdot 360^\circ = 4/200 \cdot 360^\circ = 7.2^\circ$ . Similarly, we can calculate the angle of the other segments and obtain a pie chart as shown in Fig. 2.6b using the `pie` command in R

```
pie(table(sv))
```



**Fig. 2.6** Pie charts

*Remark 2.3.2* Note that the area of a segment is *not* proportional to the absolute frequency of the respective category. Instead, the area of the segment is proportional to the angle  $f_j \cdot 360^\circ$  (and depends also on the radius of the whole circle). It has been argued that this may cause improper interpretations as the human eye may catch the segment's area more easily than the angle of a segment. Pie charts should therefore be used with caution.

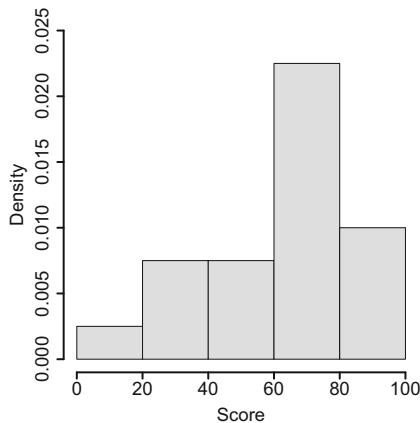
### 2.3.3 Histogram

If a variable consists of a large number of different values, the number of categories used to construct bar charts will consequently be large too. A bar chart may thus not give a clear summary when applied to a continuous variable. Instead, a **histogram** is the appropriate choice to represent the distribution of values of continuous variables. It is based on the idea to categorize the data into different groups and plot the bars for each category with height  $h_j = f_j/d_j$ , where  $d_j = e_j - e_{j-1}$  denotes the width of the  $j$ 'th class interval or category. An important consideration for this concept is that the area of the bars (=height  $\times$  width) is proportional to the relative frequency. This means that the widths of the bars need not necessarily to be the same because different widths can be adjusted with different heights of the bars.

*Example 2.3.5* Consider Example 2.1.2, where  $n = 20$  people were divided into five class intervals 0–20, 21–40, 41–60, 61–80, and 81–100 based on their performance in a written driving licence examination. The frequency table is given as

Class intervals	0–20	21–40	41–60	61–80	81–100
Absolute freq	$n_1 = 1$	$n_2 = 3$	$n_3 = 3$	$n_4 = 9$	$n_5 = 4$
Relative freq	$f_1 = \frac{1}{20}$	$f_2 = \frac{3}{20}$	$f_3 = \frac{3}{20}$	$f_4 = \frac{9}{20}$	$f_5 = \frac{5}{20}$
Height $f_j/d_j$	$h_1 = \frac{1}{400}$	$h_2 = \frac{3}{400}$	$h_3 = \frac{3}{400}$	$h_4 = \frac{9}{400}$	$h_5 = \frac{4}{400}$

**Fig. 2.7** Histogram for the scores of the people



The histogram for this grouped data set has five categories and therefore it has five bars. Since the widths of class intervals are the same, the heights of the bars are proportional to the relative frequency of the respective category. The resulting histogram is displayed in Fig. 2.7.

*Example 2.3.6* Recall Example 2.2.3 and the variable “pizza delivery time”. Table 2.3 shows the summary of the grouped data and the values needed to calculate the histogram. Figure 2.8a shows the histogram with equal widths of delivery time intervals. We see a symmetric distribution of the pizza delivery times, but many delivery times exceeding the target time of 30 min. If the histogram is required to have different widths for different bars, i.e. different delivery time intervals for different categories, then it can also be constructed as shown in Fig. 2.8b. This representation is different from Fig. 2.8a. The following commands in R are used to construct the histograms for absolute and relative frequencies, respectively:

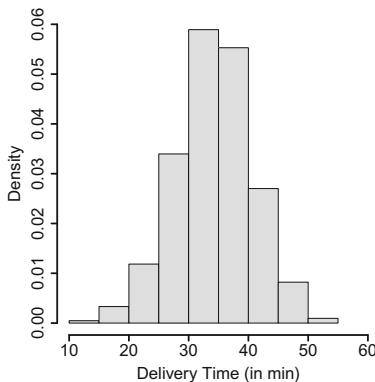
```
hist(time)          # show abs. frequencies
hist(time, freq=F) # show rel. frequencies
```



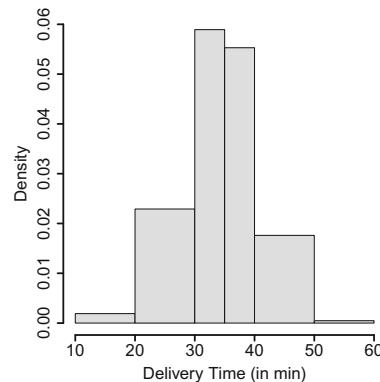
*Remark 2.3.3* The R command `truehist()` from the library MASS presents an alternative to the `hist()` command. The default specifications are somewhat different, and many users prefer it to the command `hist`.

**Table 2.3** Values needed to calculate the histogram for the grouped pizza delivery time data

Delivery time	$j$	$e_{j-1}$	$e_j$	$d_j$	$f_j$	$h_j$
[0; 10]	1	0	10	10	0.0000	0.00000
(10; 15]	2	10	15	5	0.0024	0.00047
(15; 20]	3	15	20	5	0.0166	0.00332
(20; 25]	4	20	25	5	0.0592	0.01185
(25; 30]	5	25	30	5	0.1698	0.03397
(30; 35]	6	30	35	5	0.2946	0.05893
(35; 40]	7	35	40	5	0.2765	0.05529
(40; 45]	8	40	45	5	0.1351	0.02701
(45; 50]	9	45	50	5	0.0411	0.00821
(50; 55]	10	50	55	5	0.0047	0.00094



(a) With same width for each category

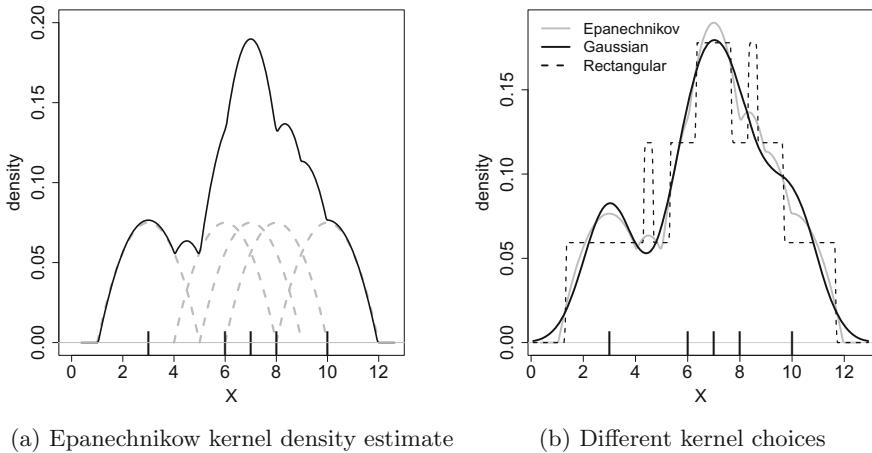


(b) With different widths per category

**Fig. 2.8** Histogram for pizza delivery time

## 2.4 Kernel Density Plots

A disadvantage of histograms is that continuous data is categorized artificially. The choice of the class intervals is crucial for the final look of the graph. A more elegant way to deal with this problem is to smooth the histogram in the sense that each observation may contribute to different classes with different weights, and the distribution is represented by a continuous function rather than a step function. A **kernel density plot** can be produced by using the following function:



**Fig. 2.9** Construction of kernel density plots

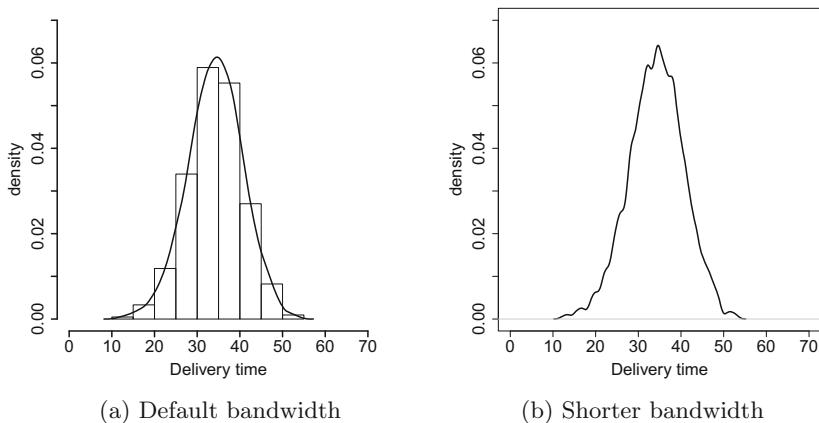
$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad h > 0, \quad (2.12)$$

where  $n$  is the sample size,  $h$  is the bandwidth, and  $K$  is a kernel function, for example

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } -1 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases} \quad (\text{rectangular kernel})$$

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{if } |x| < 1 \\ 0 & \text{elsewhere.} \end{cases} \quad (\text{Epanechnikov kernel})$$

To better understand this concept, consider Fig. 2.9a. The tick marks on the  $x$ -axis represent five observations: 3, 6, 7, 8, and 10. On each observation  $x_i$  as well as its surrounding values, we apply a kernel function, which is the Epanechnikov kernel in the figure. This means that we have five functions (grey, dashed lines), which refer to the five observations. These functions are largest at the observation itself and become gradually smaller as the distance from the observation increases. Summing up the functions, as described in Eq. (2.12), yields the solid black line, which is the kernel density plot of the five observations. It is a smooth curve, which represents the data distribution. The degree of smoothness can be controlled by the bandwidth  $h$ , which is chosen as 2 in Fig. 2.9a.



**Fig. 2.10** Kernel density plot for delivery time

The choice of the kernel may affect the overall look of the plot. Above, we have given the functions for the rectangular and Epanechnikov kernels. However, another common function for kernel density plots is the normal distribution function, which is introduced in Sect. 8.2.2, see Fig. 2.9b for a comparison of different kernels. The kernel which is based on the normal distribution is called the “Gaussian kernel” and is the default in R, where a kernel density plot can be produced combining the `plot` and `density` commands:

```
example <- c(3,6,7,8,10)
plot(density(example, kernel='gaussian'))
```



Please note that kernel functions are not defined arbitrarily and need to satisfy certain conditions, such as those required for probability density functions as explained in Chap. 7, Theorem 7.2.1.

*Example 2.4.1* Let us consider the pizza data which we introduced earlier and in Appendix A.4. We can summarize the delivery time by using a kernel density plot using the R command `plot(density(time))` and compare it with a histogram, see Fig. 2.10a. We see that the delivery times are symmetric around 35 min. If we shorten the bandwidth to a half of the default bandwidth (option `adjust=0.5`), the kernel density plot becomes more wiggly, which is illustrated in Fig. 2.10b.

## 2.5 Key Points and Further Issues

### Note:

- ✓ Bar charts and histograms are not the same graphical tools. Bar charts visualize the categories of nominal or ordinal variables whereas histograms visualize the distribution of continuous variables. A bar chart does not require to have ordered values on the  $x$ -axis, but a histogram always requires the values on the  $x$ -axis to be on a continuous scale and to be ordered. The interpretation of a histogram is simplified if the class intervals are equally sized, since then the heights of the rectangles of the histogram are proportional to the absolute or relative frequencies.
- ✓ The ECDF can be used only for ordinal and continuous variables, see Sect. 7.2 for the theoretical background of the cumulative distribution function.
- ✓ A pie chart summarizes observations from a discrete (nominal, ordinal or grouped continuous) variable. It is only useful if the number of different values (categories) is small. It is to be kept in mind that the area of each segment is not proportional to the absolute frequency of the respective category. The angle of the segment is proportional to the relative frequency of the respective category.
- ✓ Other possibilities to visualize the distribution of variables are, for example, box plots (Sect. 3.3) and stratified plots (Sects. 4.1.3, 4.3.1, and 4.4).

---

## 2.6 Exercises

*Exercise 2.1* Consider the results of the national elections in South Africa in 2014 and 2009:

Party	Results 2014 (%)	Results 2009 (%)
ANC (African National Congress)	62.15	65.90
DA (Democratic Alliance)	22.23	16.66
EFF (Economic Freedom Fighters)	6.35	–
IFP (Inkatha Freedom Party)	2.40	4.55
COPE (Congress of the People)	0.67	7.42
Others	6.20	5.47

- Summarize the results of the 2014 elections in a bar chart. Do it manually and by using *R*.
- How would you compare the results of the 2009 and 2014 elections? Offer a simple solution that can be represented in a single plot. Construct this plot in *R*.

*Exercise 2.2* Consider a variable  $X$  describing the time until the first goal was scored in the matches of the 2006 football World Cup competition. Only matches with at least one goal are considered, and goals during the  $x$ th minute of extra time are denoted as  $90 + x$ :

6	24	90+1	8	4	25	3	83	89	34	25	24	18	6
23	10	28	4	63	6	60	5	40	2	22	26	23	26
44	49	34	2	33	9	16	55	23	13	23	4	8	26
70	4	6	60	23	90+5	28	49	6	57	33	56	7	

- (a) What is the scale of  $X$ ?
- (b) Write down the frequency table of  $X$  based on the following categories:  $[0, 15)$ ,  $[15, 30)$ ,  $[30, 45)$ ,  $[45, 60)$ ,  $[60, 75)$ ,  $[75, 90)$ ,  $[90, 96)$ .
- (c) Draw the histogram for  $X$  with intervals relating to the groups from the frequency table.
- (d) Now use  $R$  to reproduce the histogram. Compare the histogram to a kernel density plot of your choice.
- (e) Calculate the empirical cumulative distribution function for the grouped data.
- (f) Use  $R$  to plot the ECDF (via a step function) for
  - (i) the original data and
  - (ii) the grouped data.
- (g) Consider the grouped data. Now assume that the values within each interval are distributed uniformly. Determine the proportion of first goals which occurred
  - (i) in the first half, i.e. during the first 45 min,
  - (ii) in the last 10 min or during the extra time,
  - (iii) between the 20th and 65th min, i.e. what is  $H(20 \leq X \leq 65)$ ?
- (h) Determine the time point at which in 80 % of the matches the first goal was scored at or before this time point.

*Exercise 2.3* Suppose we have the following information to construct a histogram for a continuous variable with 2000 observations:

$j$	$e_{j-1}$	$e_j$	$d_j$	$h_j$
1	0	1	1	0.125
2	1	4	3	0.125
3	4	7	3	0.125
4	7	8	1	0.125

- (a) Determine the relative frequencies for each interval.
- (b) Determine the absolute frequencies.

*Exercise 2.4* A university survey was conducted on 500 first-year students to obtain knowledge about the size of their accommodation (in square metres).

$j$	Size of accommodation ( $\text{m}^2$ ) $e_{j-1} \leq x \leq e_j$	$F(x)$
1	8–14	0.25
2	14–22	0.40
3	22–34	0.75
4	34–50	0.97
5	50–82	1.00

- (a) Determine the absolute frequencies for each category.
- (b) What proportion of people live in a flat of at least  $34 \text{ m}^2$ ?

*Exercise 2.5* Consider a survey in which 100 people were asked to rate on a scale from 1 to 10 how much they agree with the statement that “there is too much football on television”. The results are summarized below:

Score	0	1	2	3	4	5	6	7	8	9	10
Responses	0	1	3	8	8	27	30	11	6	4	2

- (a) Calculate and draw the ECDF of the scores.
- (b) Determine  $F(3)$  and  $F(9)$ .
- (c) Consider the situation, where the data is summarized in the two categories “disagree” ( $\text{score} \leq 5$ ) and “agree” ( $\text{score} > 5$ ). What would the ECDF look like under the approach outlined in (2.11)? Determine  $F(3)$  and  $F(9)$  for the summarized data.
- (d) Explain the differences between (b) and (c).

*Exercise 2.6* It is possible to produce professional graphics in *R*. However, it is advantageous to go beyond the default options. To demonstrate this, consider Example 2.1.3 about the pizza delivery data, which is described in Appendix A.4.

- (a) Set the working directory in *R* (`setwd()`), read in the data (`read.csv()`), and attach the data. Draw a histogram of the variable “temperature”. Type `?hist`, and view the options. Adjust the histogram so that you are satisfied with (i) axes labelling, (ii) axes range, and (iii) colour. Now use the `lines()` command to add a dashed vertical line at  $65^\circ\text{C}$  (which is the minimum temperature the pizza should have at the time of delivery).
- (b) Consider a different approach, which constructs plots by means of multiple layers using `ggplot2`. You need an Internet connection to install the package using the command `install.packages('ggplot2')`. Browse through the help

pages on <http://docs.ggplot2.org/current/>. Look specifically at the examples for `ggplot`, `qplot`, `scale_histogram`, and `scale_y_continuous`. Try to understand the roles of “aesthetics” and “geoms”. Now, after loading the library via `library(ggplot2)`, create a `ggplot` object for the pizza data, which declares “temperature” to be the  $x$ -variable. Now add a layer with `geom_histogram` to create a histogram with interval width of 2.5 and dark grey bars which are 50 % transparent. Change the  $y$ -axis labelling by adding the relevant layer using `scale_y_continuous`. Plot the graph.

- (c) Now create a normal bar chart for the variable “driver” in *R*. Type `?barplot` and `?par` to see the options one can pass on to `barchart()` to adjust the graph. Make the graph look good.
- (d) Now create the same bar chart with `ggplot2`. Use `qplot` instead of `ggplot` to create the plot. Use an option which makes each bar to consist of segments relating to the day of delivery, so that one can see the number of deliveries by driver to highlight during which days the drivers delivered most often. Browse through “themes” and “scales” on the help page, and add layers that make the background black and white and the bars on a grey scale.

→ Solutions to all exercises in this chapter can be found on p. [325](#)

\*Source Toutenburg, H., Heumann, C., *Deskriptive Statistik*, 7th edition, 2009, Springer, Heidelberg

---

# Measures of Central Tendency and Dispersion

# 3

A data set may contain many variables and observations. However, we are not always interested in each of the measured values but rather in a summary which interprets the data. Statistical functions fulfil the purpose of summarizing the data in a meaningful yet concise way.

*Example 3.0.1* Suppose someone from Munich (Germany) plans a holiday in Bangkok (Thailand) during the month of December and would like to get information about the weather when preparing for the trip. Suppose last year's maximum temperatures during the day (in degrees Celsius) for December 1–31 are as follows:

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,  
25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

How do we draw conclusions from this data? Looking at the individual values gives us a feeling about the temperatures one can experience in Bangkok, but it does not provide us with a clear summary. It is evident that the average of these 31 values as “Sum of all values/Total number of observations”  $(22 + 24 + \dots + 28 + 29)/31 = 26.48$  is meaningful in the sense that we know what temperature to expect “on average”. To choose the right clothing for the holidays, we may also be interested in knowing the temperature range to understand the variability in temperature, which is between 21 and 31 °C. Summarizing 31 individual values with only three numbers (26.48, 21, and 31) will provide sufficient information to plan the holidays.

In this chapter, we focus on the most important statistical concepts to summarize data: these are measures of central tendency and variability. The applications of each measure depend on the scale of the variable of interest, see Appendix D.1 for a detailed summary.

### 3.1 Measures of Central Tendency

A natural human tendency is to make comparisons with the “average”. For example, a student scoring 40 % in an examination will be happy with the result if the average score of the class is 25 %. If the average class score is 90 %, then the student may not feel happy even if he got 70 % right. Some other examples of the use of “average” values in common life are mean body height, mean temperature in July in some town, the most often selected study subject, the most popular TV show in 2015, and average income. Various statistical concepts refer to the “average” of the data, but the right choice depends upon the nature and scale of the data as well as the objective of the study. We call statistical functions which describe the average or centre of the data **location parameters** or **measures of central tendency**.

#### 3.1.1 Arithmetic Mean

The **arithmetic mean** is one of the most intuitive measures of central tendency. Suppose a variable of size  $n$  consists of the values  $x_1, x_2, \dots, x_n$ . The arithmetic mean of this data is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (3.1)$$

In informal language, we often speak of “the average” or just “the mean” when using the formula (3.1).

To calculate the arithmetic mean for grouped data, we need the following frequency table:

Class intervals $a_j$	$a_1 = e_0 - e_1$	$a_2 = e_1 - e_2$	...	$a_k = e_{k-1} - e_k$
Absolute freq. $n_j$	$n_1$	$n_2$	...	$n_k$
Relative freq. $f_j$	$f_1$	$f_2$	...	$f_k$

Note that  $a_1, a_2, \dots, a_k$  are the  $k$  class intervals and each interval  $a_j$  ( $j = 1, 2, \dots, k$ ) contains  $n_j$  observations with  $\sum_{j=1}^k n_j = n$ . The relative frequency of the  $j$ th class is  $f_j = n_j/n$  and  $\sum_{j=1}^k f_j = 1$ . The mid-value of the  $j$ th class interval is defined as  $m_j = (e_{j-1} + e_j)/2$ , which is the mean of the lower and upper limits of the interval. The **weighted arithmetic mean** for grouped data is defined as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k n_j m_j = \sum_{j=1}^k f_j m_j. \quad (3.2)$$

*Example 3.1.1* Consider again Example 3.0.1 where we looked at the temperature in Bangkok during December. The measurements were

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,  
25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

The arithmetic mean is therefore

$$\bar{x} = \frac{22 + 24 + 21 + \dots + 28 + 29}{31} = 26.48^{\circ}\text{C}.$$

In R, the arithmetic mean can be calculated using the `mean` command:

```
weather <- c(22,24,21,,30,28,29)
mean(weather)
[1] 26.48387
```



Let us assume the data in Example 3.0.1 is summarized in categories as follows:

Class intervals	< 20	(20 – 25]	(25, 30]	(30, 35]	> 35
Absolute frequencies	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
Relative frequencies	$f_1 = 0$	$f_2 = \frac{12}{31}$	$f_3 = \frac{18}{31}$	$f_4 = \frac{1}{31}$	$f_5 = 0$

We can calculate the (weighted) arithmetic mean as

$$\bar{x} = \sum_{j=1}^k f_j m_j = 0 + \frac{12}{31} \cdot 22.5 + \frac{18}{31} \cdot 27.5 + \frac{1}{31} \cdot 32.5 + 0 \approx 25.7.$$

In R, we use the `weighted.mean` function to obtain the result. The function requires to specify the (hypothesized) means for each group, for example the middle values of the class intervals, as well as the weights.

```
weighted.mean(c(22.5,27.5,32.5),c(12/31,18/31,1/31))
```



Interestingly, the results of the mean and the weighted mean differ. This is because we use the middle of each class as an approximation of the mean within the class. The implication is that we assume that the values are uniformly distributed within each interval. This assumption is obviously not met. If we had knowledge about the mean in each class, like in this example, we would obtain the correct result as follows:

$$\bar{x} = \sum_{j=1}^k f_j \bar{x}_j = 0 + \frac{12}{31} \cdot 23.83333 + \frac{18}{31} \cdot 28 + \frac{1}{31} \cdot 32.5 + 0 = 26.48387.$$

However, the weighted mean is meant to estimate the arithmetic mean in those situations where only grouped data is available. It is therefore typically used to obtain an approximation of the true mean.

### Properties of the Arithmetic Mean.

- (i) The sum of the deviations of each variable around the arithmetic mean is zero:

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0. \quad (3.3)$$

- (ii) If the data is linearly transformed as  $y_i = a + bx_i$ , where  $a$  and  $b$  are known constants, it holds that

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n (a + bx_i) = \frac{1}{n} \sum_{i=1}^n a + \frac{b}{n} \sum_{i=1}^n x_i = a + b\bar{x}. \quad (3.4)$$

*Example 3.1.2* Recall Examples 3.0.1 and 3.1.1 where we considered the temperatures in December in Bangkok. We measured them in degrees Celsius, but someone from the USA might prefer to know them in degrees Fahrenheit. With a linear transformation, we can create a new temperature variable as

$$\text{Temperature in } ^\circ\text{F} = 32 + 1.8 \text{ Temperature in } ^\circ\text{C}.$$

Using  $\bar{y} = a + b\bar{x}$ , we get  $\bar{y} = 32 + 1.8 \cdot 26.48 \approx 79.7^\circ\text{F}$ .

### 3.1.2 Median and Quantiles

The median is the value which divides the observations into two equal parts such that at least 50 % of the values are greater than or equal to the median and at least 50 % of the values are less than or equal to the median. The median is denoted by  $\tilde{x}_{0.5}$ ; then, in terms of the empirical cumulative distribution function, the condition  $F(\tilde{x}_{0.5}) = 0.5$  is satisfied. Consider the  $n$  observations  $x_1, x_2, \dots, x_n$  which can be ordered as  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . The calculation of the median depends on whether the number of observations  $n$  is odd or even. When  $n$  is odd, then  $\tilde{x}_{0.5}$  is the middle ordered value. When  $n$  is even, then  $\tilde{x}_{0.5}$  is the arithmetic mean of the two middle ordered values:

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even.} \end{cases} \quad (3.5)$$

*Example 3.1.3* Consider again Examples 3.0.1–3.1.2 where we evaluated the temperature in Bangkok in December. The ordered values  $x_{(i)}$ ,  $i = 1, 2, \dots, 31$ , are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31	
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	

We have  $n = 31$ , and therefore  $\tilde{x}_{0.5} = x_{((n+1)/2)} = x_{((31+1)/2)} = x_{(16)} = 26$ . Therefore, at least 50 % of the 31 observations are greater than or equal to 26 and at least 50 % are less than or equal to 26. If one value was missing, let us say the last observation, then the median would be calculated as  $\frac{1}{2}(x_{(30/2)} + x_{(30/2+1)}) = \frac{1}{2}(26 + 26) = 26$ . In R, we would have obtained the results using the `median` command:

`median(weather)`



If we deal with grouped data, we can calculate the median under the assumption that the values within each class are equally distributed. Let  $K_1, K_2, \dots, K_k$  be  $k$  classes with observations of size  $n_1, n_2, \dots, n_k$ , respectively. First, we need to determine which class is the median class, i.e. the class that includes the median. We define the median class as the class  $K_m$  for which

$$\sum_{j=1}^{m-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^m f_j \geq 0.5 \quad (3.6)$$

hold. Then, we can determine the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m \quad (3.7)$$

where  $e_{m-1}$  denotes the lower limit of the interval  $K_m$  and  $d_m$  is the width of the interval  $K_m$ .

*Example 3.1.4* Recall Example 3.1.1 where we looked at the grouped temperature data:

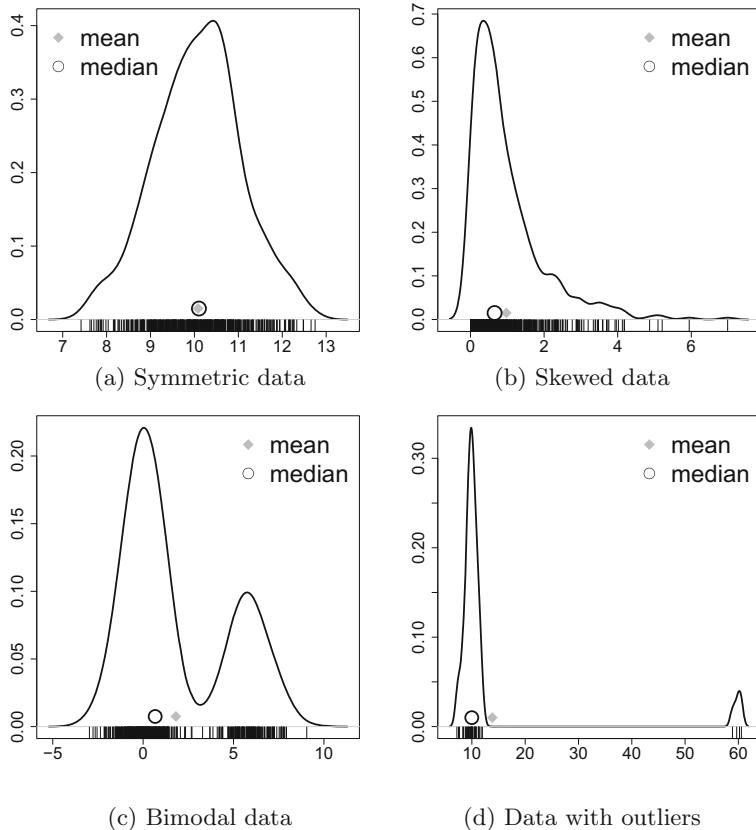
Class intervals	<20	(20–25]	(25, 30]	(30, 35]	>35
$n_j$	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
$f_j$	$f_1 = 0$	$f_2 = \frac{12}{31}$	$f_3 = \frac{18}{31}$	$f_4 = \frac{1}{31}$	$f_5 = 0$
$\sum_j f_j$	0	$\frac{12}{31}$	$\frac{30}{31}$	1	1

For the third class ( $m = 3$ ), we have

$$\sum_{j=1}^{m-1} f_j = \frac{12}{31} < 0.5 \quad \text{and} \quad \sum_{j=1}^m f_j = \frac{30}{31} \geq 0.5.$$

We can therefore calculate the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m = 25 + \frac{0.5 - \frac{12}{31}}{\frac{18}{31}} \cdot 5 \approx 25.97.$$



**Fig. 3.1** Arithmetic mean and median for different data

**Comparing the Mean with the Median.** In the above examples, the mean and the median turn out to be quite similar to each other. This is because we looked at data which is symmetrically distributed around its centre, i.e. on average, we can expect 26°C with deviations that are similar above and below the average temperature. A similar example is given in Fig. 3.1a: we see that the raw data is summarized by using ticks at the bottom of the graph and by using a kernel density estimator. The mean and the median are similar here because the distribution of the observations is symmetric around the centre. If we have skewed data (Fig. 3.1b), then the mean and the median may differ. If the data has more than one centre, such as in Fig. 3.1c, neither the median nor the mean has meaningful interpretations. If we have outliers (Fig. 3.1d), then it is wise to use the median because the mean is sensitive to outliers. These examples show that depending on the situation of interest either the mean, the median, both or neither of them can be useful.

**Quantiles.** Quantiles are a generalization of the idea of the median. The median is the value which splits the data into two equal parts. Similarly, a quantile partitions the data into other proportions. For example, a 25 %-quantile splits the data into two parts such that at least 25 % of the values are less than or equal to the quantile and at least 75 % of the values are greater than or equal to the quantile. In general, let  $\alpha$  be a number between zero and one. The  $(\alpha \times 100)\%$ -quantile, denoted as  $\tilde{x}_\alpha$ , is defined as the value which divides the data in proportions of  $(\alpha \times 100)\%$  and  $(1 - \alpha) \times 100\%$  such that at least  $\alpha \times 100\%$  of the values are less than or equal to the quantile and at least  $(1 - \alpha) \times 100\%$  of the values are greater than or equal to the quantile. In terms of the empirical cumulative distribution function, we can write  $F(\tilde{x}_\alpha) = \alpha$ . It follows immediately that for  $n$  observations, at least  $n\alpha$  values are less than or equal to  $\tilde{x}_\alpha$  and at least  $n(1 - \alpha)$  observations are greater than or equal to  $\tilde{x}_\alpha$ . The median is the 50%-quantile  $\tilde{x}_{0.5}$ . If  $\alpha$  takes the values 0.1, 0.2, ..., 0.9, the quantiles are called **deciles**. If  $\alpha \cdot 100$  is an integer number (e.g.  $\alpha \times 100 = 95$ ), the quantiles are called **percentiles**, i.e. the data is divided into 100 equal parts. If  $\alpha$  takes the values 0.2, 0.4, 0.6, and 0.8, the quantiles are known as **quintiles** and they divide the data into five equal parts. If  $\alpha$  takes the values 0.25, 0.5, and 0.75, the quantiles are called **quartiles**.

Consider  $n$  ordered observations  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . The  $\alpha \cdot 100\%$ -quantile  $\tilde{x}_\alpha$  is calculated as

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{if } n\alpha \text{ is not an integer number,} \\ & \text{choose } k \text{ as the smallest integer } > n\alpha, \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{if } n\alpha \text{ is an integer.} \end{cases} \quad (3.8)$$

*Example 3.1.5* Recall Examples 3.0.1–3.1.4 where we evaluated the temperature in Bangkok in December. The ordered values  $x_{(i)}$ ,  $i = 1, 2, \dots, 31$  are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

To determine the quartiles, i.e. the 25, 50, and 75 % quantiles, we calculate  $n\alpha$  as  $31 \cdot 0.25 = 7.75$ ,  $31 \cdot 0.5 = 15.5$ , and  $31 \cdot 0.75 = 23.25$ . Using (3.8), it follows that

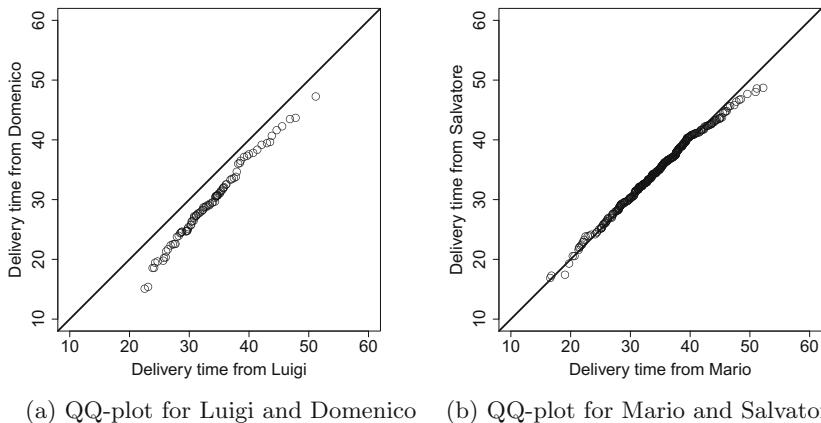
$$\begin{aligned}\tilde{x}_{0.25} &= x_{(8)} = 25, & \tilde{x}_{0.5} &= x_{(16)} = 26, \\ \tilde{x}_{0.75} &= x_{(24)} = 29.\end{aligned}$$

In R, we obtain the same results using the `quantile` function. The `probs` argument is used to specify  $\alpha$ . By default, the quartiles are reported.

```
quantile(weather)
quantile(weather, probs=c(0, 0.25, 0.5, 0.75, 1))
```



However, please note that R offers nine different ways to obtain quantiles, each of which can be chosen by the `type` argument. See Hyndman and Fan (1996) for more details.



**Fig. 3.2** QQ-plots for the pizza delivery time for different drivers

### 3.1.3 Quantile–Quantile Plots (QQ-Plots)

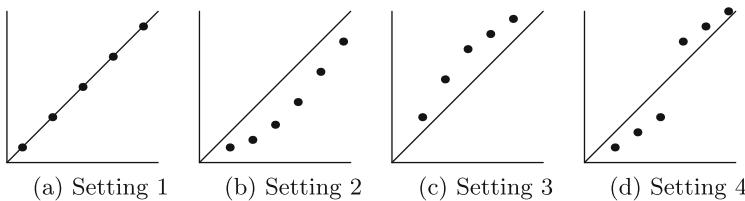
If we plot the quantiles of two variables against each other, we obtain a Quantile–Quantile plot (QQ-plot). This provides a simple summary of whether the distributions of the two variables are similar with respect to their location or not.

*Example 3.1.6* Consider again the pizza data which is described in Appendix A.4. We may be interested in the delivery time for different drivers to see if their performance is the same. Figure 3.2a shows a QQ-plot for the delivery time of driver Luigi and the delivery time of driver Domenico. Each point refers to the  $\alpha\%$  quantile of both drivers. If the point lies on the bisection line, then they are identical and we conclude that the quantiles of the both drivers are the same. If the point is below the line, then the quantile is higher for Luigi, and if the point is above the line, then the quantile is lower for Luigi. So if all the points lie exactly on the line, we can conclude that the distributions of both the drivers are the same. We see that all the reported quantiles lie below the line, which implies that all the quantiles of Luigi have higher values than those of Domenico. This means that not only on an average, but also in general, the delivery times are higher for Luigi. If we look at two other drivers, as displayed in Fig. 3.2b, the points lie very much on the bisection line. We can therefore conclude that the delivery times of these two drivers do not differ much.

In R, we can generate QQ-plots by using the `qqplot` command:

`qqplot()`





**Fig. 3.3** Different patterns for a QQ-plot

As a summary, let us consider four important patterns:

- If all the pairs of quantiles lie (nearly) on a straight line at an angle of 45 % from the  $x$ -axis, then the two samples have similar distributions (Fig. 3.3a).
- If the  $y$ -quantiles are lower than the  $x$ -quantiles, then the  $y$ -values have a tendency to be lower than the  $x$ -values (Fig. 3.3b).
- If the  $x$ -quantiles are lower than the  $y$ -quantiles, then the  $x$ -values have a tendency to be lower than the  $y$ -values (Fig. 3.3c).
- If the QQ-plot is like Fig. 3.3d, it indicates that there is a break point up to which the  $y$ -quantiles are lower than the  $x$ -quantiles and after that point, the  $y$ -quantiles are higher than the  $x$ -quantiles.

### 3.1.4 Mode

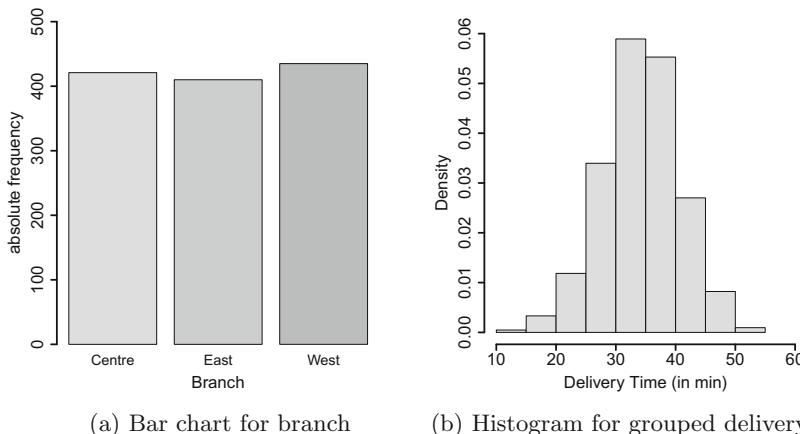
Consider a situation in which an ice cream shop owner wants to know which flavour of ice cream is the most popular among his customers. Similarly, a footwear shop owner may like to find out what design and size of shoes are in highest demand. To answer this type of questions, one can use the mode which is another measure of central tendency.

The mode  $\bar{x}_M$  of  $n$  observations  $x_1, x_2, \dots, x_n$  is the value which occurs the most compared with all other values, i.e. the value which has maximum absolute frequency. It may happen that two or more values occur with the same frequency in which case the mode is not uniquely defined. A formal definition of the mode is

$$\bar{x}_M = a_j \Leftrightarrow n_j = \max \{n_1, n_2, \dots, n_k\}. \quad (3.9)$$

The mode is typically applied to any type of variable for which the number of different values is not too large. If continuous data is summarized in groups, then the mode can be used as well.

*Example 3.1.7* Recall the pizza data set described in Appendix A.4. The pizza delivery service has three branches, in the East, West, and Centre, respectively. Suppose we want to know which branch delivers the most pizzas. We find that most of the deliveries have been made in the West, see Fig. 3.4a; therefore the mode is  $\bar{x}_M = \text{West}$ . Similarly, suppose we also want to find the mode for the categorized pizza delivery time: if we group the delivery time in intervals of 5 min, then we see that the most frequent delivery time is the interval “30–35” min, see Fig. 3.4b. The mode is therefore  $\bar{x}_M = [30, 35]$ .

**Fig. 3.4** Results from the pizza data set

### 3.1.5 Geometric Mean

Consider  $n$  observations  $x_1, x_2, \dots, x_n$  which are all positive and collected on a quantitative variable. The geometric mean  $\bar{x}_G$  of this data is defined as

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^n x_i} = \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}}. \quad (3.10)$$

The geometric mean plays an important role in fields where we are interested in products of observations, such as when we look at percentage changes in quantities. We illustrate its interpretation and use by looking at the average growth of a quantity in the sense that we allow a starting value, such as a certain amount of money or a particular population, to change over time. Suppose we have a starting value at some baseline time point 0 (zero), which may be denoted as  $B_0$ . At time  $t$ , this value may have changed and we therefore denote it as  $B_t$ ,  $t = 1, 2, \dots, T$ . The ratio of  $B_t$  and  $B_{t-1}$ ,

$$x_t = \frac{B_t}{B_{t-1}},$$

is called the  $t$ th growth factor. The growth rate  $r_t$  is defined as

$$r_t = ((x_t - 1) \cdot 100) \%$$

and gives us an idea about the growth or decline of our value at time  $t$ . We can summarize these concepts in the following table:

Time	Inventory	Growth factor	Growth rate
$t$	$B_t$	$x_t$	$r_t$
0	$B_0$	—	—
1	$B_1$	$x_1 = B_1/B_0$	$((x_1 - 1) \cdot 100) \%$
2	$B_2$	$x_2 = B_2/B_1$	$((x_2 - 1) \cdot 100) \%$
⋮	⋮	⋮	⋮
$T$	$B_T$	$x_T = B_T/B_{T-1}$	$((x_T - 1) \cdot 100) \%$

We can calculate  $B_t$  ( $t = 1, 2, \dots, T$ ) by using the growth factors:

$$B_t = B_0 \cdot x_1 \cdot x_2 \cdot \dots \cdot x_t.$$

The average growth factor from  $B_0$  to  $B_T$  is the geometric mean or geometric average of the growth factors:

$$\begin{aligned} \bar{x}_G &= \sqrt[T]{x_1 \cdot x_2 \cdot \dots \cdot x_T} \\ &= \sqrt[T]{\frac{B_0 \cdot x_1 \cdot x_2 \cdot \dots \cdot x_T}{B_0}} \\ &= \sqrt[T]{\frac{B_T}{B_0}}. \end{aligned} \quad (3.11)$$

Therefore,  $B_t$  at time  $t$  can be calculated as  $B_t = B_0 \cdot \bar{x}_G^t$ .

*Example 3.1.8* Suppose someone wants to deposit money, say €1000, in a bank. The bank advisor proposes a 5-year savings plan with the following plan for interest rates: 1 % in the first year, 1.5 % in the second year, 2.5 % in the third year, and 3 % in the last 2 years. Now he would like to calculate the average growth factor and average growth rate for the invested money. The concept of the geometric mean can be used as follows:

Year	Euro	Growth factor	Growth rate (%)
0	1000	—	—
1	1010	1.01	1.0
2	1025.15	1.015	1.5
3	1050.78	1.025	2.5
4	1082.30	1.03	3.0
5	1114.77	1.03	3.0

The geometric mean is calculated as

$$\bar{x}_G = (1.01 \cdot 1.015 \cdot 1.025 \cdot 1.03 \cdot 1.03)^{\frac{1}{5}} = 1.021968$$

which means that he will have on average about 2.2 % growth per year. The savings after 5 years can be calculated as

$$\text{€ } 1000 \cdot 1.021968^5 = \text{€ } 1114.77.$$

It is easy to compare two different saving plans with different growth strategies using the geometric mean.

### 3.1.6 Harmonic Mean

The harmonic mean is typically used whenever different  $x_i$  contribute to the mean with a different weight  $w_i$ , i.e. when we implicitly assume that the weight of each  $x_i$  is not one. It can be calculated as

$$\bar{x}_H = \frac{w_1 + w_2 + \cdots + w_k}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \cdots + \frac{w_k}{x_k}} = \frac{\sum_{i=1}^k w_i}{\sum_{i=1}^k \frac{w_i}{x_i}}. \quad (3.12)$$

For example, when calculating the average speed, each weight relates to the relative distance travelled,  $n_i/n$ , with speed  $x_i$ . Using  $w_i = n_i/n$  and  $\sum_i w_i = \sum_i n_i/n = 1$ , the harmonic mean can be written as

$$\bar{x}_H = \frac{1}{\sum_{i=1}^k \frac{w_i}{x_i}}. \quad (3.13)$$

*Example 3.1.9* Suppose an investor bought shares worth €1000 for two consecutive months. The price for a share was €50 in the first month and €200 in the second month. What is the average purchase price? The number of shares purchased in the first month is  $1000/50 = 20$ . The number of shares purchased in the second month is  $1000/200 = 5$ . The total number of shares purchased is thus  $20 + 5 = 25$ , and the total investment is €2000. It is evident that the average purchase price is  $2000/25 = €80$ . This is in fact the harmonic mean calculated as

$$\bar{x}_H = \frac{1}{\frac{0.5}{50} + \frac{0.5}{200}} = 80$$

because the weight of each purchase is  $n_i/n = 1000/2000 = 0.5$ . If the investment was €1200 in the first month and €800 in the second month, then we could use the harmonic mean with weights  $1200/2000 = 0.6$  and  $800/2000 = 0.4$ , respectively, to obtain the results.

## 3.2 Measures of Dispersion

Measures of central tendency, as introduced earlier, give us an idea about the location where most of the data is concentrated. However, two different data sets may have the same value for the measure of central tendency, say the same arithmetic means, but they may have different concentrations around the mean. In this case, the location measures may not be adequate enough to describe the distribution of the data. The concentration or dispersion of observations around any particular value is another property which characterizes the data and its distribution. We now introduce statistical methods which describe the **variability** or **dispersion** of data.

*Example 3.2.1* Suppose three students Christine, Andreas, and Sandro arrive at different times in the class to attend their lectures. Let us look at their arrival time in the class after or before the starting time of lecture, i.e. let us look how early or late they were (in minutes).

Week	1	2	3	4	5	6	7	8	9	10
Christine	0	0	0	0	0	0	0	0	0	0
Andreas	-10	+10	-10	+10	-10	+10	-10	+10	-10	+10
Sandro	3	5	6	2	4	6	8	4	5	7

We see that Christine always arrives on time (time difference of zero). Andreas arrives sometimes 10 min early and sometimes 10 min late. However, the arithmetic mean of both students is the same—on average, they both arrive on time! This interpretation is obviously not meaningful. The difference between both students is the variability in arrival times that cannot be measured with the mean or median. For this reason, we need to introduce measures of dispersion (variability). With the knowledge of both location and dispersion, we can give a much more nuanced comparison between the different arrival times. For example, consider the third student Sandro. He is always late; sometimes more, sometimes less. However, while on average he comes late, his behaviour is more predictable than that of Andreas. Both location and dispersion are needed to give a fair comparison.

*Example 3.2.2* Consider another example in which a supplier for the car industry needs to deliver 10 car doors with an exact width of 1.00 m. He supplies 5 doors with a width of 1.05 m and the remaining 5 doors with a width of 0.95 m. The arithmetic mean of all the 10 doors is 1.00 m. Based on the arithmetic mean, one may conclude that all the doors are good but the fact is that none of the doors are usable as they will not fit into the car. This knowledge can be summarized by a measure of dispersion.

The above examples highlight that the distribution of a variable needs to be characterized by a measure of dispersion in addition to a measure of location (central tendency). Now we introduce various measures of dispersion.

### 3.2.1 Range and Interquartile Range

Consider a variable  $X$  with  $n$  observations  $x_1, x_2, \dots, x_n$ . Order these  $n$  observations as  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . The range is a measure of dispersion defined as the difference between the maximum and minimum value of the data as

$$R = x_{(n)} - x_{(1)}. \quad (3.14)$$

The **interquartile range** is defined as the difference between the 75th and 25th quartiles as

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}. \quad (3.15)$$

It covers the centre of the distribution and contains 50 % of the observations.

*Remark 3.2.1* Note that the interquartile range is defined as the interval  $[\tilde{x}_{0.25}; \tilde{x}_{0.75}]$  in some literature. However, in line with most of the statistical literature, we define the interquartile range to be a measure of dispersion, i.e. the difference between  $\tilde{x}_{0.75}$  and  $\tilde{x}_{0.25}$ .

*Example 3.2.3* Recall Examples 3.0.1–3.1.5 where we looked at the temperature in Bangkok during December. The ordered values  $x_{(i)}$ ,  $i = 1, \dots, 31$ , are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

We obtained the quantiles in Example 3.1.5 as  $\tilde{x}_{0.25} = 25$  and  $\tilde{x}_{0.75} = 29$ . The interquartile range is therefore  $d_Q = 29 - 25 = 4$ , which means that 50 % of the data is centred between 25 and 29 °C. The range is  $R = 31 - 21 = 10$  °C, meaning that the temperature is varying at most by 10 °C. In R, there are several ways to obtain quartiles, minimum and maximum values, e.g. by using `min`, `max`, `quantiles`, `range`, among others. All numbers can be easily obtained by the `summary` command which we recommend using.

`summary(weather)`



### 3.2.2 Absolute Deviation, Variance, and Standard Deviation

Another measure of dispersion is the variance. The variance is one of the most important measures in statistics and is needed throughout this book. We use the idea of “absolute deviation” to give some more background and motivation for understanding the variance as a measure of dispersion, followed by some examples.

Consider the deviations of  $n$  observations around a certain value “A” and combine them together, for instance, via the arithmetic mean of all the deviations:

$$D = \frac{1}{n} \sum_{i=1}^n (x_i - A). \quad (3.16)$$

This measure has the drawback that the deviations  $(x_i - A)$ ,  $i = 1, 2, \dots, n$ , can be either positive or negative and, consequently, their sum can potentially be very small or even zero. Using  $D$  as a measure of variability is therefore not a good idea since  $D$  may be small even for a large variability in the data.

Using absolute values of the deviations solves this problem, and we introduce the following measure of dispersion:

$$D(A) = \frac{1}{n} \sum_{i=1}^n |x_i - A|. \quad (3.17)$$

It can be shown that the absolute deviation attains its minimum when  $A$  corresponds to the median of the data:

$$D(\tilde{x}_{0.5}) = \frac{1}{n} \sum_{i=1}^n |x_i - \tilde{x}_{0.5}|. \quad (3.18)$$

We call  $D(\tilde{x}_{0.5})$  the **absolute median deviation**. When  $A = \bar{x}$ , we speak of the **absolute mean deviation** given by

$$D(\bar{x}) = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|. \quad (3.19)$$

Another solution to avoid the positive and negative signs of deviation in (3.16) is to consider the squares of deviations  $x_i - A$ , rather than using the absolute value. This provides another measure of dispersion as

$$s^2(A) = \frac{1}{n} \sum_{i=1}^n (x_i - A)^2 \quad (3.20)$$

which is known as the **mean squared error** (MSE) with respect to  $A$ . The MSE is another important measure in statistics, see Chap. 9, Eq. (9.4), for details. It can be shown that  $s^2(A)$  attains its minimum value when  $A = \bar{x}$ . This is the (sample) **variance**

$$\tilde{s}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (3.21)$$

After expanding  $\tilde{s}^2$ , we can write (3.21) as

$$\tilde{s}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2. \quad (3.22)$$

The positive square root of the variance is called the (sample) **standard deviation**, defined as

$$\tilde{s} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (3.23)$$

The standard deviation has the same unit of measurement as the data whereas the unit of the variance is the square of the units of the observations. For example, if  $X$  is weight, measured in kg, then  $\bar{x}$  and  $\tilde{s}$  are also measured in kg, while  $\tilde{s}^2$  is measured in  $\text{kg}^2$  (which may be more difficult to interpret). The variance is a measure which we use in other chapters to obtain measures of association between variables and to

draw conclusions from a sample about a population of interest; however, the standard deviation is typically preferred for a descriptive summary of the dispersion of data.

The standard deviation measures how much the observations vary or how they are dispersed around the arithmetic mean. A low value of the standard deviation indicates that the values are highly concentrated around the mean. A high value of the standard deviation indicates lower concentration of the observations around the mean, and some of the observed values may even be far away from the mean. If there are extreme values or outliers in the data, then the arithmetic mean is more sensitive to outliers than the median. In such a case, the absolute median deviation (3.18) may be preferred over the standard deviation.

*Example 3.2.4* Consider again Example 3.2.1 where we evaluated the arrival times of Christine, Andreas, and Sandro in their lecture. Using the arithmetic mean, we concluded that both Andreas and Christine arrive on time, whereas Sandro is always late; however, we saw that the variation of arrival times differs substantially among the three students. To describe and quantify this variability formally, we calculate the variance and absolute median deviation:

$$\tilde{s}_C^2 = \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x})^2 = \frac{1}{10} ((0 - 0)^2 + \dots + (0 - 0)^2) = 0$$

$$\tilde{s}_A^2 = \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x})^2 = \frac{1}{10} ((-10 - 0)^2 + \dots + (10 - 0)^2) \approx 111.1$$

$$\tilde{s}_S^2 = \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x})^2 = \frac{1}{10} ((3 - 5)^2 + \dots + (7 - 5)^2) \approx 3.3$$

$$D(\tilde{x}_{0.5,C}) = \frac{1}{10} \sum_{i=1}^n |x_i - \tilde{x}_{0.5}| = |0 - 0| + \dots + |0 - 0| = 0$$

$$D(\tilde{x}_{0.5,A}) = \frac{1}{10} \sum_{i=1}^n |x_i - \tilde{x}_{0.5}| = |-10 - 0| + \dots + |10 - 0| = 10$$

$$D(\tilde{x}_{0.5,S}) = \frac{1}{10} \sum_{i=1}^n |x_i - \tilde{x}_{0.5}| = |3 - 5| + \dots + |7 - 5| = 1.4.$$

We observe that the variation/dispersion/variability is the lowest for Christine and highest for Andreas. Both median absolute deviation and variance allow a comparison between the two students. If we take the square root of the variance, we obtain the standard deviation. For example,  $\tilde{s}_S = \sqrt{3.3} \approx 1.8$ , which means that the average difference of the observations from the arithmetic mean is 1.8.

In *R*, we can use the `var` command to calculate the variance. However, note that *R* uses  $1/(n - 1)$  instead of  $1/n$  in calculating the variance. The idea behind the multiplication by  $1/(n - 1)$  in place of  $1/n$  is discussed in Chap. 9, see also Theorem 9.2.1.

**Variance for Grouped Data.** The variance for grouped data can be calculated using

$$s_b^2 = \frac{1}{n} \sum_{j=1}^k n_j (a_j - \bar{x})^2 = \frac{1}{n} \left( \sum_{j=1}^k n_j a_j^2 - n\bar{x}^2 \right) = \frac{1}{n} \sum_{j=1}^k n_j a_j^2 - \bar{x}^2, \quad (3.24)$$

where  $a_j$  is the middle value of the  $j$ th interval. However, when the data is artificially grouped and the knowledge about the original ungrouped data is available, we can also use the arithmetic mean of the  $j$ th class:

$$s_b^2 = \frac{1}{n} \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2. \quad (3.25)$$

The two expressions (3.24) and (3.25) represent the **variance between the different classes**, i.e. they describe the variability of the class specific means  $\bar{x}_j$ , weighted by the size of each class  $n_j$ , around the overall mean  $\bar{x}$ . It is evident that the variance *within* each class is not taken into account in these formulae. The variability of measurements in each class, i.e. the variability of  $\forall x_i \in K_j$ , is another important component to determine the overall variance in the data. It is therefore not surprising that using only the between variance  $s_b^2$  will underestimate the total variance and therefore

$$s_b^2 \leq s^2. \quad (3.26)$$

If the data within each class is known, we can use the Theorem of Variance Decomposition (see p. 136 for the theoretical background) to determine the variance. This allows us to represent the total variance as the sum of the **variance between the different classes** and the **variance within the different classes** as

$$\tilde{s}^2 = \underbrace{\frac{1}{n} \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2}_{\text{between}} + \underbrace{\frac{1}{n} \sum_{j=1}^k n_j \tilde{s}_j^2}_{\text{within}}. \quad (3.27)$$

In (3.27),  $\tilde{s}_j^2$  is the variance of the  $j$ th class:

$$\tilde{s}_j^2 = \frac{1}{n_j} \sum_{x_i \in K_j} (x_i - \bar{x}_j)^2. \quad (3.28)$$

The proof of (3.27) is given in Appendix C.1, p. 423.

*Example 3.2.5* Recall the weather data used in Examples 3.0.1–3.2.3 and the grouped data specified as follows:

Class intervals	<20	(20–25]	(25, 30]	(30, 35]	>35
$n_j$	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
$\bar{x}_j$	–	23.83	28	31	–
$\tilde{s}_j^2$	–	1.972	2	0	–

We know that  $\bar{x} = 26.48$  and  $n = 31$ . The first step is to calculate the mean and variances in each class using (3.28). We then obtain  $\bar{x}_j$  and  $s_j^2$  as listed above. The within and between variances are as follows:

$$\begin{aligned}\frac{1}{n} \sum_{j=1}^k n_j \tilde{s}_j^2 &= \frac{1}{31} (12 \cdot 1.972 + 18 \cdot 2 + 1 \cdot 0) \approx 1.925 \\ \frac{1}{n} \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 &= \frac{1}{31} (12 \cdot [23.83 - 26.48]^2 + 18 \cdot [28 - 26.48]^2 \\ &\quad + 1 \cdot [31 - 26.48]^2) \approx 4.71.\end{aligned}$$

The total variance is therefore  $\tilde{s}^2 \approx 6.64$ . Estimating the variance using all 31 observations would yield the same results. However, it becomes clear that without knowledge about the variance within each class, we cannot reliably estimate  $\tilde{s}^2$ . In the above example, the variance between the classes is 3 times lower than the total variance which is a serious underestimation.

**Linear Transformations.** Let us consider a linear transformation  $y_i = a + bx_i$  ( $b \neq 0$ ) of the original data  $x_i$ , ( $i = 1, 2, \dots, n$ ). We get the arithmetic mean of the transformed data as  $\bar{y} = a + b\bar{x}$  and for the variance:

$$\begin{aligned}\tilde{s}_y^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{b^2}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= b^2 \tilde{s}_x^2.\end{aligned}\tag{3.29}$$

*Example 3.2.6* Let  $x_i$ ,  $i = 1, 2, \dots, n$ , denote measurements on time. These data could have been recorded and analysed in hours, but we may be interested in a summary in minutes. We can make a linear transformation  $y_i = 60x_i$ . Then,  $\bar{y} = 60\bar{x}$  and  $\tilde{s}_y^2 = 60^2 \tilde{s}_x^2$ . If the mean and variance of the  $x_i$ 's have already been obtained, then the mean and variance of the  $y_i$ 's can be obtained directly using these transformations.

**Standardization.** A variable is called standardized if its mean is zero and its variance is 1. Standardization can be achieved by using the following transformation:

$$y_i = \frac{x_i - \bar{x}}{\tilde{s}_x} = -\frac{\bar{x}}{\tilde{s}_x} + \frac{1}{\tilde{s}_x} x_i = a + bx_i.\tag{3.30}$$

It follows that  $\bar{y} = \sum_{i=1}^n (x_i - \bar{x})/\tilde{s}_x = 0$  and  $\tilde{s}_y^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / \tilde{s}_x^2 = 1$ . There are many statistical methods which require standardization, see, for example, Sect. 10.3.1 for details in the context of statistical tests.

*Example 3.2.7* Let  $X$  be a variable which measures air pollution by using the concentration of atmospheric particulate matter (in  $\mu\text{g}/\text{m}^3$ ). Suppose we have the following 10 measurements:

30    25    12    45    50    52    38    39    45    33.

We calculate  $\bar{x} = 36.9$ ,  $\tilde{s}_x^2 = 136.09$ , and  $\tilde{s}_x = 11.67$ . To get a standardized variable  $Y$ , we transform all the observations  $x_i$ 's as

$$y_i = \frac{x_i - \bar{x}}{\tilde{s}_x} = -\frac{\bar{x}}{\tilde{s}_x} + \frac{1}{\tilde{s}_x}x_i = -\frac{36.9}{11.67} + \frac{1}{11.67}x_i = -3.16 + 0.086x_i.$$

Now  $y_1 = -3.16 + 0.086 \cdot 30 = -0.58$ ,  $y_2 = -3.16 + 0.086 \cdot 25 = -1.01$ , ..., are the standardized observations. The `scale` command in *R* allows standardization, and we can obtain the standardized observations corresponding to the 10 measurements as

```
air <- c(30,25,12,45,50,52,38,39,45,33)
scale(air)
```



Please note that the `scale` command uses  $1/(n - 1)$  for calculating the variance, as already outlined above. Thus, the results provided by `scale` are not identical to those using (3.30).

### 3.2.3 Coefficient of Variation

Consider a situation where two different variables have arithmetic means  $\bar{x}_1$  and  $\bar{x}_2$  with standard deviations  $\tilde{s}_1$  and  $\tilde{s}_2$ , respectively. Suppose we want to compare the variability of hotel prices in Munich (measured in euros) and London (measured in British pounds). How can we provide a fair comparison? Since the prices are measured in different units, and therefore likely have arithmetic means which differ substantially, it does not make much sense to compare the standard deviations directly. The coefficient of variation  $v$  is a measure of dispersion which uses both the standard deviation and mean and thus allows a fair comparison. It is properly defined only when all the values of a variable are measured on a ratio scale and are positive such that  $\bar{x} > 0$  holds. It is defined as

$$v = \frac{s}{\bar{x}}. \tag{3.31}$$

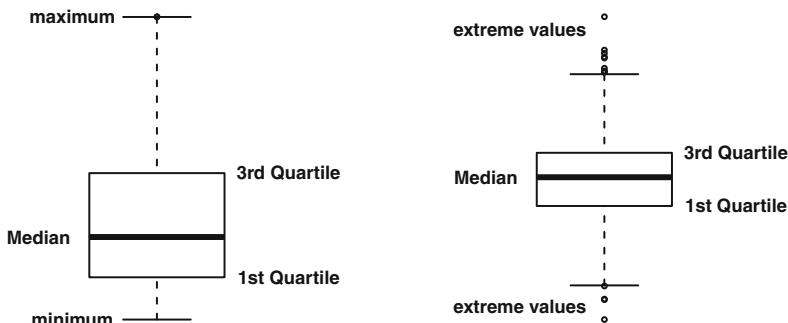
The coefficient of variation is a unit-free measure of dispersion. It is often used when the measurements of two variables are different but can be put into relation by using a linear transformation  $y_i = bx_i$ . It is possible to show that if all values  $x_i$  of a variable  $X$  are transformed into a variable  $Y$  with values  $y_i = b \cdot x_i$ ,  $b > 0$ , then  $v$  does not change.

*Example 3.2.8* If we want to compare the variability of hotel prices in two selected cities in Germany and England, we could calculate the mean prices, together with their standard deviation. Suppose a sample of prices of say 100 hotels in two selected cities in Germany and England is available and suppose we obtain the mean and standard deviations of the two cities as  $x_1 = \text{€}130$ ,  $x_2 = \text{£}230$ ,  $s_1 = \text{€}99$ , and  $s_2 = \text{£}212$ . Then,  $v_1 = 99/130 \approx 0.72$  and  $v_2 = 212/230 = 0.92$ . This indicates higher variability in hotel prices in England. However, if the data distribution is skewed or bimodal, then it may be wise not to choose the arithmetic mean as a measure of central tendency and likewise the coefficient of variation.

### 3.3 Box Plots

So far we have described various measures of central tendency and dispersion. It can be tedious to list those measures in summary tables. A simple and powerful graph is the **box plot** which summarizes the distribution of a continuous (or sometimes an ordinal) variable by using its median, quartiles, minimum, maximum, and extreme values.

Figure 3.5a shows a typical box plot. The vertical length of the box is the interquartile range  $d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$ , which shows the region that contains 50% of the data. The bottom end of the box refers to the first quartile, and the top end of the box refers to the third quartile. The thick line in the box is the median. It becomes immediately clear that the box indicates the symmetry of the data: if the median is in the middle of the box, the data should be symmetric, otherwise it is skewed. The *whiskers* at the end of the plot mark the minimum and maximum values of the data. Looking at the box plot as a whole tells us about the data distribution and the range and variability of observations. Sometimes, it may be advisable to understand which values are extreme in the sense that they are “far away” from the centre of the distribution. In many software packages, including *R*, values are defined to be extreme if they are greater than 1.5 box lengths away from the first or third quartile. Sometimes, they are called outliers. Outliers and extreme values are defined differently in some software packages and books.



(a) Box plot without extreme values

(b) Box plot with extreme values

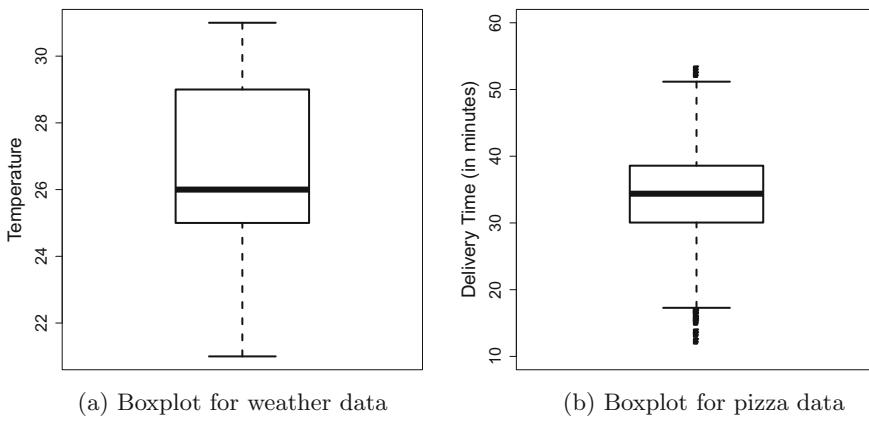
The `boxplot` command in *R* draws a box plot. The `range` option controls whether extreme values should be plotted, and if yes, how one wants to define such values.

```
boxplot(variable, range=1.5)
```



*Example 3.3.1* Recall Examples 3.0.1–3.2.5 where we looked at the temperature in Bangkok during December. We have already calculated the median ( $26^{\circ}\text{C}$ ) and the quartiles ( $25, 29^{\circ}\text{C}$ ). The minimum and maximum values are  $21^{\circ}\text{C}$  and  $31^{\circ}\text{C}$ . The box plot for this data is shown in Fig. 3.5a. One can see that the temperature distribution is slightly skewed with more variability for lower temperatures. The interquartile range is 4, and therefore, any value  $>29 + 4 \times 1.5 = 35$  or  $<25 - 4 \times 1.5 = 19$  would be an extreme value. However, there are no extreme values in the data.

*Example 3.3.2* Consider again the pizza data described in Appendix A.4. We use *R* to plot the box plot for the delivery time via `boxplot(time)` (Fig. 3.5b). We see a symmetric distribution with a median delivery time of about 35 min. Most of the deliveries took between 30 and 40 min. The extreme values indicate that there were some exceptionally short and long delivery times.



## 3.4 Measures of Concentration

A completely different concept used to describe a quantitative variable is the idea of concentration. For a variable  $X$ , it summarizes the proportion of each observation with respect to the sum of all observations  $\sum_{i=1}^n x_i$ . Let us look at a simple example to demonstrate its usefulness.

**Table 3.1** Concentration of farmland: two different situations

Farmer ( $i$ )	$x_i$ (Area, in hectare)
1	20
2	20
3	20
4	20
5	20
	$\sum_{i=1}^5 x_i = 100$
Farmer ( $i$ )	$x_i$ (Area, in hectare)
1	0
2	0
3	0
4	0
5	100
	$\sum_{i=1}^5 x_i = 100$

*Example 3.4.1* Consider a village with 5 farms. Each farmer has a farm of a certain size. How can we evaluate the land distribution? Do all farmers have a similar amount of land or do one or two farmers have a big advantage because they have considerably more space?

Table 3.1 shows two different situations: in the table on the left, we see an equal distribution of land, i.e. each farmer owns 20 hectares of farmland. This means  $X$  is *not* concentrated, rather it is equally distributed. A statistical function describing the concentration could return a value of zero in such a case. Consider another extreme where one farmer owns all the farmland and the others do not own anything, as shown on the right side of Table 3.1. This is an extreme concentration of land: one person owns everything and thus, we say the concentration is high. A statistical function describing the concentration could return a value of one in such a case.

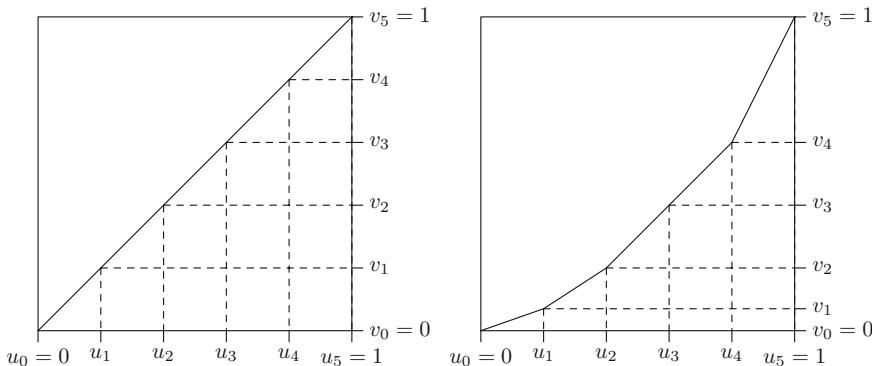
### 3.4.1 Lorenz Curve

The **Lorenz curve** is a popular method to display concentrations graphically. Consider  $n$  observations  $x_1, x_2, \dots, x_n$  of a variable  $X$ . Assume that all the observations are positive. The sum of all the observations is  $\sum_{i=1}^n x_i = n\bar{x}$  if the data is ungrouped. First, we need to order the data:  $0 \leq x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . To plot the Lorenz curve, we need

$$u_i = \frac{i}{n}, \quad i = 0, \dots, n, \tag{3.32}$$

and

$$v_i = \frac{\sum_{j=1}^i x_{(j)}}{\sum_{j=1}^n x_{(j)}}, \quad i = 1, \dots, n; \quad v_0 := 0, \tag{3.33}$$



**Fig. 3.5** Lorenz curves for no concentration (left) and some concentration (right)\*

where  $\sum_{j=1}^i x_{(j)}$  is the cumulative total of observations up to the  $i$ th observation. The idea is that  $v_i$  describe the contribution of all values  $\leq i$  in comparison with the sum of all values. Plotting  $u_i$  against  $v_i$  for all  $i$  shows how much the sum of all  $x_i$ , for all observations  $\leq i$ , contributes to the total sum. In other words, the point  $(u_i, v_i)$  says that  $u_i \cdot 100\%$  of observations contain  $v_i \cdot 100\%$  of the sum of all  $x_i$  less than or equal to  $i$ . Obviously, if all  $x_i$  are identical, the Lorenz curve will be a straight diagonal line, also known as the identity line or **line of equality**. If the  $x_i$  are of different sizes, then the Lorenz curve falls below the line of equality. This is illustrated in the following example.

*Example 3.4.2* Recall Example 3.4.1 where we looked at the distribution of farmland among 5 farmers. On the upper panel of Table 3.1, we observed an equal distribution of land among the farmers:  $x_1 = 20, x_2 = 20, x_3 = 20, x_4 = 20$ , and  $x_5 = 20$ . We obtain  $u_1 = 1/5, u_2 = 2/5, \dots, u_5 = 1$  and  $v_1 = 20/100, v_2 = 40/100, \dots, v_5 = 1$ . This yields a Lorenz curve as displayed on the left side of Fig. 3.5: there is no concentration. We can interpret each point. For example,  $(u_2, v_2) = (0.4, 0.4)$  means that 40 % of farmers own 40 % of the land.

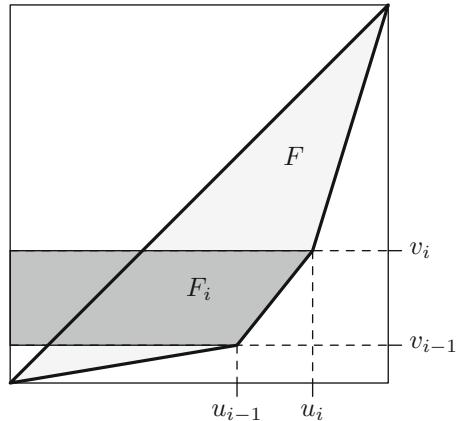
The lower panel of Table 3.1 describes the situation with strong concentration. For this table, we obtain  $u_1 = 1/5, u_2 = 2/5, \dots, u_5 = 1$  and  $v_1 = 0, v_2 = 0, \dots, v_5 = 1$ . Therefore, for example, 80 % of farmers own 0 % of the land which shows strong inequality. Most often we do not have such extreme situations. In this case, the Lorenz curve is bent towards the lower right corner of the plot, see the right side of Fig. 3.5.

We can plot the Lorenz curve in R using the Lc command in the library ineq. The Lorenz curve for the left table of Example 3.4.1 is plotted in R as follows:

```
library(ineq)
x <- c(20,20,20,20,20)
plot(Lc(x))
```



**Fig. 3.6** Lorenz curve and the Gini coefficient\*



We can use the same approach as above to obtain the Lorenz curve when we have grouped data. We simply describe the contributions for each class rather than for each observation and approximate the values in each class by using its mid-point. More formally we can write:

$$\tilde{u}_i = \sum_{j=1}^i f_j, \quad i = 1, 2, \dots, k; \quad \tilde{u}_0 := 0 \quad (3.34)$$

and

$$\tilde{v}_i = \frac{\sum_{j=1}^i f_j a_j}{\sum_{j=1}^k f_j a_j} = \frac{\sum_{j=1}^i n_j a_j}{n \bar{x}}, \quad i = 1, 2, \dots, k; \quad \tilde{v}_0 := 0. \quad (3.35)$$

### 3.4.2 Gini Coefficient

We have seen in Sect. 3.4.1 that the Lorenz curve corresponds to the identity line, that is the diagonal line of equality, for no concentration. When there is some concentration, then the curve deviates from this line. The amount of deviation depends on the strength of concentration. Suppose we want to design a measure of concentration which is 0 for no concentration and 1 for perfect (i.e. extreme) concentration. We can simply measure the area between the Lorenz curve and the identity line and multiply it by 2. For no concentration, the area will be zero and hence the measure will be zero. If there is perfect concentration, then the curve will coincide with the axes, the area will be close to 0.5, and twice the area will be close to one. The measure based on such an approach is called the Gini coefficient:

$$G = 2 \cdot F. \quad (3.36)$$

Note that  $F$  is the area between the curve and the bisection or diagonal line.

The Gini coefficient can be estimated by adding up the areas of the trapeziums  $F_i$  as displayed in Fig. 3.6:

$$F = \sum_{i=1}^n F_i - 0.5,$$

where

$$F_i = \frac{u_{i-1} + u_i}{2}(v_i - v_{i-1}).$$

It can be shown that this corresponds to

$$G = 1 - \frac{1}{n} \sum_{i=1}^n (v_{i-1} + v_i), \quad (3.37)$$

but the proof is omitted. The same formula can be used for grouped data except that  $\tilde{v}$  is used instead of  $v$ . Since

$$0 \leq G \leq \frac{n-1}{n}, \quad (3.38)$$

one may prefer to use the standardized Gini coefficient

$$G^+ = \frac{n}{n-1} G, \quad (3.39)$$

which takes a maximum value of 1.

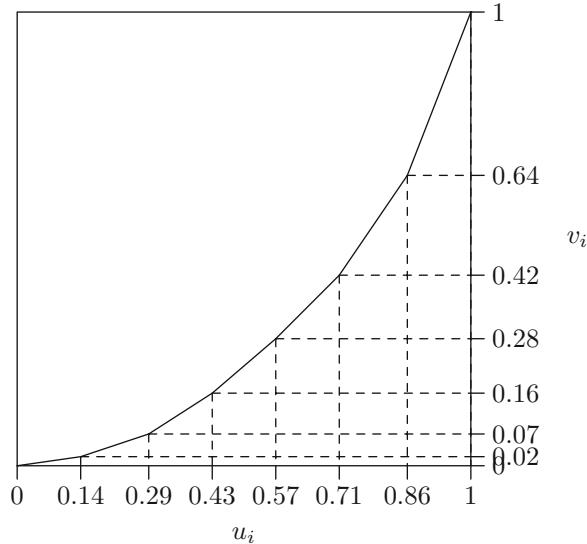
*Example 3.4.3* We return to our farmland example. Suppose we have 7 farmers with farms of different sizes:

Farmer		1	2	3	4	5	6	7
Farmland size $x_i$		20	14	59	9	36	23	3

Using the ordered values, we can calculate  $u_i$  and  $v_i$  using (3.32) and (3.33):

$i$	$x_{(i)}$	$u_i$	$v_i$
1	3	$\frac{1}{7} = 0.1429$	$\frac{3}{164} = 0.0183$
2	9	$\frac{2}{7} = 0.2857$	$\frac{12}{164} = 0.0732$
3	14	$\frac{3}{7} = 0.4286$	$\frac{26}{164} = 0.1585$
4	20	$\frac{4}{7} = 0.5714$	$\frac{46}{164} = 0.2805$
5	23	$\frac{5}{7} = 0.7143$	$\frac{69}{164} = 0.4207$
6	36	$\frac{6}{7} = 0.8571$	$\frac{105}{164} = 0.6402$
7	59	$\frac{7}{7} = 1.0000$	$\frac{164}{164} = 1.0000$

**Fig. 3.7** Lorenz curve for Example 3.4.3\*



The Lorenz curve is displayed in Fig. 3.7. Using this information, it is easy to calculate the Gini coefficient:

$$G = 1 - \frac{1}{7}(0.0183 + [0.0183 + 0.0732] + [0.0732 + 0.1585] + [0.1585 + 0.2805] \\ + [0.2805 + 0.4207] + [0.4207 + 0.6402] + [0.6402 + 1]) = 0.402$$

We know that  $G = 0.4024 \leq \frac{6}{7} = \frac{n-1}{n}$ . To standardize the coefficient, we therefore have to use (3.39):

$$G^+ = \frac{7}{6}G = \frac{7}{6} \cdot 0.4024 = 0.4695.$$

In R, we can obtain the non-standardized Gini Coefficient using the `ineq` function in the library `ineq`.

```
library(ineq)
farm <- c(20,14,59,9,36,23,3)
ineq(farm)
```



### 3.5 Key Points and Further Issues

**Note:**

- ✓ A summary on how to descriptively summarize data is given in Appendix D.1.
- ✓ The median is preferred over the arithmetic mean when the data distribution is skewed or there are extreme values.
- ✓ If data of a continuous variable is grouped, and the original ungrouped data is not known, additional assumptions are needed to calculate measures of central tendency and dispersion. However, in some cases, these assumptions may not be satisfied, and the formulae provided may give imprecise results.
- ✓ QQ-plots are not only descriptive summaries but can also be used to test modelling assumptions, see Chap. 11.9 for more details.
- ✓ The distribution of a continuous variable can be easily summarized using a box plot.

---

### 3.6 Exercises

*Exercise 3.1* A hiking enthusiast has a new app for his smartphone which summarizes his hikes by using a GPS device. Let us look at the distance hiked (in km) and maximum altitude (in m) for the last 10 hikes:

Distance	12.5	29.9	14.8	18.7	7.6	16.2	16.5	27.4	12.1	17.5
Altitude	342	1245	502	555	398	670	796	912	238	466

- (a) Calculate the arithmetic mean and median for both distance and altitude.
- (b) Determine the first and third quartiles for both the distance and the altitude variables. Discuss the shape of the distribution given the results of (a) and (b).
- (c) Calculate the interquartile range, absolute median deviation, and standard deviation for both variables. What is your conclusion about the variability of the data?
- (d) One metre corresponds to approximately 3.28 ft. What is the average altitude when measured in feet rather than in metres?
- (e) Draw and interpret the box plot for both distance and altitude.
- (f) Assume distance is measured as only short (5–15 km), moderate (15–20 km), and long (20–30 km). Summarize the grouped data in a frequency table. Calculate the weighted arithmetic mean under the assumption that the raw data is not

known. Determine the weighted median under the assumption that the values within each class are equally distributed.

- (g) What is the variance for the grouped data when the raw data is known, i.e. when one has knowledge about the variance in each class? How does it compare with the variance one obtains when the raw data is unknown?
- (h) Use *R* to reproduce the results of (a), (b), (c), (e), and (f).

*Exercise 3.2* A gambler notes down his wins and losses (in €) from playing 10 games of roulette in a casino.

Round	1	2	3	4	5	6	7	8	9	10
Won/Lost	200	600	-200	-200	-200	-100	-100	-400	0	

- (a) Assume  $\bar{x} = -€90$  and  $s = €294.7881$ . What is the result of round 10?
- (b) Determine the mode and the interquartile range.
- (c) A different gambler plays 33 rounds of roulette. His results are  $\bar{x} = €12$  and  $s = €1000$ . Is it meaningful to compare the variability of results of the two players by using the coefficient of variation? If yes, determine the coefficients of variation; if no, why is a comparison not possible?

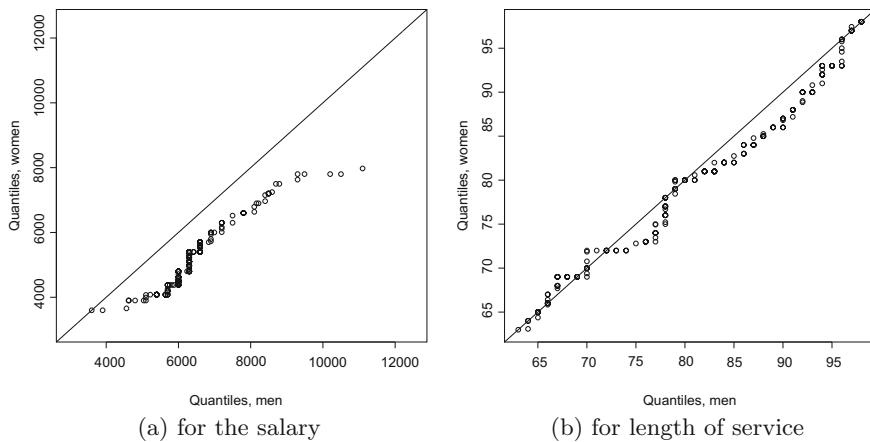
*Exercise 3.3* A fashion boutique has summarized its daily sales of designer socks in different groups: men's socks, women's socks, and children's socks. Unfortunately, the data for men's socks was lost. Determine the missing values.

	$n$	Arithmetic mean in €	Standard deviation in €
Women's wear	45	16	$\sqrt{6}$
Men's wear	?	?	?
Children's wear	20	7.5	$\sqrt{3}$
Total	100	15	$\sqrt{19.55}$

*Exercise 3.4* The number of members of a millionaires' club were as follows:

Year	2011	2012	2013	2014	2015	2016
Members	23	24	27	25	30	28

- (a) What is the average growth rate of the membership?
- (b) Based on the results of (a), how many members would one expect in 2018?



**Fig. 3.8** QQ-plots

- (c) The president of the club is interested in the number of members in 2025, the year when his presidency ends. Would it make sense to predict the number of members for 2025?

In 2015, the members invested €250 million on the stock market. 10 members contributed 16% of the investment sum, 8 members contributed €60 million, 8 members contributed €70 million, and another 4 members contributed the remaining amount.

- (d) Draw the Lorenz curve for this data.
  - (e) Calculate and interpret the standardized Gini coefficient.

**Exercise 3.5** Consider the monthly salaries  $Y$  (in Swiss francs) of a well-reputed software company, as well as the length of service (in months,  $X$ ), and gender ( $Z$ ). Figure 3.8 shows the QQ-plots for both  $Y$  and  $X$  given  $Z$ . Interpret both graphs.

**Exercise 3.6** There is no built-in function in R to calculate the mode of a variable. Program such a function yourself. Hint: type ?table and ?names to recall the functionality of these functions. Combine them in an intelligent way.

**Exercise 3.7** Consider a country in which 90% of the wealth is owned by 20% of the population, the so-called upper class. For simplicity, let us assume that the wealth is distributed equally within this class.

- (a) Draw the Lorenz curve for this country.
  - (b) Now assume a revolution takes place in the country and all members of the upper class have to give away their wealth which is then distributed equally across the remaining population. Draw the Lorenz curve for this scenario.
  - (c) What would the curve from (b) look like if the entire upper class left the country?

*Exercise 3.8* A bus route in the mountainous regions of Romania has a length of 418 km. The manager of the bus company serving the route wants his buses to finish a trip within 8 h. The bus travels the first 180 km with an average speed of 48 km/h, the next 117 km with an average speed of 37 km/h, and the last section with an average speed of 52 km/h.

- (a) What is the average speed with which the bus travels?
- (b) Will the bus finish the trip in time?

*Exercise 3.9* Four friends have a start-up company which sells vegan ice cream. Their initial financial contributions are as follows:

Person	1	2	3	4
Contribution (in €)	800	10300	4700	2220

- (a) Calculate and draw the Lorenz curve.
- (b) Determine and interpret the standardized Gini coefficient.
- (c) Does  $G^+$  change if each of the friends contributes only half the amount of money? If yes, how much? If no, why not?
- (d) Use R to draw the above Lorenz curve and to calculate the Gini coefficient.

*Exercise 3.10* Recall the pizza delivery data which is described in Appendix A.4. Use R to read in and analyse the data.

- (a) Calculate the mean, median, minimum, maximum, first quartile, and third quartile for all quantitative variables.
- (b) Determine and interpret the 99 % quantile for delivery time and temperature.
- (c) Write a function which calculates the absolute mean deviation. Use the function to calculate the absolute mean deviation of temperature.
- (d) Scale the delivery time and calculate the mean and variance for this variable.
- (e) Draw a box plot for delivery time and temperature. The box plots should not highlight extreme values.
- (f) Use the cut command to create a new variable which summarizes delivery time in steps of 10 min. Calculate the arithmetic mean of this variable.
- (g) Reproduce the QQ-plots shown in Example 3.1.6.

→ Solutions to all exercises in this chapter can be found on p. [333](#)

\*Source Toutenburg, H., Heumann, C., *Deskriptive Statistik*, 7th edition, 2009, Springer, Heidelberg

In Chaps. 2 and 3 we discussed how to analyse a single variable using graphs and summary statistics. However, in many situations we may be interested in the interdependence of two or more variables. For example, suppose we want to know whether male and female students in a college have any preference between the subjects mathematics and biology, i.e. if there is any evidence that male students prefer mathematics over biology and female students prefer biology over mathematics or vice versa. Suppose we choose an equal number of male and female students and ask them about their preferred subject. We expect that if there is no association between the two variables “gender of student” (male or female) and “subject” (mathematics or biology), then an equal proportion of male and female students should choose the subjects biology and mathematics, respectively. Any difference in the proportions may indicate a preference of males or females for a particular subject. Similarly, in another example, we may want to find out whether female employees of an organization are paid less than male employees or vice versa. Let us assume again that we choose an equal number of male and female employees and assume further that the salary is measured as a binary variable (low- versus high-salary group). We then expect that if there is no gender discrimination, the number of male and female employees in the lower- and higher-salary groups in the organization should be approximately equal. In both examples, the variables considered are binary and nominal (although the salary can also be seen as ordinal) and the data is summarized in terms of frequency measures. There may, however, be situations in which we are interested in associations between ordinal or continuous variables. Consider a data set in which height, weight, and age of infants are given. Usually, the height and weight of infants increase with age. Also, the height of infants increases with their weight and vice versa. Clearly, there is an interrelation or association among the three variables. In another example, two persons have to judge participants of a dance competition and rank them according to their performance. Now if we want to learn about the fairness in the judgment, we expect that both the judges give similar ranks to each candidate,

i.e. both judges give high ranks to good candidates and low ranks to not so good candidates. We are therefore interested in studying the association between the ranks given by the two judges. In all these examples, the intention lies in measuring the degree of association between two (or more) variables. We therefore need to study different ways of measuring the association level for different types of variables. In this chapter, we present measures and graphical summaries for the association of two variables—dependent on their scale.

## 4.1 Summarizing the Distribution of Two Discrete Variables

When both variables are discrete, then it is possible to list all combinations of values of the two variables and to count how often these combinations occur in the data. Consider the salary example in the introduction to this chapter in which both the variables were binary. There are four possible combinations of variable categories (female and low-salary group, female and high-salary group, male and low-salary group, and male and high-salary group). A complete description of the joint occurrence of these two variables can be given by counting, for each combination, the number of units for which this combination is measured. In the following, we generalize this concept to two variables where each can have an arbitrary (but fixed) number of values or categories.

### 4.1.1 Contingency Tables for Discrete Data

Suppose we have data on two discrete variables. This data can be described in a two-dimensional **contingency table**.

*Example 4.1.1* An airline conducts a customer satisfaction survey. The survey includes questions about travel class and satisfaction levels with respect to different categories such as seat comfort, in-flight service, meals, safety, and other indicators. Consider the information on  $X$ , denoting the travel class (Economy = “E”, Business = “B”, First = “F”), and “Y”, denoting the overall satisfaction with the flight on a scale from 1 to 4 as 1 (poor), 2 (fair), 3 (good), and 4 (very good). A possible response from 12 customers may look as follows:

		Passenger number											
		1	2	3	4	5	6	7	8	9	10	11	12
Travel class	E	E	E	B	E	B	F	E	E	B	E	B	
	Satisfaction	2	4	1	3	1	2	4	3	2	4	3	3

We can calculate the absolute frequencies for each of the combination of observed values. For example, there are 2 passengers (passenger numbers 3 and 5) who were

**Table 4.1** Contingency table for travel class and satisfaction

		Overall rating of flight quality				Total (row)
		Poor	Fair	Good	Very good	
Travel class	Economy	2	2	2	1	7
	business	0	1	2	1	4
	first	0	0	0	1	1
	Total (column)	2	3	4	3	12

flying in economy class and rated the flight quality as poor, there were no passengers from both business class and first class who rated the flight quality as poor; there were 2 passengers who were flying in economy class and rated the quality as fair (2), and so on. Table 4.1 is a two-dimensional table summarizing this information.

Note that we not only summarize the joint frequency distribution of the two variables but also the distributions of the individual variables. Summing up the rows and columns of the table gives the respective frequency distributions. For example, the last column of the table demonstrates that 7 passengers were flying in economy class, 4 passengers were flying in business class and 1 passenger in first class.

Now we extend this example and discuss a general framework to summarize the absolute frequencies of two discrete variables in contingency tables. We use the following notations: Let  $x_1, x_2, \dots, x_k$  be the  $k$  classes of a variable  $X$  and let  $y_1, y_2, \dots, y_l$  be the  $l$  classes of another variable  $Y$ . We assume that both  $X$  and  $Y$  are discrete variables. It is possible to summarize the absolute frequencies  $n_{ij}$  related to  $(x_i, y_j)$ ,  $i = 1, 2, \dots, k$ ,  $j = 1, 2, \dots, l$ , in a  $k \times l$  **contingency table** as shown in Table 4.2.

**Table 4.2**  $k \times l$  contingency table

		Y						Total (rows)
		$y_1$	...	$y_j$	...	$y_l$		
$X$	$x_1$	$n_{11}$	...	$n_{1j}$	...	$n_{1l}$	$n_{1+}$	
	$x_2$	$n_{21}$	...	$n_{2j}$	...	$n_{2l}$	$n_{2+}$	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	
	$x_i$	$n_{i1}$	...	$n_{ij}$	...	$n_{il}$	$n_{i+}$	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	
	$x_k$	$n_{k1}$	...	$n_{kj}$	...	$n_{kl}$	$n_{k+}$	
	Total (columns)	$n_{+1}$	...	$n_{+j}$	...	$n_{+l}$	$n$	

We denote the sum of the  $i$ th row as  $n_{i+} = \sum_{j=1}^l n_{ij}$  and the sum over the  $j$ th column as  $n_{+j} = \sum_{i=1}^k n_{ij}$ . The total number of observations is therefore

$$n = \sum_{i=1}^k n_{i+} = \sum_{j=1}^l n_{+j} = \sum_{i=1}^k \sum_{j=1}^l n_{ij}. \quad (4.1)$$

*Remark 4.1.1* Note that it is also possible to use the relative frequencies  $f_{ij} = n_{ij}/n$  instead of the absolute frequencies  $n_{ij}$  in Table 4.2, see Example 4.1.2.

### 4.1.2 Joint, Marginal, and Conditional Frequency Distributions

When the data on two variables are summarized in a contingency table, there are several concepts which can help us in studying the characteristics of the data. For example, how the values of both the variables behave jointly, how the values of one variable behave when another variable is kept fixed etc. These features can be studied using the concepts of joint frequency distribution, marginal frequency distribution, and conditional frequency distribution. If relative frequency is used instead of absolute frequency, then we speak of the joint relative frequency distribution, marginal relative frequency distribution, and conditional relative frequency distribution.

**Definition 4.1.1** Using the notations of Table 4.2, we define the following:

The frequencies  $n_{ij}$  represent the **joint frequency distribution** of  $X$  and  $Y$ .

The frequencies  $n_{i+}$  represent the **marginal frequency distribution** of  $X$ .

The frequencies  $n_{+j}$  represent the **marginal frequency distribution** of  $Y$ .

We define  $f_{i|j}^{X|Y} = n_{ij}/n_{+j}$  to be the **conditional frequency distribution** of  $X$  given  $Y = y_j$ .

We define  $f_{j|i}^{Y|X} = n_{ij}/n_{i+}$  to be the **conditional frequency distribution** of  $Y$  given  $X = x_i$

The frequencies  $f_{ij}$  represent the **joint relative frequency distribution** of  $X$  and  $Y$ .

The frequencies  $f_{i+} = \sum_{j=1}^l f_{ij}$  represent the **marginal relative frequency distribution** of  $X$ .

The frequencies  $f_{+j} = \sum_{i=1}^k f_{ij}$  represent the **marginal relative frequency distribution** of  $Y$ .

We define  $f_{i|j}^{X|Y} = f_{ij}/f_{+j}$  to be the **conditional relative frequency distribution** of  $X$  given  $Y = y_j$ .

We define  $f_{j|i}^{Y|X} = f_{ij}/f_{i+}$  to be the **conditional relative frequency distribution** of  $Y$  given  $X = x_i$ .

**Table 4.3** Contingency table for travel class and satisfaction

		Overall rating of flight quality				Total (rows)
		Poor	Fair	Good	Very good	
Travel class	Economy	10	33	15	4	62
	Business	0	3	20	2	25
	First	0	0	5	8	13
	Total (columns)	10	36	40	14	100

Note that for a bivariate joint frequency distribution, there will only be two marginal (or relative) frequency distributions but possibly more than two conditional (or relative) frequency distributions.

*Example 4.1.2* Recall the setup of Example 4.1.1. We now collect and evaluate the responses of 100 customers (instead of 12 passengers as in Example 4.1.1) regarding their choice of the travel class and their overall satisfaction with the flight quality.

The data is provided in Table 4.3 where each of the cell entries illustrates how many out of 100 passengers answered  $x_i$  and  $y_j$ : for example, the first entry “10” indicates that 10 passengers were flying in economy class and described the overall service quality as poor.

- The marginal frequency distributions are displayed in the last column and last row, respectively. For example, the marginal distribution of  $X$  refers to the frequency table of “travel class” ( $X$ ) and tells us that 62 passengers were flying in economy class, 25 in business class, and 13 in first class. Similarly, the marginal distribution of “overall rating of flight quality” ( $Y$ ) tells us that 10 passengers rated the quality as poor, 36 as fair, 40 as good, and 14 as very good.
- The conditional frequency distributions give us an idea about the behaviour of one variable when the other one is kept fixed. For example, the conditional distribution of the “overall rating of flight quality” ( $Y$ ) among passengers who were flying in economy class ( $f_{Y|X=\text{Economy}}$ ) gives  $f_{1|1}^{Y|X} = 10/62 \approx 16\%$  which means that approximately 16 % of the customers in economy class are rating the quality as poor,  $f_{2|1}^{Y|X} = 33/62 \approx 53\%$  of the customers in economy class are rating the quality as fair,  $f_{3|1}^{Y|X} = 15/62 \approx 24\%$  of the customers in economy class are rating the quality as good and  $f_{4|1}^{Y|X} = 4/62 \approx 7\%$  of the customers in economy class are rating the quality as very good. Similarly,  $f_{3|2}^{Y|X} = 20/25 \approx 80\%$  which means that 80 % of the customers in business class are rating the quality as good and so on.
- The conditional frequency distribution of the “travel class” ( $X$ ) of passengers given the “overall rating of flight quality” ( $Y$ ) is obtained by  $f_{X|Y=\text{Satisfaction level}}$ . For example,  $f_{X|Y=\text{good}}$  gives  $f_{1|3}^{X|Y} = 15/40 = 37.5\%$  which means that 37.5 %

of the passengers who rated the flight to be good travelled in economy class,  $f_{2|3}^{X|Y} = 20/40 = 50\%$  of the passengers who rated the flight to be good travelled in business class and  $f_{3|3}^{X|Y} = 5/40 = 12.5\%$  of the passengers who rated the flight to be good travelled in first class.

- In total, we have 100 customers and hence

$$\begin{aligned} n &= \sum_{i=1}^k n_{i+} = 62 + 25 + 13 = \sum_{j=1}^l n_{+j} = 10 + 36 + 40 + 14 \\ &= \sum_{i=1}^k \sum_{j=1}^l n_{ij} = 10 + 33 + 15 + 4 + 3 + 20 + 2 + 5 + 8 = 100 \end{aligned}$$

- Alternatively, we can summarize  $X$  and  $Y$  using the relative frequencies as follows:

		Overall rating of flight quality				Total (rows)
		Poor	Fair	Good	Very good	
Travel class	Economy	$\frac{10}{100}$	$\frac{33}{100}$	$\frac{15}{100}$	$\frac{4}{100}$	$\frac{62}{100}$
	Business	0	$\frac{3}{100}$	$\frac{20}{100}$	$\frac{2}{100}$	$\frac{25}{100}$
	First	0	0	$\frac{5}{100}$	$\frac{8}{100}$	$\frac{13}{100}$
Total (columns)		$\frac{10}{100}$	$\frac{36}{100}$	$\frac{40}{100}$	$\frac{14}{100}$	1

To produce the frequency table without the marginal distributions, we can use the *R* command `table(X, Y)`. To obtain the full contingency table including the marginal distributions in *R*, one can use the function `addmargins()`. For the relative frequencies, the function `prop.table()` can be used. In summary, a full contingency table is obtained by using

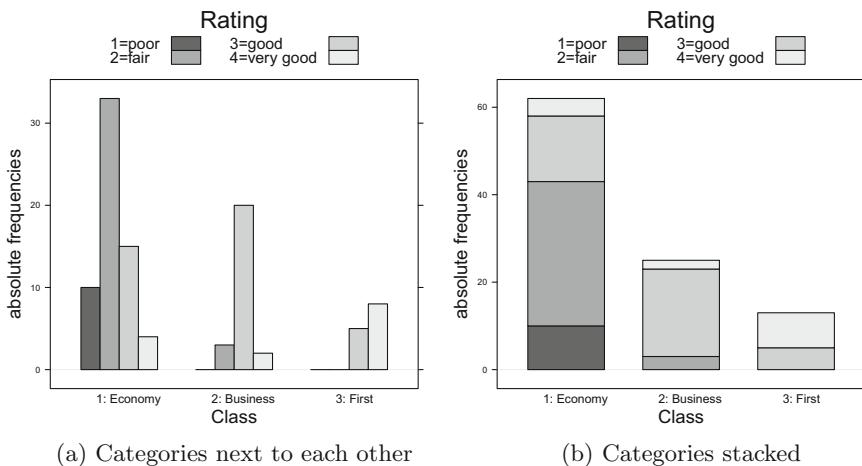
```
addmargins(table(X, Y))
addmargins(prop.table(table(X, Y)))
```



### 4.1.3 Graphical Representation of Two Nominal or Ordinal Variables

Bar charts (see Sect. 2.3.1) can be used to graphically summarize the association between two nominal or two ordinal variables. The bar chart is drawn for  $X$  and the categories of  $Y$  are represented by separated bars or stacked bars for each category of  $X$ . In this way, we summarize the joint distribution of the contingency table.

*Example 4.1.3* Consider Example 4.1.2. There are 62 passengers flying in the economy class. From these 62 passengers, 10 rated the quality of the flight as poor, 33 as fair, 15 as good, and 4 as very good. This means for  $X = x_1$  (= Economy), we can

**Fig. 4.1** Bar charts for travel class and rating of satisfaction

either place 4 bars next to each other, as in Fig. 4.1a, or we can stack them on top of each other, as in Fig. 4.1b. The same can be done for the other categories of  $X$ , see Fig. 4.1. Stacked and stratified bar charts are prepared in R by calling the library `lattice` and using the function `bar chart`. In detail, one needs to specify:

```
Class <- c(rep('1: Economy',62),rep('2: Business',25),
rep('3: First',13))
Rating <- c(rep('1=poor',10),rep('2=fair',33),...)
library(lattice)
barchart(table(Class,Rating),horizontal=FALSE,stack=FALSE)
barchart(table(Class,Rating),horizontal=FALSE,stack=TRUE)
```



*Remark 4.1.2* There are several other options in R to specify stratified bar charts. We refer the interested reader to Exercise 2.6 to explore how the R package `ggplot2` can be used to make such graphics. Sometimes it can also be useful to visualize the difference of two variables and not stack or stratify the bars, see Exercise 2.1.

**Independence and Expected Frequencies** An important statistical concept is **independence**. In this section, we touch upon its descriptive aspects, see Chaps. 6 (Sect. 6.5) and 7 (Sect. 7.5) for more theoretical details. Two variables are considered to be independent if the observations on one variable do not influence the observations on the other variable. For example, suppose two different persons roll a die separately; then, the outcomes of their rolls do not depend on each other. So we can say that the two observations are independent. In the context of contingency tables, two variables are independent of each other when the joint relative frequency equals the product of the marginal relative frequencies of the two variables, i.e. the

**Table 4.4** Observed and expected absolute frequencies for the airline survey

		Overall rating of flight quality				
		Poor	Fair	Good	Very good	Total
Travel class	Economy	10 (6.2)	33 (22.32)	15 (24.8)	4 (8.68)	62
	Business	0 (2.5)	3 (9.0)	20 (10.0)	2 (3.5)	25
	First	0 (1.3)	0 (4.68)	5 (5.2)	8 (1.82)	13
	Total	10	36	40	14	100

following equation holds:

$$f_{ij} = f_{i+} f_{+j} . \quad (4.2)$$

The **expected absolute frequencies under independence** are obtained by

$$\tilde{n}_{ij} = n f_{ij} = n \frac{n_{i+} n_{+j}}{n} = \frac{n_{i+} n_{+j}}{n} . \quad (4.3)$$

Note that the absolute frequencies are always integers but the expected absolute frequencies may not always be integers.

*Example 4.1.4* Recall Example 4.1.2. The expected absolute frequencies for the contingency table can be calculated using (4.3). For example,

$$\tilde{n}_{11} = \frac{62 \cdot 10}{100} = 6.2, \quad \tilde{n}_{12} = \frac{62 \cdot 36}{100} = 22.32 \quad \text{etc.}$$

Table 4.4 lists both the observed absolute frequency and expected absolute frequency (in brackets).

To calculate the expected absolute frequencies in R, we can access the “expected” object returned from a  $\chi^2$ -test applied to the respective contingency table as follows:

```
chisq.test(table(Class,Rating))$expected
```



A detailed motivation and explanation of this command is given in Sect. 10.8.

## 4.2 Measures of Association for Two Discrete Variables

When two variables are not independent, then they are associated. Their association can be weak or strong. Now we describe some popular measures of association. Measures of association describe the degree of association between two variables and can have a direction as well. Note that if variables are defined on a nominal scale, then nothing can be said about the direction of association, only about the strength.

Let us first consider a  $2 \times 2$  contingency table which is a special case of a  $k \times l$  contingency table, see Table 4.5.

**Table 4.5**  $2 \times 2$  contingency table

		$Y$		Total (row)
		$y_1$	$y_2$	
$X$	$x_1$	$a$	$b$	$a + b$
	$x_2$	$c$	$d$	$c + d$
	Total (column)	$a + c$	$b + d$	$n$

**Table 4.6**  $2 \times 2$  contingency table

		Persons		
		Not affected	Affected	Total (row)
Vaccination	Vaccinated	90	10	100
	Not vaccinated	40	60	100
	Total (column)	130	70	200

The variables  $X$  and  $Y$  are independent if

$$\frac{a}{a+c} = \frac{b}{b+d} = \frac{a+b}{n} \quad (4.4)$$

or equivalently if

$$a = \frac{(a+b)(a+c)}{n}. \quad (4.5)$$

Note that some other forms of the conditions (4.4)–(4.5) can also be derived in terms of  $a, b, c$ , and  $d$ .

*Example 4.2.1* Suppose a vaccination against flu (influenza) is given to 200 persons. Some of the persons may get affected by flu despite the vaccination. The data is summarized in Table 4.6. Using the notations of Table 4.5, we have  $a = 90, b = 10, c = 40, d = 60$ , and thus,  $(a+b)(a+c)/n = 100 \cdot 130/200 = 65$  which is less than  $a = 90$ . Hence, being affected by flu is not independent of the vaccination, i.e. whether one is vaccinated or not has an influence on getting affected by flu. In the vaccinated group, only 10 of 100 persons are affected by flu while in the group not vaccinated 60 of 100 persons are affected. Another interpretation is that if independence holds, then we would expect 65 persons to be not affected by flu in the vaccinated group but we observe 90 persons. This shows that vaccination has a protective effect.

To gain a better understanding about the strength of association between two variables, we need to develop the concept of dependence and independence further. The following three subsections illustrate this in more detail.

### 4.2.1 Pearson's $\chi^2$ Statistic

We now introduce Pearson's  $\chi^2$  statistic which is used for measuring the association between variables in a contingency table and plays an important role in the construction of statistical tests, see Sect. 10.8. The  $\chi^2$  statistic or  $\chi^2$  coefficient for a  $k \times l$  contingency table is given as

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}. \quad (4.6)$$

A simpler formula for  $2 \times 2$  contingency tables is

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}. \quad (4.7)$$

The idea behind the  $\chi^2$  coefficient is that when the relationship between two variables is stronger, then the deviations between observed and expected frequencies are expected to be higher (because the expected frequencies are calculated assuming independence) and this indicates a stronger relationship between the two variables. If observed and expected frequencies are identical or similar, then this is an indication that the association between the two variables is weak and the variables may even be independent. The  $\chi^2$  statistic for a  $k \times l$  contingency table sums up all the differences between the observed and expected frequencies, squares them, and scales them with respect to the expected frequencies. The squaring of the difference makes the statistic independent of the positive and negative signs of the difference between observed and expected frequencies. The range of values for  $\chi^2$  is

$$0 \leq \chi^2 \leq n(\min(k, l) - 1). \quad (4.8)$$

Note that  $\min(k, l)$  is the minimum function and simply returns the smaller of the two numbers  $k$  and  $l$ . For example,  $\min(3, 4)$  returns the value 3. Consequently the values of  $\chi^2$  obtained from (4.6) can be compared with the range from (4.8). A value of  $\chi^2$  close to zero indicates a weak association and a value of  $\chi^2$  close to  $n(\min(k, l) - 1)$  indicates a strong association between the two variables. Note that the range of  $\chi^2$  depends on  $n$ ,  $k$  and  $l$ , i.e. the sample size and the dimension of the contingency table.

The  $\chi^2$  statistic is a *symmetric* measure in the sense that its value does not depend on which variable is defined as  $X$  and which as  $Y$ .

*Example 4.2.2* Consider Examples 4.1.2 and 4.1.4. Using the values from Table 4.4, we can calculate the  $\chi^2$  statistic as

$$\chi^2 = \frac{(10 - 6.2)^2}{6.2} + \frac{(33 - 22.32)^2}{22.32} + \cdots + \frac{(8 - 1.82)^2}{1.82} = 57.95064$$

The maximum possible value for the  $\chi^2$  statistic is  $100(\min(4, 3) - 1) = 200$ . Thus,  $\chi^2 \approx 57$  indicates a moderate association between "travel class" and "overall rating of flight quality" of the passengers. In R, we obtain this result as follows:

```
chisq.test(table(Class,Rating))$statistic
```



### 4.2.2 Cramer's V Statistic

A problem with Pearson's  $\chi^2$  coefficient is that the range of its maximum value depends on the sample size and the size of the contingency table. These values may vary in different situations. To overcome this problem, the coefficient can be standardized to lie between 0 and 1 so that it is independent of the sample size as well as the dimension of the contingency table. Since  $n(\min(k, l) - 1)$  was the maximal value of the  $\chi^2$  statistic, dividing  $\chi^2$  by this maximal value automatically leads to a scaled version with maximal value 1. This idea is used by Cramer's  $V$  statistic which for a  $k \times l$  contingency table is given by

$$V = \sqrt{\frac{\chi^2}{n(\min(k, l) - 1)}}. \quad (4.9)$$

The closer the value of  $V$  gets to 1, the stronger the association between the two variables.

*Example 4.2.3* Consider Example 4.2.2. The obtained  $\chi^2$  statistic is 57.95064. To obtain Cramer's  $V$ , we just need to calculate

$$V = \sqrt{\frac{\chi^2}{n(\min(k, l) - 1)}} = \sqrt{\frac{57.95064}{100(3 - 1)}} \approx 0.54. \quad (4.10)$$

This indicates a moderate association between "travel class" and "overall rating of flight quality" because 0.54 lies in the middle of 0 and 1. In R, there are two options to calculate  $V$ : (i) to calculate the  $\chi^2$  statistic and then adjust it as in (4.9), (ii) to use the functions `assocstats` and `xtabs` contained in the package `vcd` as follows:

```
library(vcd)
assocstats(xtabs(~Class+Rating))
```



### 4.2.3 Contingency Coefficient C

Another option to standardize  $\chi^2$  is given by a corrected version of Pearson's contingency coefficient:

$$C_{\text{corr}} = \frac{C}{C_{\max}} = \sqrt{\frac{\min(k, l)}{\min(k, l) - 1}} \sqrt{\frac{\chi^2}{\chi^2 + n}}, \quad (4.11)$$

with

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \text{and} \quad C_{\max} = \sqrt{\frac{\min(k, l) - 1}{\min(k, l)}}. \quad (4.12)$$

It always lies between 0 and 1. The closer the value of  $C$  is to 1, the stronger the association.

*Example 4.2.4* We know from Example 4.2.2 that the  $\chi^2$  statistic for travel class and satisfaction level is 57.95064. To calculate  $C_{\text{corr}}$ , we need the following calculations:

$$C = \sqrt{\frac{57.95064}{57.95064 + 100}} = 0.606, \quad C_{\max} = \sqrt{\frac{\min(4, 3) - 1}{\min(4, 3)}} = \sqrt{\frac{2}{3}} = 0.816,$$

$$C_{\text{corr}} = \frac{C}{C_{\max}} = \frac{0.606}{0.816} \approx 0.74.$$

There is a moderate to strong association between “travel class” and “overall rating of flight quality” of the passengers. We can compute  $C$  in *R* using the *vcd* package as follows:

```
library(vcd)
Cmax = sqrt((min(c(3,4))-1)/min(c(3,4)))
assocstats(xtabs(~Class+Rating))$cont/Cmax
```



#### 4.2.4 Relative Risks and Odds Ratios

We now introduce the concepts of odds ratios and relative risks. Consider a  $2 \times 2$  contingency table as introduced in Table 4.5. Now suppose we have two variables  $X$  and  $Y$  with their conditional distributions  $f_{i|j}^{X|Y}$  and  $f_{j|i}^{Y|X}$ . In the context of a  $2 \times 2$  contingency table,  $f_{1|1}^{X|Y} = n_{11}/n_{+1}$ ,  $f_{1|2}^{X|Y} = n_{12}/n_{+2}$ ,  $f_{2|2}^{X|Y} = n_{22}/n_{+2}$ , and  $f_{2|1}^{X|Y} = n_{21}/n_{+1}$ . The relative risks are defined as the ratio of two conditional distributions, for example

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{a/(a+c)}{b/(b+d)} \quad \text{and} \quad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{n_{21}/n_{+1}}{n_{22}/n_{+2}} = \frac{c/(a+c)}{d/(b+d)}. \quad (4.13)$$

The odds ratio is defined as the ratio of these relative risks from (4.13) as

$$OR = \frac{f_{1|1}^{X|Y}/f_{1|2}^{X|Y}}{f_{2|1}^{X|Y}/f_{2|2}^{X|Y}} = \frac{f_{1|1}^{X|Y} f_{2|2}^{X|Y}}{f_{2|1}^{X|Y} f_{1|2}^{X|Y}} = \frac{ad}{bc}. \quad (4.14)$$

Alternatively, the odds ratio can be defined as the ratio of the chances for “disease”,  $a/b$  (number of smokers with the disease divided by the number of non-smokers with the disease), and no disease,  $c/d$  (number of smokers with no disease divided by the number of non-smokers with no disease).

The relative risks compare proportions, while the odds ratio compares odds.

*Example 4.2.5* A classical example refers to the possible association of smoking with a particular disease. Consider the following data on 240 individuals:

		Smoking Yes	No	Total (row)
Disease	Yes	34	66	100
	No	22	118	140
Total (column)		56	184	240

We calculate the following relative risks:

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{34/56}{66/184} \approx 1.69 \quad \text{and} \quad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{22/56}{118/184} \approx 0.61. \quad (4.15)$$

Thus, the proportion of individuals with the disease is 1.69 times higher among smokers when compared with non-smokers. Similarly, the proportion of healthy individuals is 0.61 times smaller among smokers when compared with non-smokers.

The relative risks are calculated to compare the proportion of sick or healthy patients between smokers and non-smokers. Using these two relative risks, the odds ratio is obtained as

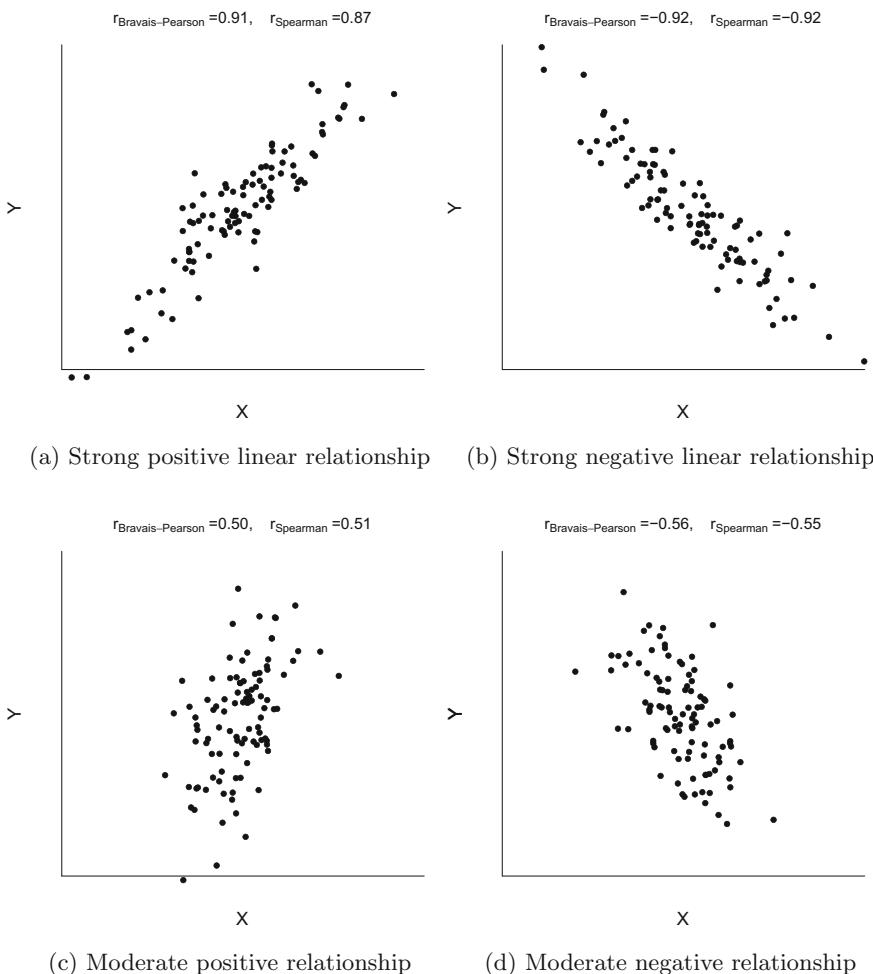
$$OR = \frac{34 \times 118}{66 \times 22} = 2.76.$$

We can interpret this outcome as follows: (i) the chances of smoking are 2.76 times higher for individuals with the disease compared with healthy individuals (follows from definition (4.14)). We can also say that (ii) the chances of having the particular disease is 2.76 times higher for smokers compared with non-smokers. If we interchange either one of the “Yes” and “No” columns or the “Yes” and “No” rows, we obtain  $OR = 1/2.76 \approx 0.36$ , giving us further interpretations: (iii) the chances of smoking are 0.36 times lower for individuals without disease compared with individuals with the disease, and (iv) the chance of having the particular disease is 0.36 times lower for non-smokers compared with smokers. Note that all four interpretations are correct and one needs to choose the right interpretation in the light of the experimental situation and the question of interest.

## 4.3 Association Between Ordinal and Continuous Variables

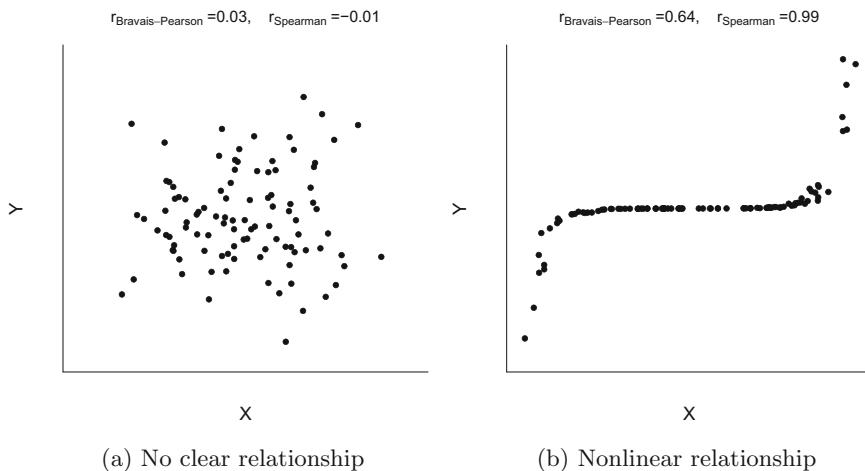
### 4.3.1 Graphical Representation of Two Continuous Variables

A simple way to graphically summarize the association between two continuous variables is to plot the paired observations of the two variables in a two-dimensional coordinate system. If  $n$  paired observations for two continuous variables  $X$  and  $Y$  are available as  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , then all such observations can be plotted

**Fig. 4.2** Scatter plots

in a single graph. This graph is called a **scatter plot**. Such a plot reveals possible relationships and trends between the two variables. For example, Figs. 4.2 and 4.3 show scatter plots with six different types of association.

- Figure 4.2a shows increasing values of  $Y$  for increasing values of  $X$ . We call this relationship positive association. The relationship between  $X$  and  $Y$  is nearly linear because all the points lie around a straight line.
- Figure 4.2b shows decreasing values of  $Y$  for increasing values of  $X$ . We call this relationship negative association.
- Figure 4.2c tells us the same as Fig. 4.2a, except that the positive association is weaker.



**Fig. 4.3** Continues Fig. 4.2—more scatter plots

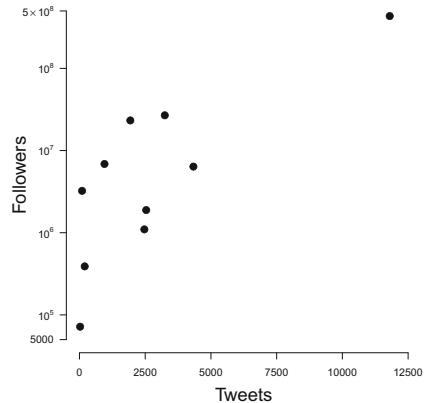
- Figure 4.2d tells us the same as Fig. 4.2b, except that the negative association is weaker.
  - Figure 4.3a shows that as the  $X$ -values increase, the values of  $Y$  neither increase nor decrease. This indicates that there is no clear relationship between  $X$  and  $Y$  and highlights the lack of association between  $X$  and  $Y$ .
  - Figure 4.3b illustrates a nonlinear relationship between  $X$ - and  $Y$ -values.

*Example 4.3.1* To explore the possible relationship between the overall number of tweets with the number of followers on Twitter, we take a sample of 10 prime ministers and heads of state in different countries as of June 2014 and obtain the following data:

Name	Tweets	Followers
Angela Merkel	25	7194
Barack Obama	11,800	43,400,000
Jacob Zuma	99	324,000
Dilma Rousseff	1934	2,330,000
Sauli Niinistö	199	39,000
Vladimir Putin	2539	189,000
Francois Hollande	4334	639,000
David Cameron	952	688,000
Enrique P. Nieto	3245	2,690,000
John Key	2468	110,000

The tweets are denoted by  $x_i$  and the followers are denoted by  $y_i$ ,  $i = 1, 2, \dots, 10$ . We plot paired observations  $(x_i, y_i)$  into a cartesian coordinate system. For example, we plot  $(x_1, y_1) = (25, 7194)$  for Angela Merkel,  $(x_2, y_2) = (11, 800, 43, 400, 000)$

**Fig. 4.4** Scatter plot between tweets and followers



for Barack Obama, and so on. Figure 4.4 shows the scatter plot for the number of tweets and the number of followers (on a log-scale).

One can see that there is a positive association between the number of tweets and the number of followers. This does, however, *not* imply a causal relationship: it is not necessarily *because* someone tweets more he/she has more followers or *because* someone has more followers he/she tweets more; the scatter plot just describes that those with more tweets have more followers. In *R*, we produce this scatter plot by the `plot` command:

```
tweets <- c(25,11800,99,...)
followers <- c(7194,43400000,...)
plot(tweets,followers)
```



### 4.3.2 Correlation Coefficient

Suppose two variables  $X$  and  $Y$  are measured on a continuous scale and are linearly related like  $Y = a + b X$  where  $a$  and  $b$  are constant values. The **correlation coefficient**  $r(X, Y) = r$  measures the degree of *linear* relationship between  $X$  and  $Y$  using

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}, \quad (4.16)$$

with

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = n\tilde{s}_X^2, \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = n\tilde{s}_Y^2, \quad (4.17)$$

and

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}. \quad (4.18)$$

Karl Pearson (1857–1936) presented the first rigorous treatment of correlation and acknowledged Auguste Bravais (1811–1863) for ascertaining the initial mathematical formulae for correlation. This is why the correlation coefficient is also known as the **Bravais–Pearson correlation coefficient**.

The correlation coefficient is independent of the units of measurement of  $X$  and  $Y$ . For example, if someone measures the height and weight in metres and kilograms respectively and another person measures them in centimetres and grams, respectively, then the correlation coefficient between the two sets of data will be the same. The correlation coefficient is symmetric, i.e.  $r(X, Y) = r(Y, X)$ . The limits of  $r$  are  $-1 \leq r \leq 1$ . If all the points in a scatter plot lie exactly on a straight line, then the linear relationship between  $X$  and  $Y$  is perfect and  $|r| = 1$ , see also Exercise 4.7. If the relationship between  $X$  and  $Y$  is (i) perfectly linear and increasing, then  $r = +1$  and (ii) perfectly linear and decreasing, then  $r = -1$ . The signs of  $r$  thus determine the direction of the association. If  $r$  is close to zero, then it indicates that the variables are independent or the relationship is not linear. Note that if the relationship between  $X$  and  $Y$  is nonlinear, then the degree of linear relationship may be low and  $r$  is then close to zero even if the variables are clearly not independent. Note that  $r(X, X) = 1$  and  $r(X, -X) = -1$ .

*Example 4.3.2* Look again at the scatter plots in Figs. 4.2 and 4.3. We observe strong positive linear correlation in Fig. 4.2a ( $r = 0.91$ ), strong negative linear correlation in Fig. 4.2b ( $r = -0.92$ ), moderate positive linear correlation in Fig. 4.2c ( $r = 0.50$ ), moderate negative linear association in Fig. 4.2d ( $r = -0.56$ ), no visible correlation in Fig. 4.3a ( $r = 0.03$ ), and strong nonlinear (but not so strong linear) correlation in Fig. 4.3b ( $r = 0.64$ ).

*Example 4.3.3* In a decathlon competition, a group of athletes are competing with each other in 10 different track and field events. Suppose we are interested in how the results of the 100-m race relate to the results of the long jump competition. The correlation coefficient for the 100-m race ( $X$ , in seconds) and the long jump event ( $Y$ , in metres) for 5 athletes participating in the 2004 Olympic Games (see also Appendix A.4) are listed in Table 4.7.

To calculate the correlation coefficient, we need the following summary statistics:

$$\bar{x} = \frac{1}{5}(10.85 + 10.44 + 10.50 + 10.89 + 10.62) = 10.66$$

$$\bar{y} = \frac{1}{5}(7.84 + 7.96 + 7.81 + 7.47 + 7.74) = 7.764$$

$$S_{xx} = (10.85 - 10.66)^2 + (10.44 - 10.66)^2 + \cdots + (10.62 - 10.66)^2 = 0.1646$$

**Table 4.7** Results of 100-m race and long jump of 5 athletes

$i$	$x_i$	$y_i$
Roman Sebrle	10.85	7.84
Bryan Clay	10.44	7.96
Dmitriy Karpov	10.50	7.81
Dean Macey	10.89	7.47
Chiel Warners	10.62	7.74

$$\begin{aligned} S_{yy} &= (7.84 - 7.764)^2 + (7.96 - 7.764)^2 + \cdots + (7.74 - 7.764)^2 = 0.13332 \\ S_{xy} &= (10.85 - 10.66)(7.84 - 7.764) + \cdots + (10.62 - 10.66)(7.74 - 7.764) \\ &= -0.1027 \end{aligned}$$

The correlation coefficient therefore is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-0.1027}{\sqrt{0.1646 \times 0.13332}} \approx -0.69.$$

Since  $-0.69$  is negative, we can say that (i) there is a negative correlation between the 100-m race and the long jump event, i.e., shorter running times result in longer long jump results, and (ii) this association is moderate to strong.

In R, we can obtain the results (after attaching the data) as follows:

```
cor(X.100m, X.Long.jump, method='pearson')
```



### 4.3.3 Spearman's Rank Correlation Coefficient

Consider a situation where  $n$  objects are ranked with respect to two variables  $X$  and  $Y$ . For instance, the variables could represent the opinion of two different judges in a talent competition who rank the participants with respect to their performance. This means that for each judge, the worst participant (with the lowest score  $x_i$ ) is assigned rank 1, the second worst participant (with the second lowest score  $x_i$ ) will receive rank 2, and so on. Thus, every participant has been given two ranks by two different judges. Suppose we want to measure the degree of association between the two different judgments; that is, the two different sets of ranks. We expect that under perfect agreement, both the judges give the same judgment in the sense that they give the same ranks to each candidate. However, if they are not in perfect agreement, then there may be some variation in the ranks assigned by them. To measure the degree of agreement, or, in general, the degree of association, one can use **Spearman's rank correlation coefficient**. As the name says, this correlation coefficient uses only the ranks of the values and not the values themselves. Thus, this measure is suitable for both ordinal and continuous variables. We introduce the following notations: let  $R(x_i)$  denote the rank of the  $i$ th observation on  $X$ , i.e. the rank  $x_i$  among the ordered values of  $X$ . Similarly,  $R(y_i)$  denotes the rank of the  $i$ th observation of  $y$ . The difference between the two rank values is  $d_i = R(x_i) - R(y_i)$ . Spearman's rank correlation coefficient is defined as

$$R = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}. \quad (4.19)$$

The values of  $R$  lie between  $-1$  and  $+1$  and measure the degree of correlation between the ranks of  $X$  and  $Y$ . Note that it does not matter whether we choose an ascending or descending order of the ranks, the value of  $R$  remains the same. When all the observations are assigned exactly the same ranks, then  $R = 1$  and when all the observations are assigned exactly the opposite ranks, then  $R = -1$ .

*Example 4.3.4* Look again at the scatter plots in Figs. 4.2 and 4.3. We observe strong positive correlation in Fig. 4.2a ( $R = 0.87$ ), strong negative correlation in Fig. 4.2b ( $R = -0.92$ ), moderate positive correlation in Fig. 4.2c ( $R = 0.51$ ), moderate negative association in Fig. 4.2d ( $R = -0.55$ ), no visible correlation in Fig. 4.3a ( $R = -0.01$ ), and strong nonlinear correlation in Fig. 4.3b ( $R = 0.99$ ).

*Example 4.3.5* Let us follow Example 4.3.3 a bit further and calculate Spearman's rank correlation coefficient for the first five observations of the decathlon data. Again we list the results of the 100-m race ( $X$ ) and the results of the long jump competition ( $Y$ ). In addition, we assign ranks to both  $X$  and  $Y$ . For example, the shortest time receives rank 1, whereas the longest time receives rank 5. Similarly, the shortest long jump result receives rank 1, the longest long jump result receives rank 5.

$i$	$x_i$	$R(x_i)$	$y_i$	$R(y_i)$	$d_i$	$d_i^2$
Roman Sebrle	10.85	4	7.84	4	0	0
Bryan Clay	10.44	1	7.96	5	-4	16
Dmitriy Karpov	10.50	2	7.81	3	-1	1
Dean Macey	10.89	5	7.47	1	-4	16
Chiel Warners	10.62	3	7.74	2	-1	1
Total						34

Using (4.19), Spearman's rank correlation coefficient can be calculated as

$$R = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 34}{5 \cdot 24} = -0.7.$$

We therefore have a moderate to strong negative association between the 100-m race and the long jump event. We now know that for the 5 athletes above longer running times relate to shorter jumping distances which in turn means that a good performance in one discipline implies a good performance in the other discipline. In R, we can obtain the same results by using the cor command:

```
cor(X.100m, X.Long.jump, method='spearman')
```



If two or more observations take the same values for  $x_i$  (or  $y_i$ ), then there is a **tie**. In such situations, the respective ranks can simply be averaged, though more complicated solutions also exist (one of which is implemented in the R function cor). For example, if in Example 4.3.5 Bryan Clay's was 10.50 s instead of 10.44 s, then both Bryan Clay and Dmitriy Karpov had the same time. Instead of assigning the ranks 1 and 2 to them, we assign the ranks 1.5 to each of them.

The differences between the correlation coefficient and the rank correlation coefficient are manifold: firstly, Pearson's correlation coefficient can be used for continuous variables only, but not for nominal or ordinal variables. The rank correlation coefficient can be used for either two continuous or two ordinal variables or a combination of an ordinal and a continuous variable, but not for two nominal variables. Moreover, the rank correlation coefficient responds to any type of relationship whereas

Pearson's correlation measures the degree of a linear relationship only—see also Fig. 4.3b. Another difference between the two correlation coefficients is that Pearson uses the entire information contained in the continuous data in contrast to the rank correlation coefficient which uses only ordinal information contained in the ordered data.

#### 4.3.4 Measures Using Discordant and Concordant Pairs

Another concept which uses ranks to measure the association between ordinal variables is based on **concordant** and **discordant** observation pairs. It is best illustrated by means of an example.

*Example 4.3.6* Suppose an online book store conducts a survey on their customer's satisfaction with respect to both the timeliness of deliveries ( $X$ ) and payment options ( $Y$ ). Let us consider the following  $2 \times 3$  contingency table with a summary of the responses of 100 customers. We assume that the categories for both variables can be ordered and ranks can be assigned to different categories, see the numbers in brackets in Table 4.8. There are 100 observation pairs  $(x_i, y_i)$  which summarize the response of the customers with respect to both  $X$  and  $Y$ . For example, there are 18 customers who were unsatisfied with the timeliness of the deliveries and complained that there are not enough payment options. If we compare two responses  $(x_{i_1}, y_{i_1})$  and  $(x_{i_2}, y_{i_2})$ , it might be possible that one customer is more happy (or more unhappy) than the other customer with respect to both  $X$  and  $Y$  or that one customer is more happy with respect to  $X$  but more unhappy with respect to  $Y$  (or vice versa). If the former is the case, then this is a concordant observation pair; if the latter is true, then it is a discordant pair. For instance, a customer who replied “enough” and “satisfied” is more happy than a customer who replied “not enough” and “unsatisfied” because he is more happy with respect to both  $X$  and  $Y$ .

In general, a pair is

- **concordant** if  $i_2 > i_1$  and  $j_2 > j_1$  (or  $i_2 < i_1$  and  $j_2 < j_1$ ),
- **discordant** if  $i_2 < i_1$  and  $j_2 > j_1$  (or  $i_2 > i_1$  and  $j_2 < j_1$ ),
- **tied** if  $i_1 = i_2$  (or  $j_1 = j_2$ ).

**Table 4.8** Payment options and timeliness survey with 100 participating customers

		Timeliness				Total
		Unsatisfied (1)	Satisfied (2)	Very satisfied (3)		
Payment options	Not enough (1)	7	11	26	44	
	Enough (2)	10	15	31	56	
	Total	17	26	57	100	

Obviously, if we have only concordant observations, then there is a strong positive association because a higher value of  $X$  (in terms of the ranking) implies a higher value of  $Y$ . However, if we have only discordant observations, then there is a clear negative association. The measures which are introduced below simply put the number of concordant and discordant pairs into relation. This idea is reflected in **Goodman and Kruskal's  $\gamma$**  which is defined as

$$\gamma = \frac{K}{K + D} - \frac{D}{K + D} = \frac{K - D}{K + D}, \quad (4.20)$$

where

$$K = \sum_{i < m} \sum_{j < n} n_{ij} n_{mn}, \quad D = \sum_{i < m} \sum_{j > n} n_{ij} n_{mn}$$

describe the number of concordant and discordant observation pairs, respectively. An alternative measure is **Stuart's  $\tau_c$**  given as

$$\tau_c = \frac{2 \min(k, l)(K - D)}{n^2(\min(k, l) - 1)}. \quad (4.21)$$

Both measures are standardized to lie between  $-1$  and  $1$ , where larger values indicate a stronger association and the sign indicates the direction of the association.

*Example 4.3.7* Consider Example 4.3.6. A customer who replied “enough” and “satisfied” is more happy than a customer who replied “not enough” and “unsatisfied” because the observation pairs, using ranks, are  $(2, 2)$  and  $(1, 1)$  and therefore  $i_2 > i_1$  and  $j_2 > j_1$ . There are  $7 \times 15$  such pairs. Similarly those who said “not enough” and “unsatisfied” are less happy than those who said “enough” and “very satisfied” ( $7 \times 31$  pairs). Table 4.5 summarizes the comparisons in detail.

Table 4.5a shows that  $(x_1, y_1) = (\text{not enough, unsatisfied})$  is concordant to  $(x_2, y_2) = (\text{enough, satisfied})$  and  $(x_2, y_3) = (\text{enough, very satisfied})$  and tied to  $(x_2, y_1) = (\text{enough, unsatisfied})$ ,  $(x_1, y_2) = (\text{not enough, satisfied})$ , and  $(x_1, y_3) = (\text{not enough, very satisfied})$ . Thus for these comparisons, we have 0 discordant pairs,  $(7 \times 15) + (7 \times 31)$  concordant pairs and  $7 \times (10 + 11 + 26)$  tied pairs. Table 4.5b-f show how the task can be completed. While tiresome, systematically working through the table (and making sure to not count pairs more than once) yields

$$K = 7 \times (15 + 31) + 11 \times 31 = 663$$

$$D = 10 \times (11 + 26) + 15 \times 26 = 760.$$

As a visual rule of thumb, working from the top left to the bottom right yields the concordant pairs; and working from the bottom left to the top right yields the discordant pairs. It follows that  $K = (663 - 760)/(663 + 760) \approx -0.07$  which indicates no clear relationship between the two variables. A similar result is obtained using  $\tau_c$  which is  $4 \times (760 - 663)/100^2 \approx 0.039$ . This rather lengthy task can be made much quicker by using the `ord.gamma` and `ord.tau` commands from the *R* library `youready`:

(a)	$y_1$	$y_2$	$y_3$
$x_1$		t	t
$x_2$	t	c	c

(d)	$y_1$	$y_2$	$y_3$
$x_1$	t	d	d
$x_2$		t	t

(b)	$y_1$	$y_2$	$y_3$
$x_1$	t		t
$x_2$	d	t	c

(e)	$y_1$	$y_2$	$y_3$
$x_1$	c	t	d
$x_2$	t		t

(c)	$y_1$	$y_2$	$y_3$
$x_1$	t	t	
$x_2$	d	d	t

(f)	$y_1$	$y_2$	$y_3$
$x_1$	c	c	t
$x_2$	t		t

**Fig. 4.5** Scheme to visualize concordant ( $c$ ), discordant ( $d$ ), and tied ( $t$ ) pairs in a  $2 \times 3$  contingency table

```
library(ryouready)
ex <- matrix(c(7,11,26,10,15,31),ncol=3,byrow=T)
ord.gamma(ex)
ord.tau(ex)
```



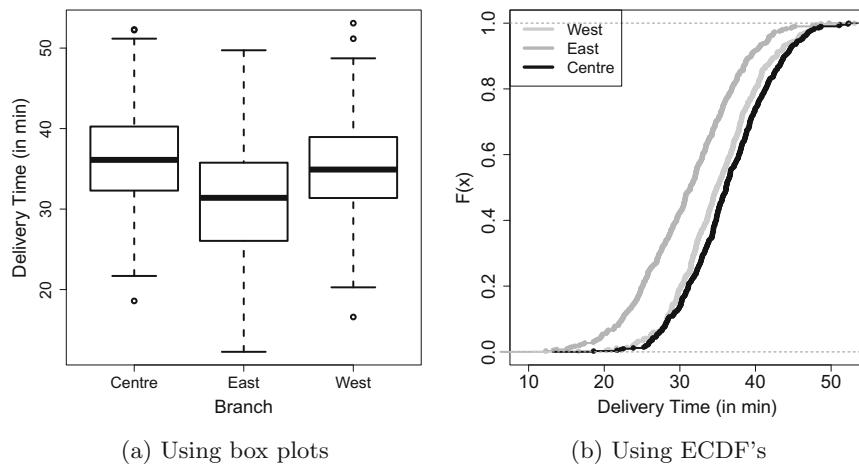
## 4.4 Visualization of Variables from Different Scales

If we want to jointly visualize the association between a variable  $X$ , which is either nominal or ordinal and another variable  $Y$ , which is continuous, then we can use any graph which is suitable for the continuous variable (see Chaps. 2 and 3) and produce it for each category of the nominal/ordinal variable. We recommend using stratified box plots or stratified ECDF's, as they are easy to read when summarized in a single figure; however, it is also possible to place histograms next to each other or on top of each other, or overlay kernel density plots, but we do not illustrate this here in more detail.

*Example 4.4.1* Consider again our pizza delivery example (Appendix A.4). If we are interested in the pizza delivery times by branch, we may simply plot the box plots and ECDF's of delivery time by branch. Figure 4.6 shows that the shortest delivery times can be observed in the branch in the East. Producing these graphs in R is straightforward: The `boxplot` command can be used for two variables by separating them with the `~` sign. For the ECDF, we have to produce a plot for each branch and overlay them with the “`add=TRUE`” option.

```
boxplot(time~branch)
plot.ecdf(time[branch=='East'])
plot.ecdf(time[branch=='West'], add=TRUE)
plot.ecdf(time[branch=='Centre'], add=TRUE)
```



**Fig. 4.6** Distribution of pizza delivery time stratified by branch

## 4.5 Key Points and Further Issues

### Note:

- ✓ How to use different measures of association:
  - 2 nominal variables → Pearson's  $\chi^2$ , relative risks, odds ratio, Cramer's  $V$ , and  $C_{\text{corr}}$
  - 2 ordinal variables → Spearman's rank correlation coefficient,  $\gamma$ ,  $\tau_c$
  - 2 continuous variables → Pearson's correlation coefficient, Spearman's correlation coefficient
- ✓ For two variables which are measured on different scales, for example continuous/ordinal or ordinal/nominal, one should use measures of association suitable for the less informative of the two scales.
- ✓ Another graphical representation of both a continuous and discrete variable is stratified confidence interval plots (error plots), see Chap. 9.

## 4.6 Exercises

*Exercise 4.1* A newspaper asks two of its staff to review the coffee quality at different trendy cafés. The coffee can be rated on a scale from 1 (miserable) to 10 (excellent). The results of the two coffee enthusiasts  $X$  and  $Y$  are as follows:

Café $i$	$x_i$	$y_i$
1	3	6
2	8	7
3	7	10
4	9	8
5	5	4

- (a) Calculate and interpret Spearman's rank correlation coefficient.
- (b) Does Spearman's  $R$  differ depending on whether ranks are assigned in a decreasing or increasing order?
- (c) Suppose the coffee can only be rated as either good ( $>5$ ) or bad ( $\leq 5$ ). Do the chances of a good rating differ between the two journalists?

*Exercise 4.2* A total of 150 customers of a petrol station are asked about their satisfaction with their car and motorbike insurance. The results are summarized below:

	Satisfied	Unsatisfied	Total
Car	33	25	58
Car (diesel engine)	29	31	60
Motorbike	12	20	32
Total	74	76	150

- (a) Determine and interpret Pearson's  $\chi^2$  statistic, Cramer's  $V$ , and  $C_{\text{corr}}$ .
- (b) Combine the categories "car" and "car (diesel engine)" and produce the corresponding  $2 \times 2$  table. Calculate  $\chi^2$  as efficiently as possible and give a meaningful interpretation of the odds ratio.
- (c) Compare the results from (a) and (b).

*Exercise 4.3* There has been a big debate about the usefulness of speed limits on public roads. Consider the following table which lists the speed limits for country roads (in miles/h) and traffic deaths (per 100 million km) for different countries in 1986 when the debate was particularly serious:

- (a) Draw the scatter plot for the two variables.
- (b) Calculate the Bravais–Pearson and Spearman correlation coefficients.

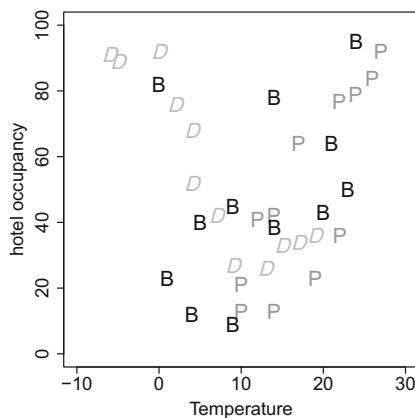
Country	Speed limit	Traffic deaths
Denmark	55	4.1
Japan	55	4.7
Canada	60	4.3
Netherlands	60	5.1
Italy	75	6.1

- (c) What are the effects on the correlation coefficients if the speed limit is given in km/h rather than miles/h ( $1 \text{ mile/h} \approx 1.61 \text{ km/h}$ )?
- (d) Consider one more observation: the speed limit for England was 70 miles/h and the death rate was 3.1.
- (i) Add this observation to the scatter plot.
  - (ii) Calculate the Bravais–Pearson correlation coefficient given this additional observation.

*Exercise 4.4* The famous passenger liner *Titanic* hit an iceberg in 1912 and sank. A total of 337 passengers travelled in first class, 285 in second class, and 721 in third class. In addition, there were 885 staff members on board. Not all passengers could be rescued. Only the following were rescued: 135 from the first class, 160 from the second class, 541 from the third class and 674 staff.

- (a) Determine and interpret the contingency table for the variables “travel class” and “rescue status”.
- (b) Use a contingency table to summarize the conditional relative frequency distributions of rescue status given travel class. Could there be an association of the two variables?
- (c) What would the contingency table from (a) look like under the independence assumption? Calculate Cramer’s  $V$  statistic. Is there any association between travel class and rescue status?
- (d) Combine the categories “first class” and “second class” as well as “third class” and “staff”. Create a contingency table based on these new categories. Determine and interpret Cramer’s  $V$ , the odds ratio, and relative risks of your choice.
- (e) Given the results from (a) to (d), what are your conclusions?

*Exercise 4.5* To study the association of the monthly average temperature (in  $^{\circ}\text{C}$ ,  $X$ ) and hotel occupation (in %,  $Y$ ), we consider data from three cities: Polenca (Mallorca, Spain) as a summer holiday destination, Davos (Switzerland) as a winter skiing destination, and Basel (Switzerland) as a business destination.



**Fig. 4.7** Temperature and hotel occupancy for the different cities

Months	Davos		Polenca		Basel	
	X	Y	X	Y	X	Y
Jan	-6	91	10	13	1	23
Feb	-5	89	10	21	0	82
Mar	2	76	14	42	5	40
Apr	4	52	17	64	9	45
May	7	42	22	79	14	39
Jun	15	36	24	81	20	43
Jul	17	37	26	86	23	50
Aug	19	39	27	92	24	95
Sep	13	26	22	36	21	64
Oct	9	27	19	23	14	78
Nov	4	68	14	13	9	9
Dec	0	92	12	41	4	12

- (a) Calculate the Bravais–Pearson correlation coefficient. The following summary statistics are available:  $\sum_{i=1}^{36} x_i y_i = 22,776$ ,  $\bar{x} = 12.22$ ,  $\bar{y} = 51.28$ ,  $s_x^2 = 76.95$ , and  $s_y^2 = 706.98$ .
- (b) Interpret the scatter plot in Fig. 4.7 which visualizes temperature and hotel occupancy for Davos (D), Polenca (P), and Basel (B).
- (c) Use  $R$  to calculate the correlation coefficient separately for each city. Interpret the results and discuss the use of the correlation coefficient if more than two variables are available.

*Exercise 4.6* Consider a neighbourhood survey on the use of a local park. Respondents were asked whether the park may be used for summer music concerts and whether dog owners should put their dogs on a lead. The results are summarized in the following contingency table:

		Put dogs on a lead			Total
		Agree	No opinion	Disagree	
Use for concerts	Agree	82	4	0	86
	No opinion	8	43	9	60
	Disagree	0	2	10	12
	Total	90	49	19	158

- (a) Calculate and interpret Goodman and Kruskal's  $\gamma$ .
- (b) Now ignore the ordinal structure of the data and calculate Cramer's  $V$ .
- (c) Create the contingency table which is obtained when the categories "no opinion" and "agree" are combined.
- (d) What is the relative risk of disagreement with summer concerts depending on the opinion about using leads?
- (e) Calculate the odds ratio and offer two interpretations of it.
- (f) Determine  $\gamma$  for the table calculated in (c).
- (g) What is your final interpretation and what may be the best measure to use in this example?

*Exercise 4.7* Consider  $n$  observations for which  $y_i = a + bx_i$ ,  $b > 0$ , holds. Show that  $r = 1$ .

*Exercise 4.8* Make yourself familiar with the Olympic decathlon data described in Appendix A.4. Read in and attach the data in *R*.

- (a) Use *R* to calculate and interpret the Bravais–Pearson correlation coefficient between the results of the discus and the shot-put events.
- (b) There are 10 continuous variables. How many different correlation coefficients can you calculate? How would you summarize them?
- (c) Apply the `cor` command to the whole data and interpret the output.
- (d) Omit the two rows which contain missing data and interpret the output again.

*Exercise 4.9* We are interested in the pizza delivery data which is described in Appendix A.4.

- (a) Read in the data and create two new binary variables which describe whether a pizza was hot ( $>65^\circ\text{C}$ ) and the delivery time was short ( $<30\text{ min}$ ). Create a contingency table for the two new variables.
- (b) Calculate and interpret the odds ratio for the contingency table from (a).

- (c) Use Cramer's  $V$ , Stuart's  $\tau_c$ , Goodman and Kruskal's  $\gamma$ , and a stacked bar chart to explore the association between the categorical time and temperature variables.
- (d) Draw a scatter plot for the continuous time and temperature variables. Determine both the Bravais–Pearson and Spearman correlation coefficients.
- (e) Use methods of your choice to explore the relationship between temperature and driver, operator, number of ordered pizzas and bill. Is it clear which of the variables influence the pizza temperature?

→ Solutions to all exercises in this chapter can be found on p. [345](#)

---

**Part II**

**Probability Calculus**

---

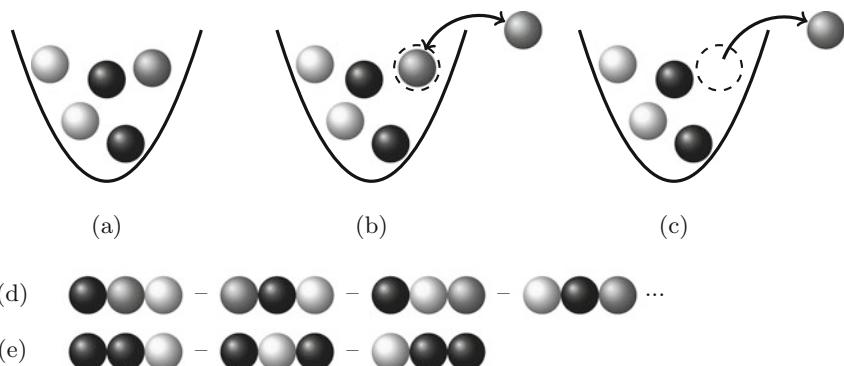
## 5.1 Introduction

Combinatorics is a special branch of mathematics. It has many applications not only in several interesting fields such as enumerative combinatorics (the classical application), but also in other fields, for example in graph theory and optimization.

First, we try to motivate and understand the role of combinatorics in statistics. Consider a simple example in which someone goes to a cafe. The person would like a hot beverage and a cake. Assume that one can choose among three different beverages, for example cappuccino, hot chocolate, and green tea, and three different cakes, let us say carrot cake, chocolate cake, and lemon tart. The person may consider different beverage and cake combinations when placing the order, for example carrot cake and cappuccino, carrot cake and tea, and hot chocolate and lemon tart. From a statistical perspective, the customer is evaluating the possible combinations before making a decision. Depending on their preferences, the order will be placed by choosing one of the combinations.

In this example, it is easy to calculate the number of possible combinations. There are three different beverages and three different cakes to choose from, leading to nine different ( $3 \times 3$ ) beverage and cake combinations. However, suppose there is a choice of 15 hot beverages and 8 different cakes. How many orders can be made? (Answer:  $15 \times 8$ ) What if the person decides to order two cakes, how will it affect the number of possible combinations of choices? It will be a tedious task to count all the possibilities. So we need a systematic approach to count such possible combinations. Combinatorics deals with the counting of different possibilities in a systematic approach.

People often use the urn model to understand the system in the counting process. The urn model deals with the drawing of balls from an urn. The balls in the urn



**Fig. 5.1** a Representation of the urn model. Drawing from the urn model b with replacement and c without replacement. Compositions of three drawn balls: d all balls are distinguishable and e some balls are not distinguishable

represent the units of a population, or the features of a population. The balls may vary in colour or size to represent specific properties of a unit or feature. We illustrate this concept in more detail in Fig. 5.1.

Suppose there are 5 balls of three different colours—two black, one grey, and two white (see Fig. 5.1a). This can be generalized to a situation in which there are  $n$  balls in the urn and we want to draw  $m$  balls. Suppose we want to know

- how many different possibilities exist to draw  $m$  out of  $n$  balls (thus determining the number of distinguishable **combinations**).

To deal with such a question, we first need to decide whether a ball will be put back into the urn after it is drawn or not. Figure 5.1b illustrates that a grey ball is drawn from the urn and then placed back (illustrated by the two-headed arrow). We say the ball is drawn *with replacement*. Figure 5.1c illustrates a different situation in which the grey ball is drawn from the urn and is *not* placed back into the urn (illustrated by the one-headed arrow). We say the ball is drawn *without replacement*.

Further, we may be interested in knowing the

- total number of ways in which the chosen set of balls can be arranged in a distinguishable order (which we will define as **permutations** later in this chapter).

To answer the question how many permutations exist, we first need to decide whether all the chosen balls are distinguishable from each other or not. For example, in Fig. 5.1d, the three chosen balls have different colours; therefore, they are distinguishable. There are many options on how they can be arranged. In contrast, some

of the chosen balls in Fig. 5.1e are the same colour, they are therefore not distinguishable. Consequently, the number of combinations is much more limited. The concept of balls and urns just represents the features of observations from a sample. We illustrate this in more detail in the following example.

*Example 5.1.1* Say a father promises his daughter three scoops of ice cream if she cleans up her room. For simplicity, let us assume the daughter has a choice of four flavours: chocolate, banana, cherry, and lemon. How many different choices does the daughter have? If each scoop has to be a different flavour she obviously has much less choice than if the scoops can have the same flavour. In the urn model, this is represented by the concept of “with/without replacement”. The urn contains 4 balls of 4 different colours which represent the ice cream flavours. For each of the three scoops, a ball is drawn to determine the flavour. If we draw with replacement, each flavour can be potentially chosen multiple times; however, if we draw without replacement each flavour can be chosen only once. Then, the number of possible combinations is easy to calculate: it is 4, i.e. (chocolate, banana, and cherry); (chocolate, banana, and lemon); (chocolate, cherry, and lemon); and (banana, cherry, and lemon). But what if we have more choices? Or if we can draw flavours multiple times? We then need calculation rules which help us counting the number of options.

Now, let us assume that the daughter picked the flavours (chocolate [C], banana [B], and lemon [L]). Like many other children, she prefers to eat her most favourite flavour (chocolate) last, and her least favourite flavour (cherry) first. Therefore, the order in which the scoops are placed on top of the cone are important! In how many different ways can the scoops be placed on top of the cone? This relates to the question of the number of distinguishable *permutations*. The answer is 6: (C,B,L)–(C,L,B)–(B,L,C)–(B,C,L)–(L,B,C)–(L,C,B). But what if the daughter did pick a flavour multiple times, e.g. (chocolate, chocolate, lemon)? Since the two chocolate scoops are non-distinguishable, there are fewer permutations: (chocolate, chocolate, and lemon)–(chocolate, lemon, and chocolate)–(lemon, chocolate, and chocolate).

The bottom line of this example is that the number of combinations/options is determined by (i) whether we draw with or without replacement (i.e. allow flavours to be chosen more than once) and (ii) whether the arrangement in a particular order (=permutation) is of any specific interest.

Consider the urn example again. Suppose three balls of different colours, black, grey, and white, are drawn. Now there are two options: The first option is to take into account the order in which the balls are drawn. In such a situation, two possible sets of balls such as (black, grey, and white) and (white, black, and grey) constitute two different sets. Such a set is called an *ordered set*. In the second option, we do not take into account the order in which the balls are drawn. In such a situation, the two possible sets of balls such as (black, grey, and white) and (white, black, and grey) are the same sets and constitute an *unordered set* of balls.

**Definition 5.1.1** A group of elements is said to be **ordered** if the order in which these elements are drawn is of relevance. Otherwise, it is called **unordered**.

*Examples.*

- Ordered samples:

- The first three places in an Olympic 100 m race are determined by the order in which the athletes arrive at the finishing line. If 8 athletes are competing with each other, the number of possible results for the first three places is of interest. In the urn language, we are taking draws without replacement (since every athlete can only have one distinct place).
- In a raffle with two prizes, the first drawn raffle ticket gets the first prize and the second raffle ticket gets the second prize.
- There exist various esoteric tarot card games which claim to foretell someone's fortune with respect to several aspects of life. The order in which the cards are shown on the table is important for the interpretation.

- Unordered samples:

- The selected members for a national football team. The order in which the selected names are announced is irrelevant.
- Out of 10 economists, 10 medical doctors, and 10 statisticians, an advisory committee consisting of 4 economists, 3 medical doctors, and 2 statisticians is elected.
- Fishing 20 fish from a lake.
- A bunch of 10 flowers made from 21 flowers of 4 different colours.

**Definition 5.1.2** The factorial function  $n!$  is defined as

$$n! = \begin{cases} 1 & \text{for } n = 0 \\ 1 \cdot 2 \cdot 3 \cdots n & \text{for } n > 0. \end{cases} \quad (5.1)$$

*Example 5.1.2* It follows from the definition of the factorial function that

$$0! = 1, \quad 1! = 1, \quad 2! = 1 \cdot 2 = 2, \quad 3! = 1 \cdot 2 \cdot 3 = 6.$$

This can be calculated in *R* as follows:

`factorial(n)`



## 5.2 Permutations

**Definition 5.2.1** Consider a set of  $n$  elements. Each ordered composition of these  $n$  elements is called a **permutation**.

We distinguish between two cases: If all the elements are distinguishable, then we speak of *permutation without replacement*. However, if some or all of the elements are not distinguishable, then we speak of *permutation with replacement*. Please note that the meaning of “replacement” here is just a convention and does not directly refer to the drawings, e.g. from the urn model considered in Example 5.1.1.

### 5.2.1 Permutations without Replacement

If all the  $n$  elements are distinguishable, then there are

$$n! \quad (5.2)$$

different compositions of these elements.

*Example 5.2.1* There were three candidate cities for hosting the 2020 Olympic Games: Tokyo (T), Istanbul (I), and Madrid (M). Before the election, there were  $3! = 6$  possible outcomes, regarding the final rankings of the cities:

$$(M, T, I), (M, I, T), (T, M, I), (T, I, M), (I, M, T), (I, T, M).$$

### 5.2.2 Permutations with Replacement

Assume that not all  $n$  elements are distinguishable. The elements are divided into groups, and these groups are distinguishable. Suppose, there are  $s$  groups of sizes  $n_1, n_2, \dots, n_s$ . The total number of different ways to arrange the  $n$  elements in  $s$  groups is:

$$\frac{n!}{n_1! n_2! n_3! \cdots n_s!} . \quad (5.3)$$

*Example 5.2.2* Consider the data in Fig. 5.1e. There are two groups consisting of two black balls ( $n_1 = 2$ ) and one white ball ( $n_2 = 1$ ). So there are the following three possible combinations to arrange the balls: (black, black, and white), (black, white, and black), and (white, black, and black). This can be determined by calculating

$$\frac{3!}{2! 1!} = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 1} = 3 .$$

### 5.3 Combinations

**Definition 5.3.1** The Binomial coefficient for any integers  $m$  and  $n$  with  $n \geq m \geq 0$  is denoted and defined as

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}. \quad (5.4)$$

It is read as “ $n$  choose  $m$ ” and can be calculated in *R* using the following command:

`choose(n, m)`



There are several calculation rules for the binomial coefficient:

$$\binom{n}{0} = 1, \quad \binom{n}{1} = n, \quad \binom{n}{m} = \binom{n}{n-m}, \quad \binom{n}{m} = \prod_{i=1}^m \frac{n+1-i}{i}. \quad (5.5)$$

We now answer the question of how many different possibilities exist to draw  $m$  out of  $n$  elements, i.e.  $m$  out of  $n$  balls from an urn. It is necessary to distinguish between the following four cases:

- (1) Combinations **without** replacement and **without** consideration of the order of the elements.
- (2) Combinations **without** replacement and **with** consideration of the order of the elements.
- (3) Combinations **with** replacement and **without** consideration of the order of the elements.
- (4) Combinations **with** replacement and **with** consideration of the order of the elements.

#### 5.3.1 Combinations without Replacement and without Consideration of the Order

When there is no replacement and the order of the elements is also not relevant, then the total number of distinguishable combinations in drawing  $m$  out of  $n$  elements is

$$\binom{n}{m}. \quad (5.6)$$

*Example 5.3.1* Suppose a company elects a new board of directors. The board consists of 5 members and 15 people are eligible to be elected. How many combinations for the board of directors exist? Since a person cannot be elected twice, we have a

situation where there is no replacement. The order is also of no importance: either one is elected or not. We can thus apply (5.6) which yields

$$\binom{15}{5} = \frac{15!}{10!5!} = 3003$$

possible combinations. This result can be obtained in *R* by using the command `choose(15, 5)`.

### 5.3.2 Combinations without Replacement and with Consideration of the Order

The total number of different combinations for the setting without replacement and with consideration of the order is

$$\frac{n!}{(n-m)!} = \binom{n}{m} m! . \quad (5.7)$$

*Example 5.3.2* Consider a horse race with 12 horses. A possible bet is to forecast the winner of the race, the second horse of the race, and the third horse of the race. The total number of different combinations for the horses in the first three places is

$$\frac{12!}{(12-3)!} = 12 \cdot 11 \cdot 10 = 1320 .$$

This result can be explained intuitively: for the first place, there is a choice of 12 different horses. For the second place, there is a choice of 11 different horses (12 horses minus the winner). For the third place, there is a choice of 10 different horses (12 horses minus the first and second horses). The total number of combinations is the product  $12 \cdot 11 \cdot 10$ . This can be calculated in *R* as follows:

`12 * 11 * 10`



### 5.3.3 Combinations with Replacement and without Consideration of the Order

The total number of different combinations with replacement and without consideration of the order is

$$\binom{n+m-1}{m} = \frac{(n+m-1)!}{m! (n-1)!} = \binom{n+m-1}{n-1} . \quad (5.8)$$

Note that these are the two representations which follow from the definition of the binomial coefficient but typically only the first representation is used in textbooks. We will motivate the second representation after Example 5.3.3.

*Example 5.3.3* A farmer has 2 fields and aspires to cultivate one out of 4 different organic products per field. Then, the total number of choices he has is

$$\binom{4+2-1}{2} = \binom{5}{2} = \frac{5!}{2! \cdot 3!} = \frac{3! \cdot 4 \cdot 5}{1 \cdot 2 \cdot 3!} = 10. \quad (5.9)$$

If 4 different organic products are denoted as a, b, c, and d, then the following combinations are possible:

$$\begin{array}{cccc} (a, a) & (a, b) & (a, c) & (a, d) \\ (b, b) & (b, c) & (b, d) & \\ (c, c) & (c, d) & & \\ (d, d) & & & \end{array}$$

Please note that, for example, (a,b) is identical to (b,a) because the order in which the products a and b are cultivated on the first or second field is not important in this example.

We now try to give an intuitive explanation of formula (5.9) using Example 5.3.3. We have  $n = 4$  products and  $m = 2$  fields and apply the following technical “trick”: we sort the combinations by the product symbols (a, b, c, or d). When we switch from one product to the next (e.g. from b to c), we make a note by adding a vertical line |. Whenever a product is skipped, we add a line too. For example, the combination (a, c) is denoted by  $a||c|$ , the combination (d, d) by  $|||dd$ , (c, c) by  $||cc|$ , and (a, a) by  $aa||$ . Therefore, the number of characters equates to the 2 chosen symbols of the set (a, b, c, d) plus the 3 vertical lines, in summary  $(4+2)-1=5$  places where  $3=n-1$  places are selected for the vertical line |. How many different line/letter combinations exist? There are 3 out of 5 possible positions for |, i.e.  $\binom{5}{3}=10$  possible combinations, and this is nothing but the right-hand side of (5.9).

### 5.3.4 Combinations with Replacement and with Consideration of the Order

The total number of different combinations for the integers  $m$  and  $n$  with replacement and when the order is of relevance is

$$n^m. \quad (5.10)$$

*Example 5.3.4* Consider a credit card with a four-digit personal identification number (PIN) code. The total number of possible combinations for the PIN is

$$n^m = 10^4 = 10,000.$$

Note that every digit in the first, second, third, and fourth places ( $m = 4$ ) can be chosen out of ten digits from 0 to 9 ( $n = 10$ ).

## 5.4 Key Points and Further Issues

### Note:

- ✓ The rules of combinatorics are as follows:

Combinations	without replacement	with replacement
without order	$\binom{n}{m}$	$\binom{n+m-1}{m}$
with order	$\binom{n}{m}m!$	$n^m$

- ✓ Combinations with and without *replacement* are also often called combinations with and without *repetition*.
- ✓ The permutation rules are as follows:

	without replacement	with replacement
Permutations	$n!$	$\frac{n!}{n_1! \cdots n_s!}$

## 5.5 Exercises

*Exercise 5.1* At a party with 10 guests, every guest shakes hands with each other guest. How many handshakes can be counted in total?

*Exercise 5.2* A language teacher is concerned about the vocabularies of his students. He thus tests 5 students in each lecture. What are the total number of possible combinations

- if a student is tested only once per lecture and
- if a student is tested more than once per lecture?

Use  $R$  to quantify numbers which you cannot calculate manually.

*Exercise 5.3* “Gobang” is a popular game in which two players set counters on a board with 381 knots. One needs to place 5 consecutive counters in a row to win the game. There are also rules on how to remove counters from the other player. Consider a match where 64 counters have already been placed on the board. How many possible combinations exist to place 64 counters on the board?

**Exercise 5.4** A shop offers a special tray of beer: “Munich’s favourites”. Customers are allowed to fill the tray, which holds 20 bottles, with any combination of Munich’s 6 most popular beers (from 6 different breweries).

- (a) What are the number of possible combinations to fill the tray?  
 (b) A customer insists of having at least one beer from each brewery in his tray.  
 How many options does he have to fill the tray?

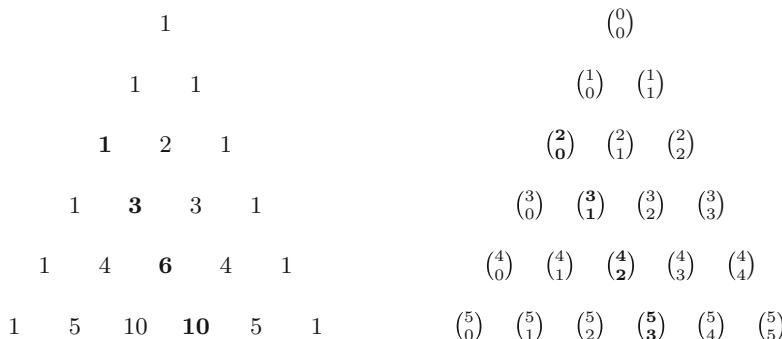
*Exercise 5.5* The FIFA World Cup 2018 in Russia consists of 32 teams. How many combinations for the top 3 teams exist when

- (a) taking into account the order of these top 3 teams and  
 (b) *without* taking into account the order of these top 3 teams?

**Exercise 5.6** An online book store assigns membership codes to each member. For administrative reasons, these codes consist of four letters between “A” and “L”. A special discount period increased the total number of members from 18,200 to 20,500. Are there enough combinations of codes left to be assigned for the new membership codes?

**Exercise 5.7** In the old scoring system of ice skating (valid until 2004), each member of a jury of 9 people judged the performance of the skaters on a scale between 0 and 6. It was a decimal scale and thus scores such as 5.1 and 5.2 were possible. Calculate the number of possible score combinations from the jury.

*Exercise 5.8* It is possible in Pascal's triangle (Fig. 5.2, left) to view each entry as the sum of the two entries directly above it. For example, the 3 on the fourth line



**Fig. 5.2** Excerpt from Pascal's triangle (*left*) and its representation by means of binomial coefficients (*right*)

from the top is the sum of the 1 and 2 above the 3. Another interpretation refers to a geometric representation of the binomial coefficient,  $\binom{n}{k}$  (Fig. 5.2, right) with  $k = 0, 1, 2, \dots$  being the column index and  $n = 0, 1, 2, \dots$  being the row index.

- (a) Show that each entry in the bold third diagonal line can be represented via  $\binom{n}{2}$ .
- (b) Now show that the sum of two consecutive entries in the bold third diagonal line always corresponds to quadratic numbers.

→ Solutions to all exercises in this chapter can be found on p. [358](#)

Let us first consider some simple examples to understand the need for probability theory. Often one needs to make a decision whether to carry an umbrella or not when leaving the house; a company might wonder whether to introduce a new advertisement to possibly increase sales or to continue with their current advertisement; or someone may want to choose a restaurant based on where he can get his favourite dish. In all these situations, randomness is involved. For example, the decision of whether to carry an umbrella or not is based on the possibility or chance of rain. The sales of the company may increase, decrease, or remain unchanged with a new advertisement. The investment in a new advertising campaign may therefore only be useful if the probability of its success is higher than that of the current advertisement. Similarly, one may choose the restaurant where one is most confident of getting the food of one's choice. In all such cases, an event may be happening or not and depending on its likelihood, actions are taken. The purpose of this chapter is to learn how to calculate such likelihoods of events happening and not happening.

---

## 6.1 Basic Concepts and Set Theory

A simple (not rigorous) definition of a **random experiment** requires that the experiment can be repeated any number of times under the same set of conditions, and its outcome is known only after the completion of the experiment. A simple and classical example of a random experiment is the tossing of a coin or the rolling of a die. When tossing a coin, it is unknown what the outcome will be, head or tail, until the coin is tossed. The experiment can be repeated and different outcomes may be observed in each repetition. Similarly, when rolling a die, it is unknown how many dots will appear on the upper surface until the die is rolled. Again, the die can be rolled repeatedly and different numbers of dots are obtained in each trial. A possible

outcome of a random experiment is called a **simple event** (or **elementary event**) and denoted by  $\omega_i$ . The set of all possible outcomes,  $\{\omega_1, \omega_2, \dots, \omega_k\}$ , is called the **sample space** and is denoted as  $\Omega$ , i.e.  $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ . Subsets of  $\Omega$  are called **events** and are denoted by capital letters such as  $A, B, C$ . The set of all simple events that are contained in the event  $A$  is denoted by  $\Omega_A$ . The event  $\bar{A}$  refers to the non-occurring of  $A$  and is called a **composite or complementary event**. Also  $\Omega$  is an event. Since it contains all possible outcomes, we say that  $\Omega$  will always occur and we call it a **sure event** or **certain event**. On the other hand, if we consider the null set  $\emptyset = \{\}$  as an event, then this event can never occur and we call it an **impossible event**. The sure event therefore is the set of all elementary events, and the impossible event is the set with no elementary events.

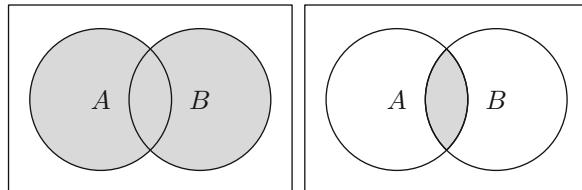
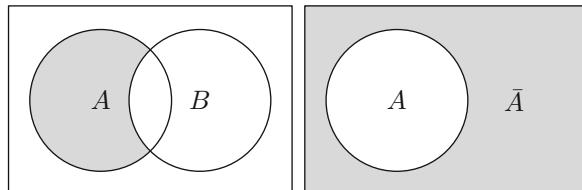
The above concepts of “events” form the basis of a definition of “probability”. Once we understand the concept of probability, we can develop a framework to make conclusions about the population of interest, using a sample of data.

*Example 6.1.1 (Rolling a die)* If a die is rolled once, then the possible outcomes are the number of dots on the upper surface: 1, 2, ..., 6. Therefore, the sample space is the set of simple events  $\omega_1 = "1"$ ,  $\omega_2 = "2"$ , ...,  $\omega_6 = "6"$  and  $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$ . Any subset of  $\Omega$  can be used to define an event. For example, an event  $A$  may be “an even number of dots on the upper surface of the die”. There are three possibilities that this event occurs:  $\omega_2, \omega_4$ , or  $\omega_6$ . If an odd number shows up, then the composite event  $\bar{A}$  occurs instead of  $A$ . If an event is defined to observe only one particular number, say  $\omega_1 = "1"$ , then it is an elementary event. An example of a sure event is “a number which is greater than or equal to 1” because any number between 1 and 6 is greater than or equal to 1. An impossible event is “the number is 7”.

*Example 6.1.2 (Rolling two dice)* Suppose we throw two dice simultaneously and an event is defined as the “number of dots observed on the upper surface of both the dice”; then, there are 36 simple events defined as (number of dots on first die, number of dots on second die), i.e.  $\omega_1 = (1, 1)$ ,  $\omega_2 = (1, 2)$ , ...,  $\omega_{36} = (6, 6)$ . Therefore  $\Omega$  is

$$\Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) \\ (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) \\ (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) \\ (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) \\ (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \\ (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}.$$

One can define different events and their corresponding sample spaces. For example, if an event  $A$  is defined as “upper faces of both the dice contain the same number of dots”, then the sample space is  $\Omega_A = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$ . If another event  $B$  is defined as “the sum of numbers on the upper faces is 6”, then

**Fig. 6.1**  $A \cup B$  and  $A \cap B^*$ **Fig. 6.2**  $A \setminus B$  and  $\bar{A} = \Omega \setminus A^*$ 

the sample space is  $\Omega_B = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$ . A sure event is “get either an even number or an odd number”; an impossible event would be “the sum of the two dice is greater than 13”.

It is possible to view events as sets of simple events. This helps to determine how different events relate to each other. A popular technique to visualize this approach is to use **Venn diagrams**. In Venn diagrams, two or more sets are visualized by circles. Overlapping circles imply that both events have one or more identical simple events. Separated circles mean that none of the simple events of event  $A$  are contained in the sample space of  $B$ . We use the following notations:

- $A \cup B$  The union of events  $A \cup B$  is the set of all simple events of  $A$  and  $B$  which occurs if at least one of the simple events of  $A$  or  $B$  occurs (Fig. 6.1, left side, grey shaded area). Please note that we use the word “or” from a statistical perspective: “ $A$  or  $B$ ” means that either a simple event from  $A$  occurs, or a simple event from  $B$  occurs, or a simple event which is part of both  $A$  and  $B$  occurs.
- $A \cap B$  The intersection of events  $A \cap B$  is the set of all simple events  $A$  and  $B$  which occur when a simple event occurs that belongs to  $A$  and  $B$  (Fig. 6.1, right side, grey shaded area).
- $A \setminus B$  The event  $A \setminus B$  contains all simple events of  $A$ , which are not contained in  $B$ . The event “ $A$  but not  $B$ ” or “ $A$  minus  $B$ ” occurs, if  $A$  occurs but  $B$  does not occur. Also  $A \setminus B = A \cap \bar{B}$  (Fig. 6.2, left side, grey shaded area).
- $\bar{A}$  The event  $\bar{A}$  contains all simple events of  $\Omega$ , which are not contained in  $A$ . The complementary event of  $A$  (which is “Not- $A$ ” or “ $\bar{A}$ ”) occurs whenever  $A$  does not occur (Fig. 6.2, right side, grey shaded area).
- $A \subseteq B$   $A$  is a subset of  $B$ . This means that all simple events of  $A$  are also part of the sample space of  $B$ .

*Example 6.1.3* Consider Example 6.1.1 where the sample space of rolling a die was determined as  $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$  with  $\omega_1 = "1"$ ,  $\omega_2 = "2"$ ,  $\dots$ ,  $\omega_6 = "6"$ .

- If  $A = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$  and  $B$  is the set of all odd numbers, then  $B = \{\omega_1, \omega_3, \omega_5\}$  and thus  $B \subseteq A$ .
- If  $A = \{\omega_2, \omega_4, \omega_6\}$  is the set of even numbers and  $B = \{\omega_3, \omega_6\}$  is the set of all numbers which are divisible by 3, then  $A \cup B = \{\omega_2, \omega_3, \omega_4, \omega_6\}$  is the collection of simple events for which the number is either even or divisible by 3 or both.
- If  $A = \{\omega_1, \omega_3, \omega_5\}$  is the set of odd numbers and  $B = \{\omega_3, \omega_6\}$  is the set of the numbers which are divisible by 3, then  $A \cap B = \{\omega_3\}$  is the set of simple events in which the numbers are odd and divisible by 3.
- If  $A = \{\omega_1, \omega_3, \omega_5\}$  is the set of odd numbers and  $B = \{\omega_3, \omega_6\}$  is the set of the numbers which are divisible by 3, then  $A \setminus B = \{\omega_1, \omega_5\}$  is the set of simple events in which the numbers are odd but not divisible by 3.
- If  $A = \{\omega_2, \omega_4, \omega_6\}$  is the set of even numbers, then  $\bar{A} = \{\omega_1, \omega_3, \omega_5\}$  is the set of odd numbers.

*Remark 6.1.1* Some textbooks also use the following notations:

$$\begin{array}{ll} A + B & \text{for } A \cup B \\ AB & \text{for } A \cap B \\ A - B & \text{for } A \setminus B. \end{array}$$

We can use these definitions and notations to derive the following properties of a particular event  $A$ :

$$\begin{array}{ll} A \cup A = A & A \cap A = A \\ A \cup \Omega = \Omega & A \cap \Omega = A \\ A \cup \emptyset = A & A \cap \emptyset = \emptyset \\ A \cup \bar{A} = \Omega & A \cap \bar{A} = \emptyset. \end{array}$$

**Definition 6.1.1** Two events  $A$  and  $B$  are *disjoint* if  $A \cap B = \emptyset$  holds, i.e. if both events cannot occur simultaneously.

*Example 6.1.4* The events  $A$  and  $\bar{A}$  are disjoint events.

**Definition 6.1.2** The events  $A_1, A_2, \dots, A_m$  are said to be mutually or pairwise disjoint, if  $A_i \cap A_j = \emptyset$  whenever  $i \neq j = 1, 2, \dots, m$ .

*Example 6.1.5* Recall Example 6.1.1. If  $A = \{\omega_1, \omega_3, \omega_5\}$  and  $B = \{\omega_2, \omega_4, \omega_6\}$  are the sets of odd and even numbers, respectively, then the events  $A$  and  $B$  are disjoint.

**Definition 6.1.3** The events  $A_1, A_2, \dots, A_m$  form a **complete decomposition** of  $\Omega$  if and only if

$$A_1 \cup A_2 \cup \dots \cup A_m = \Omega$$

and

$$A_i \cap A_j = \emptyset \quad (\text{for all } i \neq j).$$

*Example 6.1.6* Consider Example 6.1.1. The elementary events  $A_1 = \{\omega_1\}$ ,  $A_2 = \{\omega_2\}, \dots, A_6 = \{\omega_6\}$  form a complete decomposition. Other complete decompositions are, e.g.

- $A_1 = \{\omega_1, \omega_3, \omega_5\}$ ,  $A_2 = \{\omega_2, \omega_4, \omega_6\}$
- $A_1 = \{\omega_1\}$ ,  $A_2 = \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
- $A_1 = \{\omega_1, \omega_2, \omega_3\}$ ,  $A_2 = \{\omega_4, \omega_5, \omega_6\}$ .

## 6.2 Relative Frequency and Laplace Probability

There is a close connection between the relative frequency and the probability of an event. A random experiment is described by its possible outcomes, for example getting a number between 1 and 6 when rolling a die. Suppose an experiment has  $m$  possible outcomes (events)  $A_1, A_2, \dots, A_m$  and the experiment is repeated  $n$  times. Now we can count how many times each of the possible outcome has occurred. In other words, we can calculate the absolute frequency  $n_i = n(A_i)$  which is equal to the number of times an event  $A_i$ ,  $i = 1, 2, \dots, m$ , occurs. The relative frequency  $f_i = f(A_i)$  of a random event  $A_i$ , with  $n$  repetitions of the experiment, is calculated as

$$f_i = f(A_i) = \frac{n_i}{n}. \quad (6.1)$$

*Example 6.2.1* Consider roulette, a game frequently played in casinos. The roulette table consists of 37 numbers from 0 to 36. Out of these 37 numbers, 18 numbers are red, 18 are black and one (zero) is green. Players can place their bets on either a single number or a range of numbers, the colours red or black, whether the number is odd or even, among many other choices. A casino employee spins a wheel (containing pockets representing the 37 numbers) in one direction and then spins a ball over the wheel in the opposite direction. The wheel and ball gradually slow down and the ball finally settles in a pocket. The pocket number in which the ball sits down when the wheel stops is the winning number. Consider three possible outcomes  $A_1$ : “red”,  $A_2$ : “black”, and  $A_3$ : “green (zero)”. Suppose the roulette ball is spun  $n = 500$  times. All the outcomes are counted and recorded as follows:  $A_1$  occurs 240 times,  $A_2$  occurs 250 times and  $A_3$  occurs 10 times. Then, the absolute frequencies are given

by  $n_1 = n(A_1) = 240$ ,  $n_2 = n(A_2) = 250$ , and  $n_3 = n(A_3) = 10$ . We therefore get the relative frequencies as

$$f_1 = f(A_1) = \frac{240}{500} = 0.48, \quad f_2 = f(A_2) = \frac{250}{500} = 0.5,$$

$$f_3 = f(A_3) = \frac{10}{500} = 0.02.$$

If we assume that the experiment is repeated a large number of times (mathematically, this would mean that  $n$  tends to infinity) and the experimental conditions remain the same (at least approximately) over all the repetitions, then the relative frequency  $f(A)$  converges to a limiting value for  $A$ . This limiting value is interpreted as the probability of  $A$  and denoted by  $P(A)$ , i.e.

$$P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n}$$

where  $n(A)$  denotes the number of times an event  $A$  occurs out of  $n$  times.

*Example 6.2.2* Suppose a fair coin is tossed  $n = 20$  times and we observe the number of heads  $n(A_1) = 8$  times and number of tails  $n(A_2) = 12$  times. The meaning of a fair coin in this case is that the probabilities of head and tail are equal (i.e. 0.5). Then, the relative frequencies in the experiment are  $f(A_1) = 8/20 = 0.4$  and  $f(A_2) = 12/20 = 0.6$ . When the coin is tossed a large number of times and  $n$  tends to infinity, then both  $f(A_1)$  and  $f(A_2)$  will have a limiting value 0.5 which is simply the probability of getting a head or tail in tossing a fair coin.

*Example 6.2.3* In Example 6.2.1, the relative frequency of  $f(\text{red}) = f(A_1)$  tends to  $18/37$  as  $n$  tends to infinity because 18 out of 37 numbers are red.

The reader will gain a more theoretical understanding of how repeated experiments relate to expected quantities in the following chapters after learning the Theorem of Large Numbers described in Appendix A.3.

A different definition of probability was given by Pierre-Simon Laplace (1749–1827). We call an experiment a **Laplace experiment** if the number of possible simple events is finite and all the outcomes are equally probable. The probability of an arbitrary event  $A$  is then defined as follows:

**Definition 6.2.1** The proportion

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of "favourable simple events" for } A}{\text{Total number of possible simple events}} \quad (6.2)$$

is called the **Laplace probability**, where  $|A|$  is the cardinal number of  $A$ , i.e. the number of simple events contained in the set  $A$ , and  $|\Omega|$  is the cardinal number of  $\Omega$ , i.e. the number of simple events contained in the set  $\Omega$ .

The cardinal numbers  $|A|$  and  $|\Omega|$  are often calculated using the combinatoric rules introduced in Chap. 5.

*Example 6.2.4 (Example 6.1.2 continued)* The sample space contains 36 simple events. All of these simple events have equal probability  $1/36$ . To calculate the probability of the event  $A$  that the sum of the dots on the two dice is at least 4 and at most 6, we count the favourable simple events which fulfil this condition. The simple events are  $(1, 3), (2, 2), (3, 1)$  (sum is 4),  $(1, 4), (2, 3), (4, 1), (3, 2)$  (sum is 5) and  $(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)$  (sum is 6). In total, there are  $(3 + 4 + 5) = 12$  favourable simple events, i.e.

$$A = \{(1, 3), (2, 2), (3, 1), (1, 4), (2, 3), (4, 1), \\ (3, 2), (1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} .$$

The probability of the event  $A$  is therefore  $12/36 = 1/3$ .

### 6.3 The Axiomatic Definition of Probability

An important foundation for modern probability theory was established by A.N. Kolmogorov in 1933 when he proposed the following **axioms of probability**.

**Axiom 1** Every random event  $A$  has a probability in the (closed) interval  $[0, 1]$ , i.e.

$$0 \leq P(A) \leq 1.$$

**Axiom 2** The sure event has probability 1, i.e.

$$P(\Omega) = 1.$$

**Axiom 3** If  $A_1$  and  $A_2$  are disjoint events, then

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

holds.

*Remark* Axiom 3 also holds for three or more disjoint events and is called the **theorem of additivity of disjoint events**. For example, if  $A_1, A_2$ , and  $A_3$  are disjoint events, then  $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$ .

*Example 6.3.1* Suppose the two events in tossing a coin are  $A_1$ : “appearance of head” and  $A_2$ : “appearance of tail” which are disjoint. The event  $A_1 \cup A_2$ : “appearance of head or tail” has the probability

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) = 1/2 + 1/2 = 1.$$

*Example 6.3.2* Suppose an event is defined as the number of points observed on the upper surface of a die when rolling it. There are six events, i.e. the natural numbers 1, 2, 3, 4, 5, 6. These events are disjoint and they have equal probability of occurring:  $P(1) = P(2) = \dots = P(6) = 1/6$ . The probability of getting an even number is then

$$P(\text{“even number”}) = P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 1/2.$$

### 6.3.1 Corollaries Following from Kolmogorov's Axioms

We already know that  $A \cup \bar{A} = \Omega$  (sure event). Since  $A$  and  $\bar{A}$  are disjoint, using Axiom 3 we have

$$P(A \cup \bar{A}) = P(A) + P(\bar{A}) = 1.$$

Based on this, we have the following corollaries.

**Corollary 1** *The probability of the complementary event of  $A$ , (i.e.  $\bar{A}$ ) is*

$$P(\bar{A}) = 1 - P(A). \quad (6.3)$$

*Example 6.3.3* Suppose a box of 30 chocolates contains chocolates of 6 different flavours with 5 chocolates of each flavour. Suppose an event  $A$  is defined as  $A = \{\text{"marzipan flavour"}\}$ . The probability of finding a marzipan chocolate (without looking into the box) is  $P(\text{"marzipan"}) = 5/30$ . Then, the probability of the complementary event  $\bar{A}$ , i.e. the probability of not finding a marzipan chocolate is therefore

$$P(\text{"no marzipan flavour"}) = 1 - P(\text{"marzipan flavour"}) = 25/30.$$

**Corollary 2** *The probability of occurrence of an impossible event  $\emptyset$  is zero:*

$$P(\emptyset) = P(\bar{\Omega}) = 1 - P(\Omega) = 0.$$

**Corollary 3** *Let  $A_1$  and  $A_2$  be not necessarily disjoint events. The probability of occurrence of  $A_1$  or  $A_2$  is*

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2). \quad (6.4)$$

The rule in (6.4) is known as **the additive theorem of probability**. Again we use the word “or” in the statistical sense: either  $A_1$  is occurring,  $A_2$  is occurring, or both of them. This means we have to add the probabilities  $P(A_1)$  and  $P(A_2)$  but need to make sure that the simple events which are contained in both sets are not counted twice, thus we subtract  $P(A_1 \cap A_2)$ .

*Example 6.3.4* There are 10 actors acting in a play. Two actors, one of whom is male, are portraying evil characters. In total, there are 6 female actors. Let an event  $A$  describe whether the actor is male and another event  $B$  describe whether the character is evil. Suppose we want to know the probability of a randomly chosen actor being male or evil. We can then calculate

$$\begin{aligned} P(\text{actor is male or evil}) &= \\ &= P(\text{actor is male}) + P(\text{actor is evil}) - P(\text{actor is male and evil}) \\ &= \frac{4}{10} + \frac{2}{10} - \frac{1}{10} = \frac{1}{2}. \end{aligned}$$

**Corollary 4** If  $A \subseteq B$  then  $P(A) \leq P(B)$ .

*Proof* We use the representation  $B = A \cup (\bar{A} \cap B)$  where  $A$  and  $\bar{A} \cap B$  are the disjoint events. Then using Axiom 3 and Axiom 1, we get

$$P(B) = P(A) + P(\bar{A} \cap B) \geq P(A).$$

### 6.3.2 Calculation Rules for Probabilities

The introduced axioms and corollaries can be summarized as follows:

- (1)  $0 \leq P(A) \leq 1$
- (2)  $P(\Omega) = 1$
- (3)  $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ , if  $A_1$  and  $A_2$  are disjoint
- (4)  $P(\emptyset) = 0$
- (5)  $P(\bar{A}) = 1 - P(A)$
- (6)  $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
- (7)  $P(A) \leq P(B)$ , if  $A \subseteq B$

## 6.4 Conditional Probability

Consider the following example to understand the concept of conditional probability: Suppose a new medical test is developed to diagnose a particular infection of the blood. The test is conducted on blood samples from 100 randomly selected patients and the outcomes of the tests are presented in Table 6.1.

There are the following four possible outcomes:

- The blood sample has an infection and the test diagnoses it, i.e. the test is correctly diagnosing the infection.
- The blood sample does not have an infection and the test does not diagnose it, i.e. the test is correctly diagnosing that there is no infection.
- The blood sample has an infection and the test does not diagnose it, i.e. the test is incorrect in stating that there is no infection.
- The blood sample does not have an infection but the test diagnoses it, i.e. the test is incorrect in stating that there is an infection.

Table 6.2 contains the relative frequencies of Table 6.1. In the following, we interpret the relative frequencies as probabilities, i.e. we assume that the values in Table 6.2 would be observed if the number  $n$  of patients was much larger than 100.

It can be seen that the probability that a test is positive is  $P(T+) = 0.30 + 0.10 = 0.40$  and the probability that an infection is present is  $P(IP) = 0.30 + 0.15 = 0.45$ .

**Table 6.1** Absolute frequencies of test results and infection status

		Infection		Total (row)
		Present	Absent	
Test	Positive (+)	30	10	40
	Negative (-)	15	45	60
	Total (column)	45	55	Total = 100

**Table 6.2** Relative frequencies of patients and test

		Infection		Total (row)
		Present (IP)	Absent (IA)	
Test	Positive (+)	0.30	0.10	0.40
	Negative (-)	0.15	0.45	0.60
	Total (column)	0.45	0.55	Total = 1

If one already knows that the test is positive and wants to determine the probability that the infection is indeed present, then this can be achieved by the respective **conditional probability**  $P(IP|T+)$  which is

$$P(IP|T+) = \frac{P(IP \cap T+)}{P(T+)} = \frac{0.3}{0.4} = 0.75.$$

Note that  $IP \cap T+$  denotes the “relative frequency of blood samples in which the disease is present *and* the test is positive” which is 0.3.

More generally, recall Definition 4.1.1 from Chap. 4 where we defined conditional, joint, and marginal frequency distributions in contingency tables. The present example simply applies these rules to the contingency tables of relative frequencies and interprets the relative frequencies as an approximation to the probabilities of interest, as already explained.

We use the intersection operator  $\cap$  to describe events which occur for  $A = a$  and  $B = b$ . This relates to the joint relative frequencies. The marginal relative frequencies (i.e. probabilities  $P(A = a)$ ) can be observed from the column and row sums, respectively; and the conditional probabilities can be observed as the joint frequencies in relation to the marginal frequencies.

For simplicity, assume that all simple events in  $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$  are equally probable, i.e.  $P(\omega_j) = \frac{1}{k}$ ,  $j = 1, 2, \dots, k$ . Let  $A$  and  $B$  be two events containing  $n_A$  and  $n_B$  numbers of simple events. Let further  $A \cap B$  contain  $n_{AB}$  numbers of simple events. The Laplace probability using (6.2) is

$$P(A) = \frac{n_A}{k}, \quad P(B) = \frac{n_B}{k}, \quad P(A \cap B) = \frac{n_{AB}}{k}.$$

Assume that we have prior information that  $A$  has already occurred. Now we want to find out how the probability of  $B$  is to be calculated. Since  $A$  has already occurred, we know that the sample space is reduced by the number of simple events which

are contained in  $A$ . There are  $n_A$  such simple events. Thus, the total sample space  $\Omega$  is reduced by the sample space of  $A$ . Therefore, the simple events in  $A \cap B$  are those simple events which are realized when  $B$  is realized. The Laplace probability for  $B$  under the prior information on  $A$ , or under the condition that  $A$  is known, is therefore

$$P(B|A) = \frac{n_{AB}/k}{n_A/k} = \frac{P(A \cap B)}{P(A)}. \quad (6.5)$$

This can be generalized to the case when the probabilities for simple events are unequal.

**Definition 6.4.1** Let  $P(A) > 0$ . Then the **conditional probability** of event  $B$  occurring, given that event  $A$  has already occurred, is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \quad (6.6)$$

The roles of  $A$  and  $B$  can be interchanged to define  $P(A|B)$  as follows. Let  $P(B) > 0$ . The conditional probability of  $A$  given  $B$  is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (6.7)$$

We now introduce a few important theorems which are relevant to calculating conditional and other probabilities.

**Theorem 6.4.1** (Multiplication Theorem of Probability) *For two arbitrary events  $A$  and  $B$ , the following holds:*

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A). \quad (6.8)$$

This theorem follows directly from the two definitions (6.6) and (6.7) (but does not require that  $P(A) > 0$  and  $P(B) > 0$ ).

**Theorem 6.4.2** (Law of Total Probability) *Assume that  $A_1, A_2, \dots, A_m$  are events such that  $\cup_{i=1}^m A_i = \Omega$  and  $A_i \cap A_j = \emptyset$  for all  $i \neq j$ ,  $P(A_i) > 0$  for all  $i$ , i.e.  $A_1, A_2, \dots, A_m$  form a complete decomposition of  $\Omega = \cup_{i=1}^m A_i$  in pairwise disjoint events, then the probability of an event  $B$  can be calculated as*

$$P(B) = \sum_{i=1}^m P(B|A_i)P(A_i). \quad (6.9)$$

### 6.4.1 Bayes' Theorem

Bayes' Theorem gives a connection between  $P(A|B)$  and  $P(B|A)$ . For events  $A$  and  $B$  with  $P(A) > 0$  and  $P(B) > 0$ , using (6.6) and (6.7) or (6.8), we get

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A)} \frac{P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B)}. \end{aligned} \quad (6.10)$$

Let  $A_1, A_2, \dots, A_m$  be events such that  $\cup_{i=1}^m A_i = \Omega$  and  $A_i \cap A_j = \emptyset$  for all  $i \neq j$ ,  $P(A_i) > 0$  for all  $i$ , and  $B$  is another event than  $A$ , then using (6.9) and (6.10), we get

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)}. \quad (6.11)$$

The probabilities  $P(A_i)$  are called **prior probabilities**,  $P(B|A_i)$  are sometimes called **model probabilities** and  $P(A_j|B)$  are called **posterior probabilities**.

*Example 6.4.1* Suppose someone rents movies from two different DVD stores. Sometimes it happens that the DVD does not work because of scratches. We consider the following events:  $A_i$  ( $i = 1, 2$ ): “the DVD is rented from store  $i$ ”. Further let  $B$  denote the event that the DVD is working without any problems. Assume we know that  $P(A_1) = 0.6$  and  $P(A_2) = 0.4$  (note that  $A_2 = \bar{A}_1$ ) and  $P(B|A_1) = 0.95$ ,  $P(B|A_2) = 0.75$  and we are interested in the probability that a rented DVD works fine. We can then apply the Law of Total Probability and get

$$\begin{aligned} P(B) &\stackrel{(6.9)}{=} P(B|A_1)P(A_1) + P(B|A_2)P(A_2) \\ &= 0.6 \cdot 0.95 + 0.4 \cdot 0.75 = 0.87. \end{aligned}$$

We may also be interested in the probability that the movie was rented from store 1 and is working which is

$$P(B \cap A_1) \stackrel{(6.8)}{=} P(B|A_1)P(A_1) = 0.95 \cdot 0.6 = 0.57.$$

Now suppose we have a properly working DVD. What is the probability that it is rented from store 1? This is obtained as follows:

$$P(A_1|B) \stackrel{(6.7)}{=} \frac{P(A_1 \cap B)}{P(B)} = \frac{0.57}{0.87} = 0.6552.$$

Now assume we have a DVD which does not work, i.e.  $\bar{B}$  occurs. The probability that a DVD is not working given that it is from store 1 is  $P(\bar{B}|A_1) = 0.05$ . Similarly,

$P(\bar{B}|A_2) = 0.25$  for store 2. We can now calculate the conditional probability that a DVD is from store 1 given that it is not working:

$$\begin{aligned} P(A_1|\bar{B}) &\stackrel{(6.11)}{=} \frac{P(\bar{B}|A_1)P(A_1)}{P(\bar{B}|A_1)P(A_1) + P(\bar{B}|A_2)P(A_2)} \\ &= \frac{0.05 \cdot 0.6}{0.05 \cdot 0.6 + 0.25 \cdot 0.4} = 0.2308. \end{aligned}$$

The result about  $P(\bar{B})$  used in the denominator can also be directly obtained by using  $P(\bar{B}) = 1 - 0.87 = 0.13$ .

## 6.5 Independence

Intuitively, two events are independent if the occurrence or non-occurrence of one event does not affect the occurrence or non-occurrence of the other event. In other words, two events  $A$  and  $B$  are independent if the probability of occurrence of  $B$  has no effect on the probability of occurrence of  $A$ . In such a situation, one expects that

$$P(A|B) = P(A) \quad \text{and} \quad P(A|\bar{B}) = P(A).$$

Using this and (6.7), we can write

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A \cap \bar{B})}{P(\bar{B})} = P(A|\bar{B}). \end{aligned} \tag{6.12}$$

This yields:

$$\begin{aligned} P(A \cap B)P(\bar{B}) &= P(A \cap \bar{B})P(B) \\ P(A \cap B)(1 - P(B)) &= P(A \cap \bar{B})P(B) \\ P(A \cap B) &= (P(A \cap \bar{B}) + P(A \cap B))P(B) \\ P(A \cap B) &= P(A)P(B). \end{aligned} \tag{6.13}$$

This leads to the following definition of stochastic independence.

**Definition 6.5.1** Two random events  $A$  and  $B$  are called **(stochastically) independent** if

$$P(A \cap B) = P(A)P(B), \tag{6.14}$$

i.e. if the probability of simultaneous occurrence of both events  $A$  and  $B$  is the product of the individual probabilities of occurrence of  $A$  and  $B$ .

This definition of independence can be extended to the case of more than two events as follows:

**Definition 6.5.2** The  $n$  events  $A_1, A_2, \dots, A_n$  are stochastically mutually independent, if for any subset of  $m$  events  $A_{i_1}, A_{i_2}, \dots, A_{i_m}$  ( $m \leq n$ )

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \cdot \dots \cdot P(A_{i_m}) \quad (6.15)$$

holds.

A weaker form of independence is pairwise independence. If condition (6.15) is fulfilled only for two arbitrary events, i.e.  $m = 2$ , then the events are called **pairwise independent**. The difference between pairwise independence and general stochastic independence is explained in the following example.

*Example 6.5.1* Consider an urn with four balls. The following combinations of zeroes and ones are printed on the balls: 110, 101, 011, 000. One ball is drawn from the urn. Define the following events:

- $A_1$  : The first digit on the ball is 1.
- $A_2$  : The second digit on the ball is 1.
- $A_3$  : The third digit on the ball is 1.

Since there are two favourable simple events for each of the events  $A_1, A_2$  and  $A_3$ , we get

$$P(A_1) = P(A_2) = P(A_3) = \frac{2}{4} = \frac{1}{2}.$$

The probability that all the three events simultaneously occur is zero because there is no ball with 111 printed on it. Therefore,  $A_1, A_2$ , and  $A_3$  are not stochastically independent because

$$P(A_1)P(A_2)P(A_3) = \frac{1}{8} \neq 0 = P(A_1 \cap A_2 \cap A_3).$$

However,

$$\begin{aligned} P(A_1 \cap A_2) &= \frac{1}{4} = P(A_1)P(A_2), \\ P(A_1 \cap A_3) &= \frac{1}{4} = P(A_1)P(A_3), \\ P(A_2 \cap A_3) &= \frac{1}{4} = P(A_2)P(A_3), \end{aligned}$$

which means that the three events are pairwise independent.

## 6.6 Key Points and Further Issues

**Note:**

✓ We summarize some important theorems and laws:

- The Laplace probability is the ratio

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of "favourable simple events" for } A}{\text{Total number of possible simple events}}.$$

- The Law of Total Probability is

$$P(B) = \sum_{i=1}^m P(B|A_i)P(A_i).$$

- Bayes' Theorem is

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)}.$$

- $n$  events  $A_1, A_2, \dots, A_n$  are (stochastically) independent, if

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2) \cdot \dots \cdot P(A_n).$$

- ✓ In Sect. 10.8, we present the  $\chi^2$ -independence test, which can test whether discrete random variables (see Chap. 7) are independent or not.

---

## 6.7 Exercises

*Exercise 6.1*

- Suppose  $\Omega = \{0, 1, \dots, 15\}$ ,  $A = \{0, 8\}$ ,  $B = \{1, 2, 3, 5, 8, 10, 12\}$ ,  $C = \{0, 4, 9, 15\}$ . Determine  $A \cap B$ ,  $B \cap C$ ,  $A \cup C$ ,  $C \setminus A$ ,  $\Omega \setminus (B \cup A \cup C)$ .
- Now consider the three pairwise disjoint events  $E$ ,  $F$ ,  $G$  with  $\Omega = E \cup F \cup G$  and  $P(E) = 0.2$  and  $P(F) = 0.5$ . Calculate  $P(\bar{F})$ ,  $P(G)$ ,  $P(E \cap G)$ ,  $P(E \setminus E)$ , and  $P(E \cup F)$ .

*Exercise 6.2* A driving licence examination consists of two parts which are based on a theoretical and a practical examination. Suppose 25 % of people fail the practical examination, 15 % of people fail the theoretical examination, and 10 % of people fail both the examinations. If a person is randomly chosen, then what is the probability that this person

- (a) fails at least one of the examinations?
- (b) only fails the practical examination, but not the theoretical examination?
- (c) successfully passes both the tests?
- (d) fails any of the two examinations?

*Exercise 6.3* A new board game uses a twelve-sided die. Suppose the die is rolled once, what is the probability of getting

- (a) an even number?
- (b) a number greater than 9?
- (c) an even number greater than 9?
- (d) an even number or a number greater than 9?

*Exercise 6.4* The Smiths are a family of six. They are celebrating Christmas and there are 12 gifts, two for each family member. The name tags for each family member have been attached to the gifts. Unfortunately the name tags on the gifts are damaged by water. Suppose each family member draws two gifts at random. What is the probability that someone

- (a) gets his/her two gifts, rather than getting the gifts for another family member?
- (b) gets none of his/her gifts, but rather gets the gifts for other family members?

*Exercise 6.5* A chef from a popular TV cookery show sometimes puts too much salt in his pumpkin soup and the probability of this happening is 0.2. If he is in love (which he is with probability 0.3), then the probability of using too much salt is 0.6.

- (a) Create a contingency table for the probabilities of the two variables “in love” and “too much salt”.
- (b) Determine whether the two variables are stochastically independent or not.

*Exercise 6.6* Dr. Obermeier asks his neighbour to take care of his basil plant while he is away on leave. He assumes that his neighbour does not take care of the basil with a probability of  $\frac{1}{3}$ . The basil dies with probability  $\frac{1}{2}$  when someone takes care of it and with probability  $\frac{3}{4}$  if no one takes care of it.

- (a) Calculate the probability of the basil plant surviving after its owner’s leave.
- (b) It turns out that the basil eventually dies. What is the probability that Dr. Obermeier’s neighbour did not take care of the plant?

*Exercise 6.7* A bank considers changing its credit card policy. Currently 5 % of credit card owners are not able to pay their bills in any month, i.e. they never pay their bills. Among those who are generally able to pay their bills, there is still a 20 % probability that the bill is paid too late in a particular month.

- (a) What is the probability that someone is not paying his bill in a particular month?
- (b) A credit card owner did not pay his bill in a particular month. What is the probability that he never pays back the money?
- (c) Should the bank consider blocking the credit card if a customer does not pay his bill on time?

*Exercise 6.8* There are epidemics which affect animals such as cows, pigs, and others. Suppose 200 cows are tested to see whether they are infected with a virus or not. Let event  $A$  describe whether a cow has been transported by a truck recently or not and let  $B$  denote the event that a cow has been tested positive with a virus. The data are summarized in the following table:

	$B$	$\bar{B}$	Total
$A$	40	60	100
$\bar{A}$	20	80	100
Total	60	140	200

- (a) What is the probability that a cow is infected and has been transported by a truck recently?
- (b) What is the probability of having an infected cow given that it has been transported by the truck?
- (c) Determine and interpret  $P(B)$ .

*Exercise 6.9* A football practice target is a portable wall with two holes (which are the target) in it for training shots. Suppose there are two players  $A$  and  $B$ . The probabilities of hitting the target by  $A$  and  $B$  are 0.4 and 0.5, respectively.

- (a) What is the probability that at least one of the players succeeds with his shot?
- (b) What is the probability that exactly one of the players hits the target?
- (c) What is the probability that only  $B$  scores?

→ Solutions to all exercises in this chapter can be found on p. [361](#)

\*Source Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg

In the first part of the book we highlighted how to *describe* data. Now, we discuss the concepts required to draw statistical conclusions from a sample of data about a population of interest. For example, suppose we know the starting salary of a sample of 100 students graduating in law. We can use this knowledge to draw conclusions about the expected salary for the population of all students graduating in law. Similarly, if a newly developed drug is given to a sample of selected tuberculosis patients, then some patients may show improvement and some patients may not, but we are interested in the consequences for the entire population of patients. In the remainder of this chapter, we describe the theoretical concepts required for making such conclusions. They form the basis for statistical tests and inference which are introduced in Chaps. 9–11.

---

## 7.1 Random Variables

Random variables help us to view the collected data as an outcome of a random experiment. Consider the simple experiment of tossing a coin. If a coin is tossed, then one can observe either “head” ( $H$ ) or “tail” ( $T$ ). The occurrence of “head” or “tail” is random, and the exact outcome will only be known after the coin is tossed. We can toss the coin many times and obtain a sequence of outputs. For example, if a coin is tossed seven times, then one of the outcomes may be  $H, H, T, H, T, T, T$ . This outcome is the consequence of a random experiment, and it may be helpful if we can distill the sequence of outcomes in meaningful numbers. One option is to summarize them by a variable  $X$ , which takes the values  $x_1 = 1$  (denoting head) and  $x_2 = 0$  (denoting tail). We have learnt from Chap. 6 that this can be described in the framework of a random experiment where  $\Omega = \{\omega_1, \omega_2\}$  with the events  $A_1 = \{\omega_1\} = 1 = \text{head}$  and  $A_2 = \{\omega_2\} = 0 = \text{tail}$ . The random variable  $X$  is

**Table 7.1** Examples of random variables

$X$	Event	Realizations of $X$
Roll of a die	$A_i$ : number $i$ ( $i = 1, 2, \dots, 6$ )	$x = i$
Lifetime of TV	$A_i$ : survival time is $i$ months ( $i = 1, 2, \dots$ )	$x = i$
Roulette	$A_1$ : red	$x_1 = 1$
	$A_2$ : black	$x_2 = 2$
	$A_3$ : green (zero)	$x_3 = 0$

now mapped to real numbers, and therefore, it describes the possible outcome of *any* coin toss experiment. The observed outcomes  $H, H, T, H, T, T, T$  relate to a specific sample, a unique *realization* of this experiment. We can write  $X(\omega_1) = 1$  and  $X(\omega_2) = 0$  with  $\omega_1, \omega_2 \in \Omega$  and  $1, 0 \in \mathcal{R}$  where  $\mathcal{R}$  is the set of real numbers. We know that in any coin tossing experiment, the probability of head being observed is  $P(X(\omega_1) = 1) = 0.5$  and of tail being observed is  $P(X(\omega_2) = 0) = 0.5$ . We may therefore view  $X$  as a random variable which collects the possible outcomes of a random experiment and captures the uncertainty associated with them.

**Definition 7.1.1** Let  $\Omega$  represent the sample space of a random experiment, and let  $\mathcal{R}$  be the set of real numbers. A random variable is a function  $X$  which assigns to each element  $\omega \in \Omega$  one and only one number  $X(\omega) = x$ ,  $x \in \mathcal{R}$ , i.e.

$$X : \Omega \rightarrow \mathcal{R}. \quad (7.1)$$

*Example 7.1.1* The features of a die roll experiment, a roulette game, or the lifetime of a TV can all be described by a random variable, see Table 7.1. The events involve randomness, and if we have knowledge about the random process, we can assign probabilities  $P(X = x_i)$  to each event, e.g. when rolling a die, the probability of getting a “1” is  $P(X = 1) = 1/6$  and the probability of getting a “2” is  $P(X = 2) = 1/6$ .

Note that it is a convention to denote random variables by capital letters (e.g.  $X$ ) and their values by small letters (e.g.  $x$ ). It is evident from the coin tossing experiment that we need to know  $P(X = x)$  to describe the respective random variable. We assume in this chapter that we have this knowledge. However, Chaps. 9–11 show how a sample of data can be used to estimate unknown probabilities and other quantities given a prespecified uncertainty level. More generally, we can say that it is mandatory to know  $P(X \in A)$  for all possible  $A$  which are subsets of  $\mathcal{R}$ . If we choose  $A = (-\infty, x]$ ,  $x \in \mathcal{R}$ , we have

$$P(X \in A) = P(X \in (-\infty, x]) = P(-\infty < X \leq x) = P(X \leq x).$$

This consideration gives rise to the definition of the cumulative distribution function. Recall that we developed the concept of the empirical cumulative distribution function (ECDF) in Chap. 2, Sect. 2.2, but the definition there was empirical. Now, we develop it theoretically.

## 7.2 Cumulative Distribution Function (CDF)

**Definition 7.2.1** The **cumulative distribution function (CDF)** of a random variable  $X$  is defined as

$$F(x) = P(X \leq x). \quad (7.2)$$

As in Chap. 2, we can see that the CDF is useful in obtaining the probabilities related to the occurrence of random events. Note that the empirical cumulative distribution function (ECDF, Sect. 2.2) and the cumulative distribution function are closely related and therefore have a similar definition and similar calculation rules. However, in Chap. 2, we work with the cumulative distribution of *observed* values in a particular sample whereas in this chapter, we deal with random variables modelling the distribution of a general population.

The Definition 7.2.1 implies the following properties of the cumulative distribution function:

- $F(x)$  is a monotonically non-decreasing function  
(if  $x_1 \leq x_2$ , it follows that  $F(x_1) \leq F(x_2)$ ),
- $\lim_{x \rightarrow -\infty} F(x) = 0$  (the lower limit of  $F$  is 0),
- $\lim_{x \rightarrow +\infty} F(x) = 1$  (the upper limit of  $F$  is 1),
- $F(x)$  is continuous from the right, and
- $0 \leq F(x) \leq 1$  for all  $x \in \mathcal{R}$ .

Another notation for  $F(x) = P(X \leq x)$  is  $F_X(x)$ , but we use  $F(x)$ .

### 7.2.1 CDF of Continuous Random Variables

Before giving some examples about the meaning and interpretation of the CDF, we first need to consider some definitions and theorems.

**Definition 7.2.2** A random variable  $X$  is said to be **continuous** if there is a function  $f(x)$  such that for all  $x \in \mathcal{R}$

$$F(x) = \int_{-\infty}^x f(t)dt \quad (7.3)$$

holds.  $F(x)$  is the cumulative distribution function (CDF) of  $X$ , and  $f(x)$  is the probability density function (PDF) of  $x$  and  $\frac{d}{dx}F(x) = f(x)$  for all  $x$  that are continuity points of  $f$ .

**Theorem 7.2.1** For a function  $f(x)$  to be a **probability density function (PDF)** of  $X$ , it needs to satisfy the following conditions:

- (1)  $f(x) \geq 0$  for all  $x \in \mathcal{R}$ ,
- (2)  $\int_{-\infty}^{\infty} f(x)dx = 1$ .

**Theorem 7.2.2** Let  $X$  be a random variable with CDF  $F(x)$ . If  $x_1 < x_2$ , where  $x_1$  and  $x_2$  are known constants,  $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$ .

**Theorem 7.2.3** The probability of a continuous random variable taking a particular value  $x_0$  is zero:

$$P(X = x_0) = 0. \quad (7.4)$$

The proof is provided in Appendix C.2.

*Example 7.2.1* Consider the continuous random variable “waiting time for the train”. Suppose that a train arrives every 20 min. Therefore, the waiting time of a particular person is random and can be any time contained in the interval  $[0, 20]$ . We can start describing the required probability density function as

$$f(x) = \begin{cases} k & \text{for } 0 \leq x \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

where  $k$  is an unknown constant. Now, using condition (2) of Theorem 7.2.1, we have

$$1 = \int_0^{20} f(x)dx = [kx]_0^{20} = 20k$$

which needs to be fulfilled. This yields  $k = 1/20$  which is always greater than 0, and therefore, condition (1) of Theorem 7.2.1 is also fulfilled. It follows that

$$f(x) = \begin{cases} \frac{1}{20} & \text{for } 0 \leq x \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

is the probability density function describing the waiting time for the train. We can now use Definition 7.2.2 to determine the cumulative distribution function:

$$F(x) = \int_0^x f(t)dt = \int_0^x \frac{1}{20} dt = \frac{1}{20}[t]_0^x = \frac{1}{20}x.$$

Suppose we are interested in calculating the probability of a waiting time between 15 and 20 min. This can be calculated using Theorem 7.2.2:

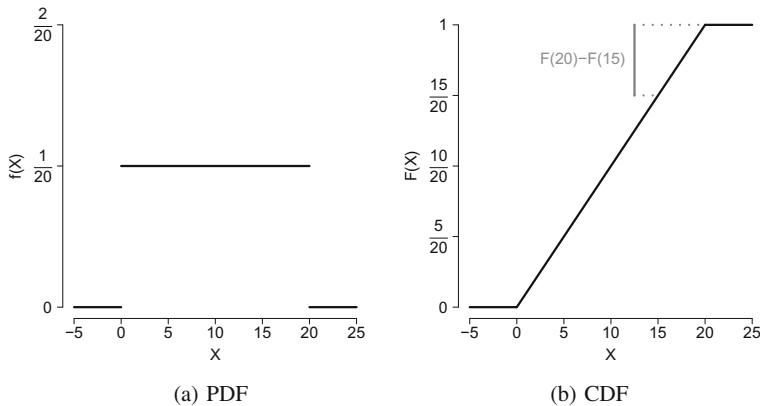
$$P(15 \leq X \leq 20) = F(20) - F(15) = \frac{20}{20} - \frac{15}{20} = 0.25.$$

We can obtain this probability from the graph of the CDF as well, see Fig. 7.1 where both the PDF and CDF of this example are illustrated.

Defining a function, for example the CDF, is simple in R: One can use the `function` command followed by specifying the variables the function evaluates in round brackets (e.g. `x`) and the function itself in braces (e.g. `x/20`). Functions can be plotted using the `curve` command:

```
cdf <- function(x){1/20 * x}
curve(cdf,from=0,to=20)
```





**Fig. 7.1** Probability density function (PDF) and cumulative distribution function (CDF) for waiting time in Example 7.2.1

Alternatively, the `plot` command can be used to plot vectors against each other; for example, after defining a function, we can define a sequence (`x<-seq(0,20,0.01)`), evaluate this sequence via the specified function (`cdf(x)`), and plot them against each other and connect the points from the sequence with a line (`plot(x,cdf(x),type='l')`).

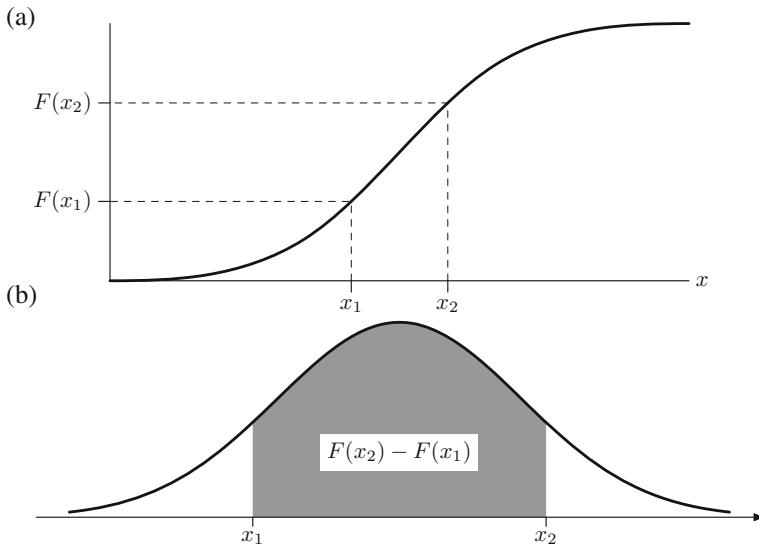
This example illustrates how the cumulative distribution function can be used to obtain probabilities of interest. Most importantly, if we want to calculate the probability that the random variable  $X$  takes values in the interval  $[x_1, x_2]$ , we simply have to look at the difference of the respective CDF values at  $x_1$  and  $x_2$ . Figure 7.2a highlights that the interval probability corresponds to the difference of the CDF values on the y-axis.

We can also use the probability density function to visualize  $P(x_1 \leq X \leq x_2)$ . We know from Theorem 7.2.1 that  $\int_{-\infty}^{\infty} f(x)dx = 1$ , and therefore, the area under the PDF equals 1. Thus, we can interpret interval probabilities as the area under the PDF between  $x_1$  and  $x_2$ . This is presented in Fig. 7.2b.

## 7.2.2 CDF of Discrete Random Variables

**Definition 7.2.3** A random variable  $X$  is defined to be **discrete** if its probability space is either finite or countable, i.e. if it takes only a finite or countable number of values. Note that a set  $V$  is said to be **countable**, if its elements can be listed, i.e. there is a one-to-one correspondence between  $V$  and the positive integers.

*Example 7.2.2* Consider the example of tossing of a coin where each trial results in either a head ( $H$ ) or a tail ( $T$ ), each occurring with the same probability



**Fig. 7.2** Graphical representation of the probability  $P(x_1 \leq X \leq x_2)$  **a** via the CDF and **b** via the PDF\*

0.5. When the coin is tossed multiple times, we may observe sequences such as  $H, T, H, H, T, H, H, T$ , and  $T, \dots$ . The sample space is  $\Omega = \{H, T\}$ . Let the random variable  $X$  denote the number of trials required to get the third head, then  $X = 4$  for the given sequence above. Clearly, the space of  $X$  is the set  $(3, 4, 5, \dots)$ . We can see that  $X$  is a discrete random variable because its space is finite and can be counted. We can also assign certain probabilities to each of these values, e.g.  $P(X = 3) = p_1$  and  $P(X = 4) = p_2$ .

**Definition 7.2.4** Let  $X$  be a discrete random variable which takes  $k$  different values. The **probability mass function** (PMF) of  $X$  is given by

$$f(X) = P(X = x_i) = p_i \quad \text{for each } i = 1, 2, \dots, k. \quad (7.5)$$

It is required that the probabilities  $p_i$  satisfy the following conditions:

- (1)  $0 \leq p_i \leq 1$ ,
- (2)  $\sum_{i=1}^k p_i = 1$ .

**Definition 7.2.5** Given (7.5), we can write the CDF of a discrete random variable as

$$F(x) = \sum_{i=1}^k I_{\{x_i \leq x\}} p_i, \quad (7.6)$$

where  $I$  is an indicator function defined as

$$I_{\{x_i \leq x\}} = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{otherwise.} \end{cases}$$

The CDF of a discrete variable is always a **step function**.

### Working with the CDF for Discrete Random variables

We can easily calculate various types of probabilities for discrete random variables using the CDF. Let  $a$  and  $b$  be some known constants, then

$$P(X \leq a) = F(a), \quad (7.7)$$

$$P(X < a) = P(X \leq a) - P(X = a) = F(a) - P(X = a), \quad (7.8)$$

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a), \quad (7.9)$$

$$P(X \geq a) = 1 - P(X < a) = 1 - F(a) + P(X = a), \quad (7.10)$$

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X < a) \\ &= F(b) - F(a) + P(X = a), \end{aligned} \quad (7.11)$$

$$P(a < X \leq b) = F(b) - F(a), \quad (7.12)$$

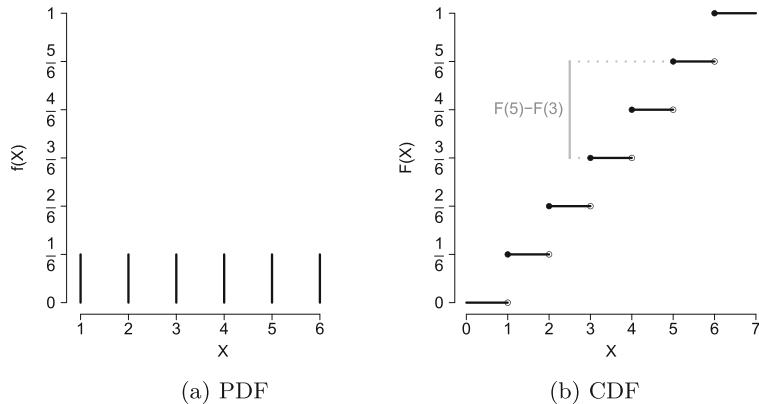
$$P(a < X < b) = F(b) - F(a) - P(X = b), \quad (7.13)$$

$$P(a \leq X < b) = F(b) - F(a) - P(X = b) + P(X = a). \quad (7.14)$$

*Remark 7.2.1* The Eqs. (7.7)–(7.14) can also be used for continuous variables, but in this case,  $P(X = a) = P(X = b) = 0$  (see Theorem 7.2.3), and therefore, Eqs. (7.7)–(7.14) can be modified accordingly.

*Example 7.2.3* Consider the experiment of rolling a die. There are six possible outcomes. If we define the random variable  $X$  as the number of dots observed on the upper surface of the die, then the six possible outcomes can be described as  $x_1 = 1, x_2 = 2, \dots, x_6 = 6$ . The respective probabilities are  $P(X = x_i) = 1/6; i = 1, 2, \dots, 6$ . The PMF and CDF are therefore defined as follows:

$$f(x) = \begin{cases} 1/6 & \text{if } x = 1 \\ 1/6 & \text{if } x = 2 \\ 1/6 & \text{if } x = 3 \\ 1/6 & \text{if } x = 4 \\ 1/6 & \text{if } x = 5 \\ 1/6 & \text{if } x = 6 \\ 0 & \text{elsewhere.} \end{cases} \quad F(x) = \begin{cases} 0 & \text{if } -\infty < x < 1 \\ 1/6 & \text{if } 1 \leq x < 2 \\ 2/6 & \text{if } 2 \leq x < 3 \\ 3/6 & \text{if } 3 \leq x < 4 \\ 4/6 & \text{if } 4 \leq x < 5 \\ 5/6 & \text{if } 5 \leq x < 6 \\ 1 & \text{if } 6 \leq x < \infty. \end{cases}$$



**Fig. 7.3** Probability density function and cumulative distribution function for rolling a die in Example 7.2.3. “●” relates to an included value and “○” to an excluded value

Both the CDF and the PDF are displayed in Fig. 7.3.

We can use the CDF to calculate any desired probability, e.g.  $P(X \leq 5) = F(5) = 5/6$ . This is shown in Fig. 7.3b where for  $X = 5$ , we obtain  $F(5) = 5/6$  when evaluating on the y-axis. Similarly,  $P(3 < X \leq 5) = F(5) - F(3) = (5/6) - (3/6) = 2/6$  can be interpreted as the difference of  $F(5)$  and  $F(3)$  on the y-axis.

### 7.3 Expectation and Variance of a Random Variable

We have seen that both the probability density function (or probability mass function) and the cumulative distribution function are helpful in characterizing the features of a random variable. Some other features of random variables are characterized by the concepts of *expectation* and *variance*.

#### 7.3.1 Expectation

**Definition 7.3.1** The expectation of a continuous random variable  $X$ , having the probability density function  $f(x)$  with  $\int |x| f(x) dx < \infty$ , is defined as

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx. \quad (7.15)$$

For a discrete random variable  $X$ , which takes the values  $x_1, x_2, \dots$  with respective probabilities  $p_1, p_2, \dots$ , the **expectation** of  $X$  is defined as

$$\mathbb{E}(X) = \sum_{i=1}^k x_i p_i = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_k P(X = x_k). \quad (7.16)$$

The *expectation* of  $X$ , i.e.  $\mathbb{E}(X)$ , is usually denoted by  $\mu = \mathbb{E}(X)$  and relates to the arithmetic mean of the distribution of the population. It reflects the central tendency of the population.

*Example 7.3.1* Consider again Example 7.2.1 where the waiting time for a train was described by the following probability density function:

$$f(x) = \begin{cases} \frac{1}{20} & \text{for } 0 \leq x \leq 20 \\ 0 & \text{otherwise.} \end{cases}$$

We can calculate the expectation as follows:

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^0 x f(x) dx + \int_0^{20} x f(x) dx + \int_{20}^{\infty} x f(x) dx \\ &= 0 + \int_0^{20} \frac{1}{20} x dx + 0 = \left[ \frac{1}{40} x^2 \right]_0^{20} = \frac{400}{40} - 0 = 10. \end{aligned}$$

The “average” waiting time for the train is therefore 10 min. This means that if a person has to wait for the train every day, then the person will experience waiting times varying randomly between 0 and 20 min and, on average, has to wait for 10 min.

*Example 7.3.2* Consider again the die roll experiment from Example 7.2.3. The probabilities for the occurrence of any  $x_i$ ,  $i = 1, 2, \dots, 6$ , are  $P(X = x_i) = 1/6$ . The expectation can thus be calculated as

$$\begin{aligned} \mathbb{E}(X) &= \sum_{i=1}^6 x_i p_i \\ &= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3) + 4 \cdot P(X = 4) \\ &\quad + 5 \cdot P(X = 5) + 6 \cdot P(X = 6) \\ &= (1 + 2 + 3 + 4 + 5 + 6) \frac{1}{6} = \frac{21}{6} = 3.5. \end{aligned}$$

### 7.3.2 Variance

The *variance* describes the variability of a random variable. It gives an idea about the concentration or dispersion of values around the arithmetic mean of the distribution.

**Definition 7.3.2** The **variance** of a random variable  $X$  is defined as

$$\text{Var}(X) = E[X - E(X)]^2. \quad (7.17)$$

The variance of a continuous random variable  $X$  is

$$\text{Var}(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx \quad (7.18)$$

where  $E(X) = \int_{-\infty}^{+\infty} xf(x) dx$ . Similarly, the variance of a discrete random variable  $X$  is

$$\text{Var}(X) = \sum_{i=1}^{\infty} (x_i - E(X))^2 p_i \quad (7.19)$$

where  $E(X) = \sum_i x_i p_i$ . The variance is usually denoted by  $\sigma^2 = \text{Var}(X)$ .

**Definition 7.3.3** The positive square root of the variance is called the **standard deviation**.

*Example 7.3.3* Recall Examples 7.2.1 and 7.3.1. We can calculate the variance of the waiting time for a train using the probability density function

$$f(x) = \begin{cases} \frac{1}{20} & \text{for } 0 \leq x \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

and  $E(X) = 10$  (already calculated in Example 7.3.1). Using (7.18), we obtain:

$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{\infty} (x - E(x))^2 f(x) dx = \int_{-\infty}^{\infty} (x - 10)^2 f(x) dx \\ &= \int_{-\infty}^0 (x - 10)^2 f(x) dx + \int_0^{20} (x - 10)^2 f(x) dx + \int_{20}^{\infty} (x - 10)^2 f(x) dx \\ &= 0 + \int_0^{20} (x - 10)^2 \cdot \frac{1}{20} dx + 0 = \int_0^{20} \frac{1}{20} (x^2 - 20x + 100) dx \\ &= \left[ \frac{1}{20} \left( \frac{1}{3}x^3 - 10x^2 + 100x \right) \right]_0^{20} = 33\frac{1}{3}. \end{aligned}$$

The standard deviation is  $\sqrt{33\frac{1}{3} \text{ min}^2} \approx 5.77 \text{ min}$ .

Recall that in Chap. 3, we introduced the sample variance and the sample standard deviation. We already know that the standard deviation has the same unit of measurement as the variable, whereas the unit of the variance is the square of the measurement unit. The standard deviation measures how the values of a random variable are dispersed around the population mean. A low value of the standard deviation indicates that the values are highly concentrated around the mean. A high value of the standard deviation indicates lower concentration of the data values around the mean, and the observed values may be far away from the mean. These considerations are helpful in making connections between random variables and samples of data, see Chap. 9 for the construction of confidence intervals.

*Example 7.3.4* Recall Example 7.3.2 where we calculated the expectation of a die roll experiment as  $E(X) = 3.5$ . With  $x_i \in \{1, 2, 3, 4, 5, 6\}$  and  $p_i = 1/6$  for all  $i = 1, 2, 3, 4, 5, 6$ , the variance for this example corresponds to

$$\begin{aligned}\text{Var}(X) &= \sum_{i=1}^6 (x_i - E(X))^2 p_i = (1 - 3.5)^2 \cdot \frac{1}{6} + (2 - 3.5)^2 \cdot \frac{1}{6} + (3 - 3.5)^2 \cdot \frac{1}{6} \\ &\quad + (4 - 3.5)^2 \cdot \frac{1}{6} + (5 - 3.5)^2 \cdot \frac{1}{6} + (6 - 3.5)^2 \cdot \frac{1}{6} \approx 2.92.\end{aligned}$$

**Theorem 7.3.1** *The variance of a random variable  $X$  can be expressed as*

$$\text{Var}(X) = E(X^2) - [E(X)]^2. \quad (7.20)$$

*The proof is given in Appendix C.2.*

*Example 7.3.5* In Examples 7.2.1, 7.3.1, and 7.3.3, we evaluated the waiting time for a train using the PDF

$$f(X) = \begin{cases} \frac{1}{20} & \text{for } 0 < X \leq 20 \\ 0 & \text{otherwise.} \end{cases}$$

We calculated the expectation and variance in Eqs. (7.15) and (7.17) as 10 min and  $33\frac{1}{3} \text{ min}^2$ , respectively. Theorem 7.3.1 tells us that we can calculate the variance in a different way as follows:

$$\begin{aligned}E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^{20} \frac{1}{20} x^2 dx \\ &= \left[ \frac{1}{60} x^3 \right]_0^{20} = 133\frac{1}{3} \\ \text{Var}(X) &= E(X^2) - [E(X)]^2 = 133\frac{1}{3} - 10^2 = 33\frac{1}{3}.\end{aligned}$$

This yields the same result as Eq. (7.18) but is much quicker.

### 7.3.3 Quantiles of a Distribution

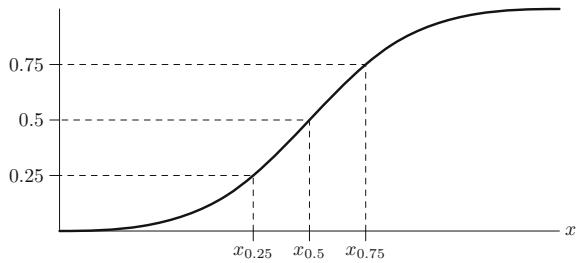
We introduced the concept of quantiles in Chap. 3, Sect. 3.1.2. Now, we define quantiles in terms of the distribution function.

**Definition 7.3.4** The value  $x_p$  for which the cumulative distribution function is

$$F(x_p) = p \quad (0 < p < 1) \quad (7.21)$$

is called the **p-quantile**.

**Fig. 7.4** First quartile, median, and third quartile\*



It follows from Definition 7.3.4 that  $x_p$  is the value which divides the cumulative distribution function into two parts: the probability of observing a value left of  $x_p$  is  $p$ , whereas the probability of observing a value right of  $x_p$  is  $1 - p$ . For example, the 0.25-quantile  $x_{0.25}$  describes the  $x$ -value for which the probability of observing  $x_{0.25}$  or any smaller value is 0.25. Figure 7.4 shows the 0.25-quantile (first quartile), the 0.5-quantile (median), and the 0.75-quantile (third quartile) in a cumulative distribution function.

*Example 7.3.6* Recall Examples 7.2.1, 7.3.1, 7.3.5 and Fig. 7.1b where we described the waiting time for a train by using the following CDF:

$$F(x) = \frac{1}{20}x.$$

The first quartile  $x_{0.25}$  is 5 because  $F(5) = 5/20 = 0.25$ . This means that the probability of waiting for the train for 5 min or less is 25 % and of waiting for longer than 5 min is 75 %.

For continuous variables, there is a unique value which describes the  $p$ -quantile. However, for discrete variables, this may not necessarily be true. In this case, the  $p$ -quantile is chosen such that

$$\begin{aligned} F(x_p) &\geq p, \\ F(x) &< p \quad \text{for } x < x_p \end{aligned}$$

holds.

*Example 7.3.7* The cumulative distribution function for rolling a die is described in Example 7.2.3 and Fig. 7.3b. The first quartile  $x_{0.25}$  is 2 because  $F(2) = 2/6 > 0.25$  and  $F(x) < 0.25$  for  $x < 2$ .

### 7.3.4 Standardization

Standardization transforms a random variable in such a way that it has an expectation of zero and a variance of one. More details on the need for standardization are discussed in Chap. 10.

**Definition 7.3.5** A random variable  $Y$  is called **standardized** when

$$\mathbb{E}(Y) = 0 \quad \text{and} \quad \text{Var}(Y) = 1.$$

**Theorem 7.3.2** Suppose a random variable  $X$  has mean  $\mathbb{E}(X) = \mu$  and  $\text{Var}(X) = \sigma^2$ . Then, it can be standardized as follows:

$$Y = \frac{X - \mu}{\sigma} = \frac{X - \mathbb{E}(X)}{\sqrt{\text{Var}(X)}}. \quad (7.22)$$

*Example 7.3.8* In Examples 7.2.1, 7.3.1, and 7.3.5, we considered the waiting time  $X$  for a train. The random variable  $X$  can take values between 0 and 20 min, and we calculated  $\mathbb{E}(X) = 10$  and  $\text{Var}(X) = 33\frac{1}{3}$ . The standardized variable of  $X$  is

$$Y = \frac{X - \mu}{\sigma} = \frac{X - 10}{\sqrt{33\frac{1}{3}}}.$$

One can show that  $\mathbb{E}(Y) = 0$  and  $\text{Var}(Y) = 1$ , see also Exercise 7.10 for more details.

## 7.4 Tschebyschev's Inequality

If we do not know the distribution of a random variable  $X$ , we can still make statements about the probability that  $X$  takes values in a certain interval (which has to be symmetric around the expectation  $\mu$ ) if the mean  $\mu$  and the variance  $\sigma^2$  of  $X$  are known.

**Theorem 7.4.1** (Tschebyschev's inequality) Let  $X$  be a random variable with  $\mathbb{E}(X) = \mu$  and  $\text{Var}(X) = \sigma^2$ . It holds that

$$P(|X - \mu| \geq c) \leq \frac{\text{Var}(X)}{c^2}. \quad (7.23)$$

This is equivalent to

$$P(|X - \mu| < c) \geq 1 - \frac{\text{Var}(X)}{c^2}. \quad (7.24)$$

The proof is given in Appendix C.2.

*Example 7.4.1* In Examples 7.2.1, 7.3.1, and 7.3.5, we have worked with a random variable which describes the waiting time for a train. We determined  $\mathbb{E}(X) = 10$  and  $\text{Var}(X) = 33\frac{1}{3}$ . We can calculate the probability of waiting between  $10 - 7 = 3$  and  $10 + 7 = 17$  min:

$$\begin{aligned} P(|X - \mu| < c) &\geq 1 - \frac{\text{Var}(X)}{c^2} \\ P(|X - 10| < 7) &\geq 1 - \frac{33\frac{1}{3}}{7^2} \approx 0.32. \end{aligned}$$

The probability is therefore at least 0.32. However, if we apply our distributional knowledge that  $F(x) = \frac{1}{20}x$  (for  $0 \leq X \leq 20$ ), then we obtain a much more precise result which is

$$P(3 < X < 17) = F(17) - F(3) = \frac{17}{20} - \frac{3}{20} = 0.7.$$

We can clearly see that Tschebyschev's inequality gives us the correct answer, that is  $P(3 < X < 17)$  is greater than 0.32. Nevertheless, the approximation to the exact probability, 0.7, is rather poor. One needs to keep in mind that only the lack of distributional knowledge makes the inequality useful.

## 7.5 Bivariate Random Variables

There are many situations in which we are interested in analysing more than one variable, say two variables. When we have more than one variable, then not only their individual distributions but also their joint distribution can be of interest. For example, we know that driving a car after drinking alcohol is not necessarily safe. If we consider two variables, the blood alcohol content  $X$  and number of car accidents  $Y$ , then we may be interested in the probability of having a high blood alcohol content *and* a car accident at the same time. If we analyse  $(X, Y)$  jointly, then we are interested in their joint **bivariate** distribution  $f_{XY}(x, y)$ . This distribution can either be discrete or continuous.

**Discrete Bivariate Random Variables.** Suppose we have two categorical variables  $X$  and  $Y$  which can take the values  $x_1, x_2, \dots, x_I$  and  $y_1, y_2, \dots, y_J$ , respectively. Their **joint probability distribution function** is characterized by

$$P(X = x_i, Y = y_j) = p_{ij} \quad (i = 1, 2, \dots, I; j = 1, 2, \dots, J)$$

with  $\sum_{i=1}^I \sum_{j=1}^J p_{ij} = 1$ . This means that the probability of observing  $x_i$  *and*  $y_j$  together is  $p_{ij}$ . We can summarize this information in a contingency table as follows:

		$Y$				
		1	2	...	J	
$X$ :	1	$p_{11}$	$p_{12}$	...	$p_{1J}$	$p_{1+}$
	2	$p_{21}$	$p_{22}$	...	$p_{2J}$	$p_{2+}$
	:	:				:
	I	$p_{I1}$	$p_{I2}$	...	$p_{IJ}$	$p_{I+}$
Total		$p_{+1}$	$p_{+2}$	...	$p_{+J}$	1

Each cell contains a “piece” of the joint distribution. The entries  $p_{+1}, p_{+2}, \dots, p_{+J}$  in the bottom row of the table summarize the **marginal distribution** of  $Y$ , which is the distribution of  $Y$  without giving reference to  $X$ . The entries  $p_{1+}, p_{2+}, \dots, p_{I+}$

in the last column summarize the marginal distribution of  $X$ . The marginal distributions can therefore be expressed as

$$P(X = x_i) = \sum_{j=1}^J p_{ij} = p_{i+} \quad i = 1, 2, \dots, I,$$

$$P(Y = y_j) = \sum_{i=1}^I p_{ij} = p_{+j} \quad j = 1, 2, \dots, J.$$

The **conditional distributions** of  $X$  given  $Y = y_j$  and  $Y$  given  $X = x_i$  are given as follows:

$$P(X = x_i | Y = y_j) = p_{i|j} = \frac{p_{ij}}{p_{+j}} \quad i = 1, 2, \dots, I,$$

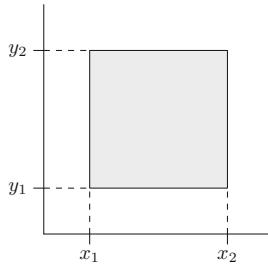
$$P(Y = y_j | X = x_i) = p_{j|i} = \frac{p_{ij}}{p_{i+}} \quad j = 1, 2, \dots, J.$$

They summarize the distribution of  $X$  for a given value of  $y_j$  (or the distribution of  $Y$  for a given value of  $x_i$ ) and play a crucial role in the construction of regression models such as the linear regression model introduced in Chap. 11. Please also recall the definitions of Sect. 4.1 where we introduced conditional and marginal distributions for data samples rather than random variables.

*Example 7.5.1* Suppose we have a contingency table on smoking behaviour  $X$  (1 = never smoking, 2 = smoking sometimes, and 3 = smoking regularly) and education level  $Y$  (1 = primary education, 2 = Secondary education, and 3 = tertiary education):

		$Y$			Total
		1	2	3	
$X$	1	0.10	0.20	0.30	0.60
	2	0.10	0.10	0.10	0.30
	3	0.08	0.01	0.01	0.10
Total		0.28	0.31	0.41	1

The cell entries represent the joint distribution of smoking behaviour and education level. We can interpret each entry as the probability of observing  $X = x_i$  and  $Y = y_j$  simultaneously. For example,  $p_{23} = P(\text{"smoking sometimes and tertiary education"}) = 0.10$ . The marginal distribution of  $X$  is contained in the last column of the table and lists the probabilities of smoking (unconditional on education level), e.g. the probability of being a non-smoker in this population is 60 %. We can also interpret the conditional distributions:  $P(X|Y = 3)$  represents the distribution of smoking behaviour among those who have tertiary education. If we are interested in the probability of smoking sometimes given tertiary education is completed, then we calculate  $P(X = 2|Y = 3) = p_{2|3} = \frac{0.10}{0.41} = 0.24$ .



**Fig. 7.5** Area covering all points of  $(X, Y)$  with  $(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2)^*$

### Continuous Bivariate Random Variables.

**Definition 7.5.1** A bivariate random variable  $(X, Y)$  is continuous if there is a function  $f_{XY}(x, y)$  such that

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^y \int_{-\infty}^x f_{XY}(x, y) dx dy \quad (7.25)$$

holds.

The function  $F_{XY}(x, y)$  is the **joint cumulative distribution function** of  $X$  and  $Y$ ; the joint distribution function is denoted by  $f_{XY}(x, y)$ , and  $f_{XY}(x, y)$  has to fulfil the usual conditions of a density function. Necessary and sufficient conditions that a function  $F_{XY}(x, y)$  is a bivariate cumulative distribution function are as follows:

$$\begin{aligned} \lim_{x \rightarrow -\infty} F_{XY}(x, y) &= 0 & \lim_{y \rightarrow -\infty} F_{XY}(x, y) &= 0 \\ \lim_{x \rightarrow \infty} F_{XY}(x, y) &= 1 & \lim_{y \rightarrow \infty} F_{XY}(x, y) &= 1 \end{aligned}$$

and  $F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \geq 0$  for all  $x_1 < x_2, y_1 < y_2$ .

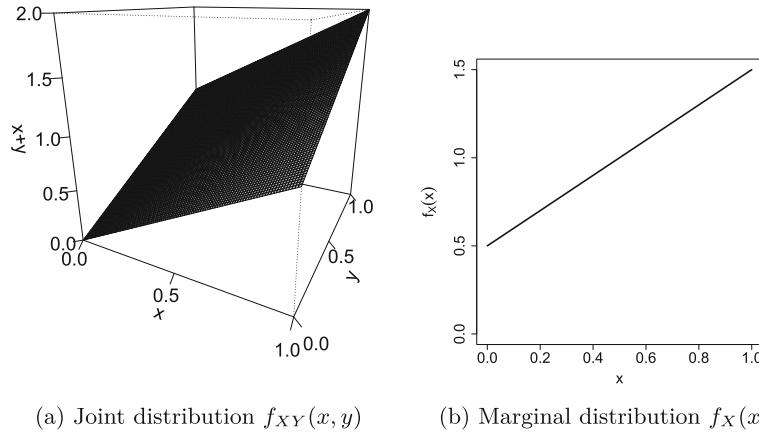
The last condition is sometimes referred to as the *rectangle inequality*. As in the univariate case, we can use the cumulative distribution function to calculate interval probabilities; similarly, we look at the rectangular area defined by  $(x_1, y_1), (x_1, y_2), (x_2, y_1)$ , and  $(x_2, y_2)$  in the bivariate case (instead of an interval  $[a, b]$ ), see Fig. 7.5.

We can calculate the desired probabilities as follows:

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x, y) dx dy.$$

The **marginal distributions** of  $X$  and  $Y$  are

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx,$$



**Fig. 7.6** Joint and marginal distribution for Example 7.5.2

respectively. Similar to the discrete case,  $f_X(x)$  and  $f_Y(y)$  also describe the distribution of  $X$  unconditional on  $Y$  and the distribution of  $Y$  unconditional on  $X$ . The **cumulative marginal distributions** are

$$F_X(x) = \int_{-\infty}^x f_X(t)dt, \quad F_Y(y) = \int_{-\infty}^y f_Y(t)dt.$$

The **conditional distributions** can be obtained by the ratio of the joint and marginal distributions:

$$f_{X|Y}(x, y) = \frac{f(x, y)}{f(y)}, \quad f_{Y|X}(x, y) = \frac{f(x, y)}{f(x)}.$$

*Example 7.5.2* Consider the function

$$f_{XY}(x, y) = \begin{cases} x + y & \text{for } 0 \leq x \leq 1, \quad 0 \leq y \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Suppose  $X$  and  $Y$  represent the concentrations of two drugs in the human body. Then,  $f_{XY}(x, y)$  may represent the sum of two drug concentrations in the human body. Since there are infinite possible realizations of both  $X$  and  $Y$ , we represent their joint distribution in a figure rather than a table, see Fig. 7.6a.

The marginal distributions for  $X$  and  $Y$  can be calculated as follows:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)dy = \int_0^1 (x + y) dy = \left[ xy + \frac{1}{2}y^2 \right]_0^1 = x + \frac{1}{2}, \\ f_Y(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)dx = \int_0^1 (x + y) dx = \left[ \frac{1}{2}x^2 + xy \right]_0^1 = y + \frac{1}{2}. \end{aligned}$$

Figure 7.6b depicts the marginal distribution for  $X$ . The slope of the marginal distribution is essentially the slope of the surface of the joint distribution shown in Fig. 7.6a. It is easy to see in this simple example that the marginal distribution of

$X$  is nothing but a cut in the surface of the joint distribution. Note that the conditional distributions  $f_{X|Y}(x, y)$  and  $f_{Y|X}(x, y)$  can be easily calculated; for example,  $f_{X|Y}(x, y) = f(x, y)/f(y) = (x + y)/(y + 0.5)$ .

### Stochastic Independence.

**Definition 7.5.2** Two continuous random variables  $X$  and  $Y$  are said to be **stochastically independent** if

$$f_{XY}(x, y) = f_X(x)f_Y(y). \quad (7.26)$$

For discrete variables, this is equivalent to

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \quad (7.27)$$

being valid for all  $(i, j)$ .

*Example 7.5.3* In Example 7.5.2, we considered the function

$$f_{XY}(x, y) = \begin{cases} x + y & \text{for } 0 \leq x \leq 1, \quad 0 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

with the marginal distributions of  $X$  and  $Y$  as  $f_X = x + 0.5$  and  $f_Y = y + 0.5$ , respectively. Since  $f_X \cdot f_Y = (x + \frac{1}{2})(y + \frac{1}{2}) \neq f_{XY}$ , it follows that  $X$  and  $Y$  are not independent. The interpretation is that the concentrations of the two drugs are not independent.

## 7.6 Calculation Rules for Expectation and Variance

**Calculation Rules for the Expectation.** For any constant values  $a$  and  $b$ , and any random variables  $X$  and  $Y$ , the following rules hold:

$$E(a) = a, \quad (7.28)$$

$$E(bX) = bE(X), \quad (7.29)$$

$$E(a + bX) = a + bE(X), \quad (7.30)$$

$$E(X + Y) = E(X) + E(Y) \text{ (additivity).} \quad (7.31)$$

The proof of rule (7.30) is given in Appendix C.2.

*Example 7.6.1* Consider again Example 7.2.3 where we illustrated how the outcome of a die roll experiment can be captured by a random variable. There were 6 events, and  $X$  could take the values  $x_1 = 1, x_2 = 2, \dots, x_6 = 6$ . The probability of the occurrence of any number was  $P(X = x_i) = 1/6$ , and the expectation was calculated as 3.5. Consider two different situations:

- (i) Suppose the die takes the value 10, 20, 30, 40, 50, and 60 instead of the values 1, 2, 3, 4, 5, and 6. The random variable  $Y = 10X$  describes this suitably, and its expectation is

$$E(Y) = E(10X) = 10E(X) = 10 \cdot 3.5 = 35$$

which follows from (7.29).

- (ii) If we are rolling two dices  $X_1$  and  $X_2$ , then the expectation for the sum of the two outcomes is

$$E(X) = E(X_1 + X_2) = E(X_1) + E(X_2) = 3.5 + 3.5 = 7$$

due to (7.31).

**Calculation Rules for the Variance.** Let  $a$  and  $b$  be any known constants and  $X$  be a random variable (discrete or continuous). Then, we have the following rules:

$$\text{Var}(a) = 0, \quad (7.32)$$

$$\text{Var}(bX) = b^2 \text{Var}(X), \quad (7.33)$$

$$\text{Var}(a + bX) = b^2 \text{Var}(X). \quad (7.34)$$

The proof of rule (7.34) is given in Appendix C.2.

*Example 7.6.2* In Examples 7.2.1, 7.3.1, 7.3.3, and 7.3.5, we evaluated a random variable describing the waiting time for a train. Now, suppose that a person first has to catch a bus to get to the train station. If this bus arrives only every 60 min, then the PDF of the random variable  $Y$  denoting the waiting time for the bus is

$$f(Y) = \begin{cases} \frac{1}{60} & \text{for } 0 < y \leq 60 \\ 0 & \text{otherwise.} \end{cases}$$

We can use Eqs. (7.15) and (7.17) to determine both the expectation and variance of  $Y$ . However, the waiting time for the bus is governed by the relation  $Y = 3X$  where  $X$  is the waiting time for the train. Therefore, we can calculate  $E(Y) = E(3X) = 3E(X) = 3 \cdot 10 = 30$  min by using rule (7.29) and the variance as  $\text{Var}(Y) = \text{Var}(3X) = 3^2 \text{Var}(X) = 9 \cdot 33\frac{1}{3} = 300$  using rule (7.33). The total waiting time is the sum of the two waiting times.

### 7.6.1 Expectation and Variance of the Arithmetic Mean

**Definition 7.6.1** We define the random variables  $X_1, X_2, \dots, X_n$  to be i.i.d. (independently identically distributed), if all  $X_i$  follow the same distribution and are stochastically independent of each other.

Let  $X_1, X_2, \dots, X_n$  be  $n$  i.i.d. random variables with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2, i = 1, 2, \dots, n$ . The arithmetic mean of these variables is given by

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

which is again a random variable that follows a distribution with certain expectation and variance. A function of random variables is called a **statistic**. By using (7.29) and (7.31), we obtain

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \mu. \quad (7.35)$$

If we apply (7.34) and recall that the variables are independent of each other, we can also calculate the variance as

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}. \quad (7.36)$$

*Example 7.6.3* If we toss a coin, we obtain either head or tail, and therefore,  $P(\text{"head"}) = P(\text{"tail"}) = \frac{1}{2}$ . If we toss the coin  $n$  times, we have for each toss

$$X_i = \begin{cases} 0 & \text{for "tail"} \\ 1 & \text{for "head"} \end{cases}, \quad i = 1, \dots, n.$$

It is straightforward to calculate the expectation and variance for each coin toss:

$$\begin{aligned} \mathbb{E}(X_i) &= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}, \\ \text{Var}(X_i) &= (0 - \frac{1}{2})^2 \cdot \frac{1}{2} + (1 - \frac{1}{2})^2 \cdot \frac{1}{2} = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

The arithmetic mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  describes the relative frequency of heads when the coin is tossed  $n$  times. We can now apply (7.35) and (7.36) to calculate

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^n 1/2 = 1/2$$

and

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \frac{1}{4} = \frac{1}{4n}.$$

With this example, the interpretation of formulae (7.35) and (7.36) becomes clearer: if the probability of head is 0.5 for a single toss, then it is also 0.5 for the mean of all tosses. If we toss a coin many times, then the variance decreases when  $n$  increases. This means that a larger sample size yields a higher precision for the calculated arithmetic mean. This observation shows the basic conclusion of the next chapter: the higher the sample size, the more secure we are of our conclusions.

## 7.7 Covariance and Correlation

The variance measures the variability of a variable. Similarly, the covariance measures the covariation or association between  $X$  and  $Y$ .

### 7.7.1 Covariance

**Definition 7.7.1** The **covariance** between  $X$  and  $Y$  is defined as

$$\varrho = \text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]. \quad (7.37)$$

The covariance is positive if, on average, larger values of  $X$  correspond to larger values of  $Y$ ; it is negative if, on average, greater values of  $X$  correspond to smaller values of  $Y$ .

The probability density function of any bivariate random variable  $(X, Y)$  is characterized by the expectation and variance of both  $X$  and  $Y$ ,

$$\begin{aligned} E(X) &= \mu_X, & \text{Var}(X) &= \sigma_X^2, \\ E(Y) &= \mu_Y, & \text{Var}(Y) &= \sigma_Y^2, \end{aligned}$$

as well as their **covariance**. We can summarize these features by using the expectation vector

$$E \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} E(X) \\ E(Y) \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

and the **covariance matrix**

$$\text{Cov} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \varrho \\ \varrho & \sigma_Y^2 \end{pmatrix}.$$

Important properties of covariance are

- (i)  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ ,
- (ii)  $\text{Cov}(X, X) = \text{Var}(X)$ ,
- (iii)  $\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$ ,
- (iv)  $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$  where  $E(XY) = \int \int xyf(x, y)dxdy$  for continuous variables and  $E(XY) = \sum_i \sum_j x_i y_j p_{ij}$  for discrete variables,
- (v) If  $X$  and  $Y$  are independent, it follows that  $E(XY) = E(X)E(Y) = \mu_X\mu_Y$ , and therefore,  $\text{Cov}(X, Y) = \mu_X\mu_Y - \mu_X\mu_Y = 0$ .

**Theorem 7.7.1** (Additivity Theorem) *The variance of the sum (subtraction) of  $X$  and  $Y$  is given by*

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2 \text{Cov}(X, Y).$$

If  $X$  and  $Y$  are independent, it follows that  $\text{Cov}(X, Y) = 0$  and therefore  $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ . We omit the proof of this theorem.

**Example 7.7.1** Recall Example 7.6.2 where we considered the waiting time  $Y$  for a bus to the train station and the waiting time  $X$  for the waiting time for a train. Suppose their joint bivariate probability density function can be written as

$$f_{XY}(x, y) = \begin{cases} \frac{1}{1200} & \text{for } 0 \leq x \leq 60, \quad 0 \leq y \leq 20 \\ 0 & \text{elsewhere.} \end{cases}$$

To calculate the covariance between  $X$  and  $Y$ , we need to calculate  $E(XY)$ :

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dx dy = \int_0^{60} \int_0^{20} xy \frac{1}{1200} dx dy \\ &= \int_0^{60} \left[ \frac{x}{1200} \frac{y^2}{2} \right]_0^{20} dy = \int_0^{60} \frac{400x}{2400} dy = \left[ \frac{1}{6} \frac{x^2}{2} \right]_0^{60} = \frac{3600}{12} = 300. \end{aligned}$$

We know from Example 7.6.2 that  $E(X) = 10$ ,  $E(Y) = 30$ ,  $\text{Var}(X) = 33\frac{1}{3}$ , and  $\text{Var}(Y) = 300$ . The covariance is thus

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 300 - 30 \cdot 10 = 0.$$

This makes sense as the waiting times for the train and the bus should be independent of each other. Using rule (7.31), we conclude that the total expected waiting time is

$$E(X + Y) = E(X) + E(Y) = 10 + 30 = 40 \text{ min.}$$

The respective variance is

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y) = 33\frac{1}{3} + 300 - 2 \cdot 0 = 333\frac{1}{3}$$

due to Theorem 7.7.1.

### 7.7.2 Correlation Coefficient

**Definition 7.7.2** The **correlation coefficient** of  $X$  and  $Y$  is defined as

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}. \quad (7.38)$$

We already know from Chap. 4 that the correlation coefficient is a measure of the degree of linear relationship between  $X$  and  $Y$ . It can take values between  $-1$  and  $1$ ,  $-1 \leq \rho(X, Y) \leq 1$ . However in Chap. 4, we considered the correlation of two samples, i.e. realizations of random variables; here, we describe the correlation coefficient of the population. If  $\rho(X, Y) = 0$ , then  $X$  and  $Y$  are said to be uncorrelated. If there is a perfect linear relationship between  $X$  and  $Y$ , then  $\rho = 1$  for a positive relationship and  $\rho = -1$  for a negative relationship, see Appendix C.2 for the proof.

**Theorem 7.7.2** *If  $X$  and  $Y$  are independent, they are also uncorrelated. However, if they are uncorrelated then they are not necessarily independent.*

*Example 7.7.2* In Example 7.6.2, we estimated the covariance between the waiting time for the bus and the waiting time for the train:  $\text{Cov}(X, Y) = 0$ . The correlation coefficient is therefore also 0 indicating no linear relationship between the waiting times for bus and train.

## 7.8 Key Points and Further Issues

### Note:

- ✓ Note that there is a difference between the empirical cumulative distribution function introduced in Chap. 2 and the CDF introduced in this chapter. In Chap. 2, we work with the cumulative distribution of observed values in a particular sample, whereas in this chapter, we deal with random variables modelling the distribution of a general population.
- ✓ The expectation and the variance of a random variable are defined as follows:

	Expectation	Variance
Discrete	$\sum_{i=1}^n x_i p_i$	$\sum_{i=1}^n (x_i - E(X))^2 p_i$
Continuous	$\int_{-\infty}^{+\infty} x f(x) dx$	$\int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx$

- ✓ Some important calculation rules are:

$$\begin{aligned} E(a + bX) &= a + bE(X); & \text{Var}(a + bX) &= b^2 \text{Var}(X); \\ E(X + Y) &= E(X) + E(Y); & \text{Var}(X \pm Y) &= \text{Var}(X) + \text{Var}(Y) \\ &&& \pm 2 \text{Cov}(X, Y) \end{aligned}$$

- ✓ Bivariate random variables  $(X, Y)$  have a joint CDF  $F_{XY}(x, y)$  which specifies the probability  $P(X \leq x; Y \leq y)$ . The conditional distribution of  $X|Y$  [ $Y|X$ ] is the PDF of  $X$  [ $Y$ ] for a given value  $Y = y$  [ $X = x$ ]. The marginal distribution of  $X$  [ $Y$ ] is the distribution of  $X$  [ $Y$ ] without referring to the values of  $Y$  [ $X$ ].

---

## 7.9 Exercises

*Exercise 7.1* Consider the following cumulative distribution function of a random variable  $X$ :

$$F(x) = \begin{cases} 0 & \text{if } x < 2 \\ -\frac{1}{4}x^2 + 2x - 3 & \text{if } 2 \leq x \leq 4 \\ 1 & \text{if } x > 4. \end{cases}$$

- What is the PDF of  $X$ ?
- Calculate  $P(X < 3)$  and  $P(X = 4)$ .
- Determine  $E(X)$  and  $\text{Var}(X)$ .

*Exercise 7.2* Joey manipulates a die to increase his chances of winning a board game against his friends. In each round, a die is rolled and larger numbers are generally an advantage. Consider the random variable  $X$  denoting the outcome of the rolled die and the respective probabilities  $P(X = 1 = 2 = 3 = 5) = 1/9$ ,  $P(X = 4) = 2/9$ , and  $P(X = 6) = 3/9$ .

- (a) Calculate and interpret the expectation and variance of  $X$ .
- (b) Imagine that the board game contains an action which makes the players use  $1/X$  rather than  $X$ . What is the expectation of  $Y = 1/X$ ? Is  $E(Y) = E(1/X) = 1/E(X)$ ?

*Exercise 7.3* An innovative winemaker experiments with new grapes and adds a new wine to his stock. The percentage sold by the end of the season depends on the weather and various other factors. It can be modelled using the random variable  $X$  with the CDF as

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 3x^2 - 2x^3 & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

- (a) Plot the cumulative distribution function with  $R$ .
- (b) Determine  $f(x)$ .
- (c) What is the probability of selling at least one-third of his wine, but not more than two thirds?
- (d) Define the CDF in  $R$  and calculate the probability of c) again.
- (e) What is the variance of  $X$ ?

*Exercise 7.4* A quality index summarizes different features of a product by means of a score. Different experts may assign different quality scores depending on their experience with the product. Let  $X$  be the quality index for a tablet. Suppose the respective probability density function is given as follows:

$$f(x) = \begin{cases} cx(2-x) & \text{if } 0 \leq x \leq 2 \\ 0 & \text{elsewhere.} \end{cases}$$

- (a) Determine  $c$  such that  $f(x)$  is a proper PDF.
- (b) Determine the cumulative distribution function.
- (c) Calculate the expectation and variance of  $X$ .
- (d) Use Tschebyschev's inequality to determine the probability that  $X$  does not deviate more than 0.5 from its expectation.

*Exercise 7.5* Consider the joint PDF for the type of customer service  $X$  (0 = telephonic hotline, 1 = Email) and of satisfaction score  $Y$  (1 = unsatisfied, 2 = satisfied, 3 = very satisfied):

$X \setminus Y$	1	2	3
0	0	$1/2$	$1/4$
1	$1/6$	$1/12$	0

- (a) Determine and interpret the marginal distributions of both  $X$  and  $Y$ .
- (b) Calculate the 75 % quantile for the marginal distribution of  $Y$ .
- (c) Determine and interpret the conditional distribution of satisfaction level for  $X = 1$ .
- (d) Are the two variables independent?
- (e) Calculate and interpret the covariance of  $X$  and  $Y$ .

*Exercise 7.6* Consider a continuous random variable  $X$  with expectation 15 and variance 4. Determine the smallest interval  $[15 - c, 15 + c]$  which contains at least 90 % of the values of  $X$ .

*Exercise 7.7* Let  $X$  and  $Y$  be two random variables for which only 6 possible events— $A_1, A_2, A_3, A_4, A_5, A_6$ —are defined:

$i$	1	2	3	4	5	6
$P(A_i)$	0.3	0.1	0.1	0.2	0.2	0.1
$X_i$	-1	2	2	-1	-1	2
$Y_i$	0	2	0	1	2	1

- (a) What is the joint PDF of  $X$  and  $Y$ ?
- (b) Calculate the marginal distributions of  $X$  and  $Y$ .
- (c) Are both variables independent?
- (d) Determine the joint PDF for  $U = X + Y$ .
- (e) Calculate  $E(U)$  and  $\text{Var}(U)$  and compare it with  $E(X) + E(Y)$  and  $\text{Var}(X) + \text{Var}(Y)$ , respectively.

*Exercise 7.8* Recall the urn model we introduced in Chap. 5. Consider an urn with eight balls: four of them are white, three are black, and one is red. Now, two balls are drawn from the urn. The random variables  $X$  and  $Y$  are defined as follows:

$$X = \begin{cases} 1 & \text{black ball} \\ 2 & \text{red ball in the first draw} \\ 3 & \text{white ball} \end{cases}$$

$$Y = \begin{cases} 1 & \text{black ball} \\ 2 & \text{red ball in the second draw} \\ 3 & \text{white ball.} \end{cases}$$

- (a) When are  $X$  and  $Y$  independent—when the two balls are drawn with replacement or without replacement?
- (b) Assume the balls are drawn such that  $X$  and  $Y$  are dependent. Use the conditional distribution  $P(Y|X)$  to determine the joint PDF of  $X$  and  $Y$ .
- (c) Calculate  $E(X)$ ,  $E(Y)$ , and  $\rho(X, Y)$ .

*Exercise 7.9* If  $X$  is the amount of money spent on food and other expenses during a day (in €) and  $Y$  is the daily allowance of a businesswoman, the joint density of these two variables is given by

$$f_{XY}(x, y) = \begin{cases} c \left( \frac{100-x}{x} \right) & \text{if } 10 \leq x \leq 100, \quad 40 \leq y \leq 100 \\ 0 & \text{elsewhere.} \end{cases}$$

- (a) Choose  $c$  such that  $f_{XY}(x, y)$  is a probability density function.
- (b) Find the marginal distribution of  $X$ .
- (c) Calculate the probability that more than €75 are spent.
- (d) Determine the conditional distribution of  $Y$  given  $X$ .

*Exercise 7.10* Consider  $n$  i.i.d. random variables  $X_i$  with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$  and the standardized variable  $Y = \frac{X-\mu}{\sigma}$ . Show that  $E(Y) = 0$  and  $\text{Var}(Y) = 1$ .

→ Solutions to all exercises in this chapter can be found on p. [365](#)

\*Source Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg

We introduced the concept of probability density and probability mass functions of random variables in the previous chapter. In this chapter, we are introducing some common standard discrete and continuous probability distributions which are widely used for either practical applications or constructing statistical methods described later in this book. Suppose we are interested in determining the probability of a certain event. The determination of probabilities depends upon the nature of the study and various prevailing conditions which affect it. For example, the determination of the probability of a head when tossing a coin is different from the determination of the probability of rain in the afternoon. One can speculate that some mathematical functions can be defined which depict the behaviour of probabilities under different situations. Such functions have special properties and describe how probabilities are distributed under different conditions. We have already learned that they are called probability distribution functions. The form of such functions may be simple or complicated depending upon the nature and complexity of the phenomenon under consideration. Let us first recall and extend the definition of independent and identically distributed random variables:

**Definition 8.0.1** The random variables  $X_1, X_2, \dots, X_n$  are called independent and identically distributed (i.i.d) if the  $X_i$  ( $i = 1, 2, \dots, n$ ) have the same marginal cumulative distribution function  $F(x)$  and if they are mutually independent.

*Example 8.0.1* Suppose a researcher plans a survey on the weight of newborn babies in a country. The researcher randomly contacts 10 hospitals with a maternity ward and asks them to randomly select 20 of the newborn babies (no twins) born in the last 6 months and records their weights. The sample therefore consists of  $10 \times 20 = 200$  baby weights. Since the hospitals and the babies are randomly selected, the babies' weights are therefore not known beforehand. The 200 weights can be denoted by the random variables  $X_1, X_2, \dots, X_{200}$ . Note that the weights  $X_i$  are random variables

because, depending on the size of the population, different samples consisting of 200 babies can be randomly selected. Also, the babies' weights can be seen as stochastically independent (an example of stochastically dependent weights would be the weights of twins if they are included in the sample). After collecting the weights of 200 babies, the researcher has a sample of 200 realized values (i.e. the weights in grams). The values are now known and denoted by  $x_1, x_2, \dots, x_{200}$ .

## 8.1 Standard Discrete Distributions

First, we discuss some standard distributions for discrete random variables.

### 8.1.1 Discrete Uniform Distribution

The discrete uniform distribution assumes that all possible outcomes have equal probability of occurrence. A more formal definition is given as follows:

**Definition 8.1.1** A discrete random variable  $X$  with  $k$  possible outcomes  $x_1, x_2, \dots, x_k$  is said to follow a discrete **uniform** distribution if the probability mass function (PMF) of  $X$  is given by

$$P(X = x_i) = \frac{1}{k}, \quad \forall i = 1, 2, \dots, k. \quad (8.1)$$

If the outcomes are the natural numbers  $x_i = i$  ( $i = 1, 2, \dots, k$ ), the mean and variance of  $X$  are obtained as

$$\mathbb{E}(X) = \frac{k+1}{2}, \quad (8.2)$$

$$\text{Var}(X) = \frac{1}{12}(k^2 - 1). \quad (8.3)$$

*Example 8.1.1* If we roll a fair die, the outcomes "1", "2", ..., "6" have equal probability of occurring, and hence, the random variable  $X$  "number of dots observed on the upper surface of the die" has a uniform discrete distribution with PMF

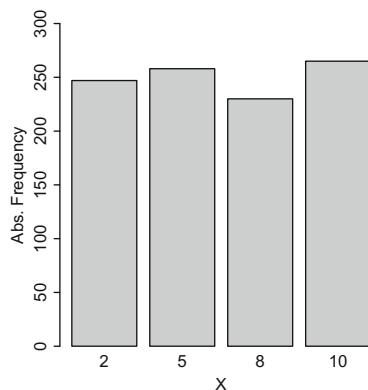
$$P(X = i) = \frac{1}{6}, \quad \text{for all } i = 1, 2, \dots, 6.$$

The mean and variance of  $X$  are

$$\mathbb{E}(X) = \frac{6+1}{2} = 3.5,$$

$$\text{Var}(X) = \frac{1}{12}(6^2 - 1) = 35/12.$$

**Fig. 8.1** Frequency distribution of 1000 generated discrete uniform random numbers with possible outcomes (2, 5, 8, 10)



Using the function `sample()` in *R*, it is easy to generate random numbers from a discrete uniform distribution. The following command generates a random sample of size 1000 from a uniform distribution with the four possible outcomes 2, 5, 8, 10 and draws a bar chart of the observed numbers. The use of the `set.seed()` function allows to reproduce the generated random numbers at any time. It is necessary to use the option `replace=TRUE` to simulate draws with replacement, i.e. to guarantee that a value can occur more than once.

```
set.seed(123789)
x <- sample(x=c(2,5,8,10), size=1000, replace=T,
prob=c(1/4,1/4,1/4,1/4))
barchart(table(x), ylim=c(0,300))
```



A bar chart of the frequency distribution of the 1000 sampled numbers with the possible outcomes (2, 5, 8, 10) using the discrete uniform distribution is given in Fig. 8.1. We see that the 1000 generated random numbers are not exactly uniformly distributed, e.g. the numbers 5 and 10 occur more often than the numbers 2 and 8. In fact, they are only approximately uniform. We expect that the deviance from a perfect uniform distribution is getting smaller as we generate more and more random numbers but will probably never be zero for a finite number of draws. The random numbers reflect the practical situation that a sample distribution is only an approximation to the theoretical distribution from which the sample was drawn. More details on how to work with random variables in *R* are given in Appendix A.3.

### 8.1.2 Degenerate Distribution

**Definition 8.1.2** A random variable  $X$  has a **degenerate distribution** at  $a$ , if  $a$  is the only possible outcome with  $P(X = a) = 1$ . The CDF in such a case is given by

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ 1 & \text{if } x \geq a. \end{cases}$$

Further,  $E(X) = a$  and  $\text{Var}(X) = 0$ .

The degenerate distribution indicates that there is only one possible fixed outcome, and therefore, no randomness is involved. It follows that we need at least two different possible outcomes to have randomness in the observations of a random variable or random experiment. The Bernoulli distribution is such a distribution where there are only two outcomes, e.g. success and failure or male and female. These outcomes are usually denoted by the values “0” and “1”.

### 8.1.3 Bernoulli Distribution

**Definition 8.1.3** A random variable  $X$  has a Bernoulli distribution if the PMF of  $X$  is given as

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0. \end{cases}$$

The cumulative distribution function (CDF) of  $X$  is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

The mean (expectation) and variance of a Bernoulli random variable are calculated as

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p \quad (8.4)$$

and

$$\text{Var}(X) = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p), \quad (8.5)$$

respectively.

A Bernoulli distribution is useful when there are only two possible outcomes, and our interest lies in any of the two outcomes, e.g. whether a customer buys a certain product or not, or whether a hurricane hits an island or not. The outcome of an event  $A$  is usually coded as 1 which occurs with probability  $p$ . If the event of interest does not occur, i.e. the complementary event  $\bar{A}$  occurs, the outcome is coded as 0 which occurs with probability  $1 - p$ . So  $p$  is the probability that the event of interest  $A$  occurs.

*Example 8.1.2* A company organizes a raffle at an end-of-year function. There are 300 lottery tickets in total, and 50 of them are marked as winning tickets. The event  $A$  of interest is “ticket wins” (coded as  $X = 1$ ), and the probability  $p$  of having a winning ticket is *a priori* (i.e. before any lottery ticket has been drawn)

$$P(X = 1) = \frac{50}{300} = \frac{1}{6} = p \quad \text{and} \quad P(X = 0) = \frac{250}{300} = \frac{5}{6} = 1 - p.$$

According to (8.4) and (8.5), the mean (expectation) and variance of  $X$  are

$$E(X) = \frac{1}{6} \quad \text{and} \quad \text{Var}(X) = \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36} \text{ respectively.}$$

### 8.1.4 Binomial Distribution

Consider  $n$  independent trials or repetitions of a Bernoulli experiment. In each trial or repetition, we may observe either  $A$  or  $\bar{A}$ . At the end of the experiment, we have thus observed  $A$  between 0 and  $n$  times. Suppose we are interested in the probability of  $A$  occurring  $k$  times, then the binomial distribution is useful.

*Example 8.1.3* Consider a coin tossing experiment where a coin is tossed ten times and the event of interest is  $A$  = “head”. The random variable  $X$  “number of heads in 10 experiments” has the possible outcomes  $k = 0, 1, \dots, 10$ . A question of interest may be: What is the probability that a head occurs in 7 out of 10 trials; or in 5 out of 10 trials? We assume that the order in which heads (and tails) appear is not of interest, only the total number of heads is of interest.

Questions of this kind are answered by the binomial distribution. This distribution can either be motivated as a repetition of  $n$  Bernoulli experiments (as in the above coin tossing example) or by the urn model (see Chap. 5): assume there are  $M$  white and  $N - M$  black balls in the urn. Suppose  $n$  balls are drawn randomly from the urn, the colour of the ball is recorded and the ball is placed back into the urn (sampling with replacement). Let  $A$  be the event of interest that a white ball is drawn from the urn. The probability of  $A$  is  $p = M/N$  (the probability of drawing a black ball is  $1 - p = (N - M)/N$ ). Since the balls are drawn with replacement, these probabilities do not change from draw to draw. Further, let  $X$  be the random variable counting the number of white balls drawn from the urn in the  $n$  experiments. Since the order of the resulting colours of balls is not of interest in this case, there are  $\binom{n}{k}$  combinations where  $k$  balls are white and  $n - k$  balls are black. Since the balls are drawn with replacement, every outcome of the  $n$  experiments is independent of all others. The probability that  $X = k$ ,  $k = 0, 1, \dots, n$ , can therefore be calculated as

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (k = 0, 1, \dots, n). \quad (8.6)$$

Please note that we can use the product  $p^k(1 - p)^{n-k}$  because the draws are independent. The binomial coefficient  $\binom{n}{k}$  is necessary to count the number of possible orders of the black and white balls.

**Definition 8.1.4** A discrete random variable  $X$  is said to follow a binomial distribution with parameters  $n$  and  $p$  if its PMF is given by (8.6). We also write  $X \sim B(n; p)$ . The mean and variance of a binomial random variable  $X$  are given by

$$\text{E}(X) = np, \quad (8.7)$$

$$\text{Var}(X) = np(1 - p). \quad (8.8)$$

*Remark 8.1.1* A Bernoulli random variable is therefore  $B(1; p)$  distributed.

*Example 8.1.4* Consider an unfair coin where the probability of observing a tail ( $T$ ) is  $p(T) = 0.6$ . Let us denote tails by “1” and heads by “0”. Suppose the coin is tossed three times. In total, there are the  $2^3 = 8$  following possible outcomes:

Outcome	$X = x$
1 1 1	3
1 1 0	2
1 0 1	2
0 1 1	2
1 0 0	1
0 1 0	1
0 0 1	1
0 0 0	0

Note that the first outcome, viz. (1, 1, 1) leads to  $x = 3$ , the next 3 outcomes, viz., (1, 1, 0), (1, 0, 1), (0, 1, 1) obtained by  $(= \binom{3}{2})$  lead to  $x = 2$ , the next 3 outcomes, viz., (1, 0, 0), ((0, 1, 0), (0, 0, 1) obtained by  $(= \binom{3}{1})$ ) lead to  $x = 1$ , and the last outcome, viz. (0, 0, 0) obtained by  $(= \binom{3}{0})$  leads to  $x = 0$ . We can, for example, calculate

$$P(X = 2) = \binom{3}{2} 0.6^2 (1 - 0.6)^1 = 0.432 \quad (\text{or } 43.2\%).$$

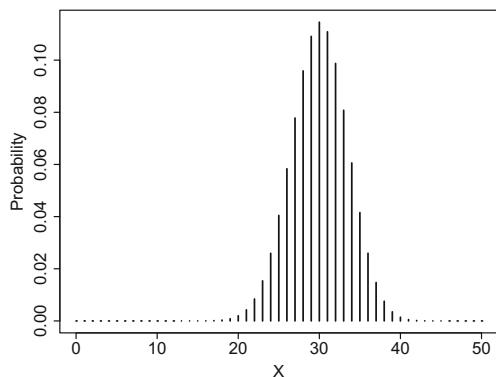
Further, the mean and variance of  $X$  are

$$\text{E}(X) = np = 3 \cdot 0.6 = 1.8, \quad \text{and} \quad \text{Var}(X) = np(1 - p) = 3 \cdot 0.6 \cdot 0.4 = 0.72.$$

Functions for the binomial distribution, as well as many other distributions, are implemented in *R*. For each of these distributions, we can easily determine the density function (PMF, PDF) for given values and parameters, determine the CDF, calculate quantiles and draw random numbers. Appendix A.3 gives more details. Nevertheless, we illustrate the concept of dealing with distributions in *R* in the following example.

*Example 8.1.5* Suppose we roll an unfair die 50 times with the probability of a tail  $p_{\text{tail}} = 0.6$ . We thus deal with a  $B(50, 0.6)$  distribution which can be plotted using the `dbinom` command. The prefix `d` stands for “density”.

**Fig. 8.2** PMF of a  $B(50, 0.6)$  distribution



```
n <- 50
p <- 0.6
k <- 0:n
pmf <- dbinom(k,n,p)
plot(k,pmf, type=h)
```



A plot of the PMF of a binomial distribution with  $n = 50$  and  $p = 0.6$  (i.e.  $B(50, 0.6)$ ) is given in Fig. 8.2.

Note that we can also calculate the CDF with *R*. We can use the `pbinom(x,n,p)` command, where the prefix `p` stands for probability, to calculate the CDF at any point. For example, suppose we are interested in  $P(X \geq 30) = 1 - F(29)$ , that is the probability of observing thirty or more tails; then we write

```
1-pbinom(29,50,0.6)
[1] 0.5610349
```



Similarly, we can determine quantiles. For instance, the 80 % quantile  $q$  which describes that  $P(X \leq q) \geq 0.8$  can be obtained by the `qbinom(q,n,p)` command as follows:

```
qbinom(0.8,50,0.6)
[1] 33
```



If we want to generate 100 random realizations from a  $B(50, 0.6)$  distribution we can use the `rbinom` command.

```
rbinom(100,50,0.6)
```



The binomial distribution has some nice properties. One of them is described in the following theorem:

**Theorem 8.1.1** Let  $X \sim B(n; p)$  and  $Y \sim B(m; p)$  and assume that  $X$  and  $Y$  are (stochastically) independent. Then

$$X + Y \sim B(n + m; p). \quad (8.9)$$

This is intuitively clear since we can interpret this theorem as describing the additive combination of two independent binomial experiments with  $n$  and  $m$  trials, with equal probability  $p$ , respectively. Since every binomial experiment is a series of independent Bernoulli experiments, this is equivalent to a series of  $n + m$  independent Bernoulli trials with constant success probability  $p$  which in turn is equivalent to a binomial distribution with  $n + m$  trials.

### 8.1.5 Poisson Distribution

Consider a situation in which the number of events is very large and the probability of success is very small: for example, the number of alpha particles emitted by a radioactive substance entering a particular region in a given short time interval. Note that the number of emitted alpha particles is very high but only a few particles are transmitted through the region in a given short time interval. Some other examples where Poisson distributions are useful are the number of flu cases in a country within one year, the number of tropical storms within a given area in one year, or the number of bacteria found in a biological investigation.

**Definition 8.1.5** A discrete random variable  $X$  is said to follow a Poisson distribution with parameter  $\lambda > 0$  if its PMF is given by

$$P(X = x) = \frac{\lambda^x}{x!} \exp(-\lambda) \quad (x = 0, 1, 2, \dots). \quad (8.10)$$

We also write  $X \sim Po(\lambda)$ . The mean and variance of a Poisson random variable are identical:

$$E(X) = \text{Var}(X) = \lambda.$$

*Example 8.1.6* Suppose a country experiences  $X = 4$  tropical storms on average per year. Then the probability of suffering from only two tropical storms is obtained by using the Poisson distribution as

$$P(X = 2) = \frac{\lambda^x}{x!} \exp(-\lambda) = \frac{4^2}{2!} \exp(-4) = 0.146525.$$

If we are interested in the probability that not more than 2 storms are experienced, then we can apply rules (7.7)–(7.13) from Chap. 7:  $P(X \leq 2) = P(X = 2) + P(X = 1) + P(X = 0) = F(2) = 0.2381033$ . We can calculate  $P(X = 1)$  and  $P(X = 0)$  from (8.10) or using R. Similar to Example 8.1.5, we use the prefix d to obtain the PMF and the prefix p to work with the CDF, i.e. we can use `dpois(x, λ)` and `ppois(x, λ)` to determine  $P(X = x)$  and  $P(X \leq x)$ , respectively.

```
dpois(2,4) + dpois(1,4) + dpois(0,4)
[1] 0.2381033
ppois(2,4)
[1] 0.2381033
```



### 8.1.6 Multinomial Distribution

We now consider random experiments where  $k$  distinct or disjoint events  $A_1, A_2, \dots, A_k$  can occur with probabilities  $p_1, p_2, \dots, p_k$ , respectively, with the restriction  $\sum_{j=1}^k p_j = 1$ . For example, if eight parties compete in a political election, we may be interested in the probability that a person votes for party  $A_j$ ,  $j = 1, 2, \dots, 8$ . Similarly one might be interested in the probability whether tuberculosis is detected in the lungs ( $A_1$ ), in other organs ( $A_2$ ), or both ( $A_3$ ). Practically, we often use the multinomial distribution to model the distribution of categorical variables. This can be interpreted as a generalization of the binomial distribution (where only two distinct events can occur) to the situation where more than two events or outcomes can occur. If the experiment is repeated  $n$  times independently, we are interested in the probability that

$A_1$  occurs  $n_1$ -times,  $A_2$  occurs  $n_2$ -times,  $\dots$ ,  $A_k$  occurs  $n_k$ -times

with  $\sum_{j=1}^k n_j = n$ . Since several events can occur, the outcome of one (of the  $n$ ) experiments is conveniently described by binary indicator variables. Let  $V_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, k$ , denote the event “ $A_j$  is observed in experiment  $i$ ”, i.e.

$$V_{ij} = \begin{cases} 1 & \text{if } A_j \text{ occurs in experiment } i \\ 0 & \text{if } A_j \text{ does not occur in experiment } i \end{cases}$$

with probabilities  $P(V_{ij} = 1) = p_j$ ,  $j = 1, 2, \dots, k$ ; then, the outcome of one experiment is a vector of length  $k$ ,

$$V_i = (V_{i1}, \dots, V_{ij}, \dots, V_{ik}) = (0, \dots, 1, \dots, 0),$$

with “1” being present in only one position, i.e. in position  $j$ , if  $A_j$  occurs in experiment  $i$ . Now, define (for each  $j = 1, \dots, k$ )  $X_j = \sum_{i=1}^n V_{ij}$ . Then,  $X_j$  is counting how often event  $A_j$  was observed in the  $n$  independent experiments (i.e. how often  $V_{ij}$  was 1 in the  $n$  experiments).

**Definition 8.1.6** The random vector  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  is said to follow a **multinomial distribution** if its PMF is given as

$$P(X_1 = n_1, X_2 = n_2, \dots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} \cdot p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \quad (8.11)$$

with the restrictions  $\sum_{j=1}^k n_j = n$  and  $\sum_{j=1}^k p_j = 1$ . We also write  $\mathbf{X} \sim M(n; p_1, \dots, p_k)$ . The mean of  $\mathbf{X}$  is the (component-wise) vector

$$\begin{aligned} E(\mathbf{X}) &= (E(X_1), E(X_2), \dots, E(X_k)) \\ &= (np_1, np_2, \dots, np_k). \end{aligned}$$

The  $(i, j)$ th element of the covariance matrix  $V(\mathbf{X})$  is

$$\text{Cov}(X_i, X_j) = \begin{cases} np_i(1 - p_i) & \text{if } i = j \\ -np_i p_j & \text{if } i \neq j. \end{cases}$$

*Remark 8.1.2* Due to the restriction that  $\sum_{j=1}^k n_j = \sum_{j=1}^k X_j = n$ ,  $X_1, \dots, X_k$  are not stochastically independent which is reflected by the negative covariance. This is also intuitively clear: if one  $X_j$  gets higher, another  $X_{j'}$ ,  $j \neq j'$ , has to become lower to satisfy the restrictions.

We use the multinomial distribution to describe the randomness of categorical variables. Suppose we are interested in the variable “political party”; there might be eight political parties, and we could thus summarize this variable by eight binary variables, each of them describing the event of party  $A_j$ ,  $j = 1, 2, \dots, 8$ , being voted for. In this sense,  $\mathbf{X} = (X_1, X_2, \dots, X_8)$  follows a multinomial distribution.

*Example 8.1.7* Consider a simple example of the urn model. The urn contains 50 balls of three colours: 25 red balls, 15 white balls, and 10 black balls. The balls are drawn from the urn with replacement. The balls are placed back into the urn after every draw, which means the draws are independent. Therefore, the probability of drawing a red ball in every draw is  $p_1 = \frac{25}{50} = 0.5$ . Analogously,  $p_2 = 0.3$  (for white balls) and  $p_3 = 0.2$  (for black balls). Consider  $n = 4$  draws. The probability of the random event of drawing “2 red balls, 1 white ball, and 1 black ball” is:

$$P(X_1 = 2, X_2 = 1, X_3 = 1) = \frac{4!}{2!1!1!}(0.5)^2(0.3)^1(0.2)^1 = 0.18. \quad (8.12)$$

We would have obtained the same result in R using the `dmultinom` function:

`dmultinom(c(2,1,1), prob=c(0.5,0.3,0.2))`



This example demonstrates that the multinomial distribution relates to an experiment with replacement and without considering the order of the draws. Instead of the urn model, consider another example where we may interpret these three probabilities as probabilities of voting for candidate  $A_j$ ,  $j = 1, 2, 3$ , in an election. Now, suppose we ask four voters about their choice, then the probability of candidate  $A_1$  receiving 2 votes, candidate  $A_2$  receiving 1 vote, and candidate  $A_3$  receiving 1 vote is 18 % as calculated in (8.12).

*Remark 8.1.3* In contrast to most of the distributions, the CDF of the multinomial distribution, i.e. the function calculating  $P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_k \leq x_k)$ , is not contained in the base R-distribution. Please note that for  $k = 2$ , the multinomial distribution reduces to the binomial distribution.

### 8.1.7 Geometric Distribution

Consider a situation in which we are interested in determining how many independent Bernoulli trials are needed until the event of interest occurs for the first time. For instance, we may be interested in how many tickets to buy in a raffle until we win for the first time, or how many different drugs to try to successfully tackle a severe migraine, etc. The geometric distribution can be used to determine the probability that the event of interest happens at the  $k$ th trial for the first time.

**Definition 8.1.7** A discrete random variable  $X$  is said to follow a geometric distribution with parameter  $p$  if its PMF is given by

$$P(X = k) = p(1 - p)^{k-1}, \quad k = 1, 2, 3, \dots \quad (8.13)$$

The mean (expectation) and variance are given by  $E(X) = 1/p$  and  $\text{Var}(X) = 1/p(1/p - 1)$ , respectively.

*Example 8.1.8* Let us consider an experiment where a coin is tossed until “head” is obtained for the first time. The probability of getting a head is  $p = 0.5$  for each toss. Using (8.13), we can determine the following probabilities:

$$\begin{aligned} P(X = 1) &= 0.5 \\ P(X = 2) &= 0.5(1 - 0.5) = 0.25 \\ P(X = 3) &= 0.5(1 - 0.5)^2 = 0.125 \\ P(X = 4) &= 0.5(1 - 0.5)^3 = 0.0625 \\ &\dots \quad \dots \end{aligned}$$

Using the command structure for obtaining PMF’s in *R* (Appendix A as well as Examples 8.1.5 and 8.1.6), we can determine the latter probability of  $P(X = 4)$  as follows:

`dgeom(3, 0.5)`



Note that the definition of  $X$  in *R* slightly differs from our definition. In *R*,  $k$  is the number of failures before the first success. This means we need to specify  $k - 1$  in the `dgeom` function rather than  $k$ . The mean and variance for this setting are

$$E(X) = \frac{1}{0.5} = 2; \quad \text{Var}(X) = \frac{1}{0.5} \left( \frac{1}{0.5} - 1 \right) = 2.$$

### 8.1.8 Hypergeometric Distribution

We can again use the urn model to motivate another distribution, the hypergeometric distribution. Consider an urn with  $N$  balls randomly drawn  $n$  balls without replacement,

$M$	white balls
$N - M$	black balls
$N$	total balls

i.e. we do not place a ball back into the urn once it is drawn. The order in which the balls are drawn is assumed to be of no interest; only the number of drawn white balls is of relevance. We define the following random variable

$X$  : “number of white balls ( $x$ ) among the  $n$  drawn balls”.

To be more precise, among the  $n$  drawn balls,  $x$  are white and  $n - x$  are black. There are  $\binom{M}{x}$  possibilities to choose  $x$  white balls from the total of  $M$  white balls, and analogously, there are  $\binom{N-M}{n-x}$  possibilities to choose  $(n - x)$  black balls from the total of  $N - M$  black balls. In total, we draw  $n$  out of  $N$  balls. Recall the probability definition of Laplace as the number of simple favourable events divided by all possible events. The number of combinations for all possible events is  $\binom{N}{n}$ ; the number of favourable events is  $\binom{M}{x} \binom{N-M}{n-x}$  because we draw, independent of each other,  $x$  out of  $M$  balls and  $n - x$  out of  $N - M$  balls. Hence, the PMF of the hypergeometric distribution is

$$P(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \quad (8.14)$$

for  $x \in \{\max(0, n - (N - M)), \dots, \min(n, M)\}$ .

**Definition 8.1.8** A random variable  $X$  is said to follow a **hypergeometric** distribution with parameters  $n, M, N$ , i.e.  $X \sim H(n, M, N)$ , if its PMF is given by (8.14).

*Example 8.1.9* The German national lottery draws 6 out of 49 balls from a rotating bowl. Each ball is associated with a number between 1 and 49. A simple bet is to choose 6 numbers between 1 and 49. If 3 or more chosen numbers correspond to the numbers drawn in the lottery, then one wins a certain amount of money. What is the probability of choosing 4 correct numbers? We can utilize the hypergeometric distribution with  $x = 4$ ,  $M = 6$ ,  $N = 49$ , and  $n = 6$  to calculate such probabilities. The interpretation is that we “draw” (i.e. bet on) 4 out of the 6 winning balls and “draw” (i.e. bet on) another 2 out of the remaining 43 ( $49 - 6$ ) losing balls. In total, we draw 6 out of 49 balls. Calculating the number of the favourable combinations and all possible combinations leads to the application of the hypergeometric distribution as follows:

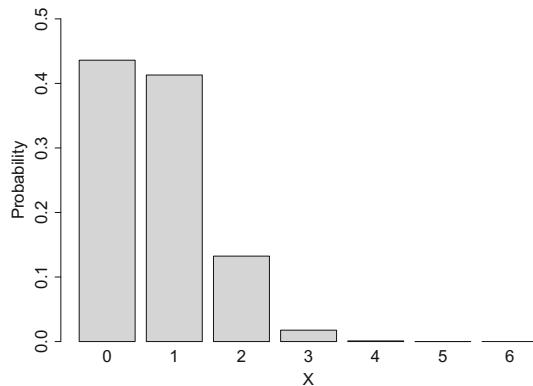
$$P(X = 4) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{6}{4} \binom{43}{2}}{\binom{49}{6}} \approx 0.001 \text{ (or } 0.1\%).$$

We would have obtained the same results using the `dhyper` command. Its arguments are `x`, `M`, `N`, `n`, and thus, we specify

`dhyper(4, 6, 43, 6)`



**Fig. 8.3** The  $H(6, 43, 6)$  distribution



The  $H(6, 43, 6)$  distribution is also visualized in Fig. 8.3. It is evident that the cumulative probability of choosing 2 or fewer correct numbers is greater than 0.9 (or 90 %), but it is very unlikely to have 3 or more numbers right. This may explain why the national lottery pays out money only for 3 or more correct numbers.

## 8.2 Standard Continuous Distributions

Now, we discuss some standard probability distributions of (absolute) continuous random variables. Characteristics of continuous random variables are that the number of possible outcomes is uncountably infinite and that they have a continuous distribution function  $F(x)$ . It follows that the point probabilities are zero, i.e.  $P(X = x) = 0$ . Further, we assume a unique density function  $f$  exists, such that  $F(x) = \int_{-\infty}^x f(t)dt$ .

### 8.2.1 Continuous Uniform Distribution

A continuous analogue to the discrete uniform distribution is the continuous uniform distribution on a closed interval in  $\mathbb{R}$ .

**Definition 8.2.1** A continuous random variable  $X$  is said to follow a (continuous) **uniform distribution** in the interval  $[a, b]$ , i.e.  $X \sim U(a, b)$ , if its probability density function (PDF) is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \quad (a < b) \\ 0 & \text{otherwise.} \end{cases}$$

The mean and variance of  $X \sim U(a, b)$  are

$$\mathrm{E}(X) = \frac{a+b}{2} \quad \text{and} \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12},$$

respectively.

*Example 8.2.1* Suppose a train arrives at a subway station regularly every 10 min. If a passenger arrives at the station without knowing the timetable, then the waiting time to catch the train is uniformly distributed with density

$$f(x) = \begin{cases} \frac{1}{10} & \text{if } 0 \leq x \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

The “average” waiting time is  $E(X) = (10 + 0)/2 = 5$  min. The probability of waiting for the train for less than 3 min is obviously 0.3 (or 30 %) which can be calculated in *R* using the `punif(x, a, b)` command (see also Appendix A.3):

`punif(3, 0, 10)`

R

### 8.2.2 Normal Distribution

The normal distribution is one of the most important distributions used in statistics. The name was given by Carl Friedrich Gauss (1777–1855), a German mathematician, astronomer, geodesist, and physicist who observed that measurements in geodesy and astronomy randomly deviate in a symmetric way from their true values. The normal distribution is therefore also often called a Gaussian distribution.

**Definition 8.2.2** A random variable  $X$  is said to follow a **normal distribution** with parameters  $\mu$  and  $\sigma^2$  if its PDF is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right); \quad -\infty < x < \infty, -\infty < \mu < \infty, \sigma^2 > 0. \quad (8.15)$$

We write  $X \sim N(\mu, \sigma^2)$ . The mean and variance of  $X$  are

$$E(X) = \mu; \quad \text{and} \quad \text{Var}(X) = \sigma^2,$$

respectively. If  $\mu = 0$  and  $\sigma^2 = 1$ , then  $X$  is said to follow a **standard normal distribution**,  $X \sim N(0, 1)$ . The PDF of a standard normal distribution is given by

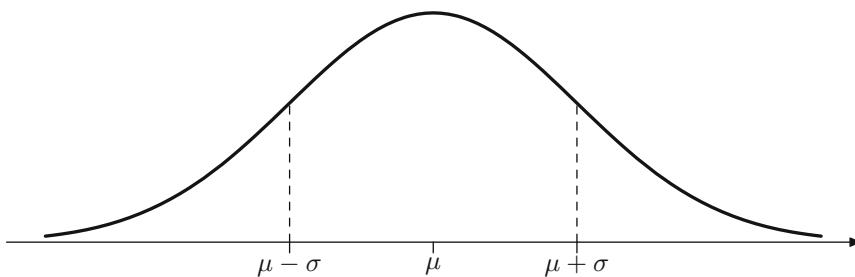
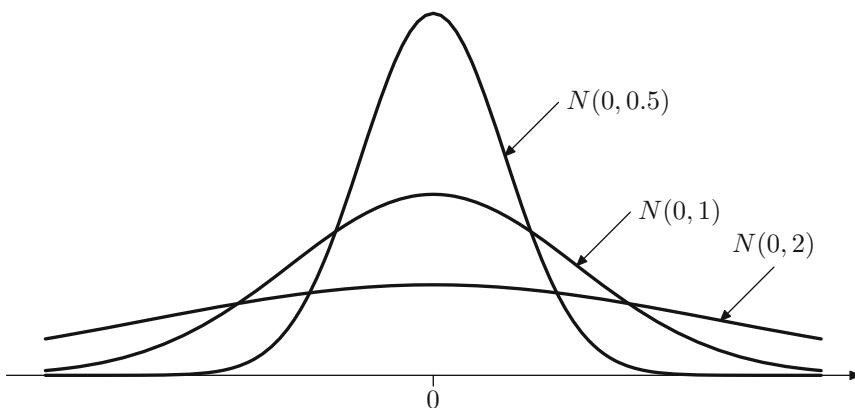
$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right); \quad -\infty < x < \infty.$$

The density of a normal distribution has its maximum (see Fig. 8.4) at  $x = \mu$ . The density is also symmetric around  $\mu$ . The inflection points of the density are at  $(\mu - \sigma)$  and  $(\mu + \sigma)$  (Fig. 8.4). A lower  $\sigma$  indicates a higher concentration around the mean  $\mu$ . A higher  $\sigma$  indicates a flatter density (Fig. 8.5).

The cumulative distribution function of  $X \sim N(\mu, \sigma^2)$  is

$$F(x) = \int_{-\infty}^x \phi(t) dt \quad (8.16)$$

which is often denoted as  $\Phi(x)$ . The value of  $\Phi(x)$  for various values of  $x$  can be obtained in *R* following the rules introduced in Appendix A.3. For example,

**Fig. 8.4** PDF of a normal distribution\***Fig. 8.5** PDF of  $N(0, 2)$ ,  $N(0, 1)$  and  $N(0, 0.5)$  distributions\*

```
pnorm(1.96, mean = 0, sd = 1)
```

R

calculates  $\Phi(1.96)$  as approximately 0.975. This means, for a standard normal distribution the probability  $P(X \leq 1.96) \approx 0.975$ .

*Remark 8.2.1* There is no explicit formula to solve the integral in Eq. (8.16). It has to be solved by numerical (or computational) methods. This is the reason why CDF tables are presented in almost all statistical textbooks, see Table C.1 in Appendix C.

*Example 8.2.2* An orange farmer sells his oranges in wooden boxes. The weights of the boxes vary and are assumed to be normally distributed with  $\mu = 15$  kg and  $\sigma^2 = \frac{9}{4}$  kg<sup>2</sup>. The farmer wants to avoid customers being unsatisfied because the boxes are too low in weight. He therefore asks the following question: What is the probability that a box with a weight of less than 13 kg is sold? Using the `pnorm(x, μ, σ)` command in R, we get

```
pnorm(13,15,sqrt(9/4))
[1] 0.09121122
```



To calculate the probability in Example 8.2.2 manually, we first have to introduce some theoretical results.

### Calculation rules for normal random variables.

Let  $X \sim N(\mu, \sigma^2)$ . Using the transformation

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad (8.17)$$

every normally distributed random variable can be transformed into a *standard* normal random variable. We call this transformation the  $Z$ -transformation. We can use this transformation to derive convenient calculation rules. The probability for  $X \leq b$  is

$$P(X \leq b) = P\left(\frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = P\left(Z \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right). \quad (8.18)$$

Consequently, the probability for  $X > a$  is

$$P(X > a) = 1 - P(X \leq a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right). \quad (8.19)$$

The probability that  $X$  realizes a value in the interval  $[a, b]$  is

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \quad (8.20)$$

Because of the symmetry of the probability density function  $\phi(x)$  around its mean 0, the following equation holds for the distribution function  $\Phi(x)$  of a standard normal random variable for any value  $a$ :

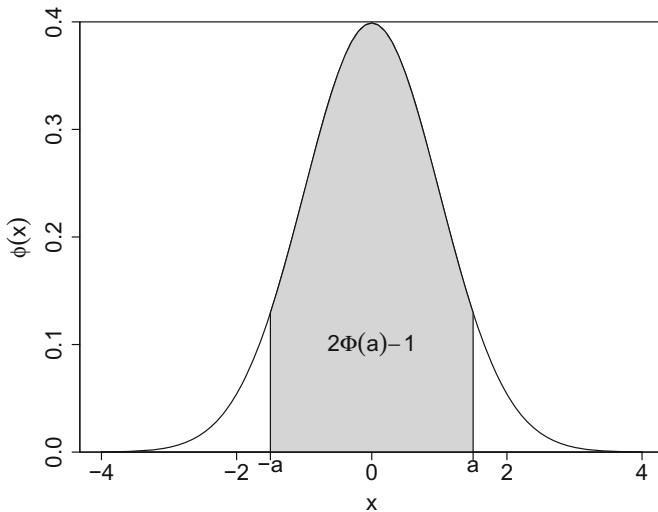
$$\Phi(-a) = 1 - \Phi(a). \quad (8.21)$$

It follows that  $P(-a < Z < a) = 2 \cdot \Phi(a) - 1$ , see also Fig. 8.6.

*Example 8.2.3* Recall Example 8.2.2 where a farmer sold his oranges. He was interested in  $P(X \leq 13)$  for  $X \sim N(15, 9/4)$ . Using (8.17), we get

$$\begin{aligned} P(X \leq 13) &= \Phi\left(\frac{13 - 15}{\sqrt{\frac{9}{4}}}\right) \\ &= \Phi\left(-\frac{4}{3}\right) = 1 - \Phi\left(\frac{4}{3}\right) \approx 0.091 \text{ (or } 9.1\%). \end{aligned}$$

To obtain  $\Phi(4/3) \approx 90.9\%$ , we could either use R (`pnorm(4/3)`) or use Table C.1.



**Fig. 8.6** Distribution function of the standard normal distribution

### Distribution of the Arithmetic Mean.

Assume that  $X \sim N(\mu, \sigma^2)$ . Consider a random sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  of independent and identically distributed random variables  $X_i$  with  $X_i \sim N(\mu, \sigma^2)$ . Then, the arithmetic mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  follows a normal distribution with mean

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu$$

and variance

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n} \quad (8.22)$$

where  $\text{Cov}(X_i, X_j) = 0$  for  $i \neq j$ . In summary, we get

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

*Remark 8.2.2* In fact, in Eq. (8.22), we have used the fact that the sum of normal random variables also follows a normal distribution, i.e.

$$(X_1 + X_2) \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

This result can be generalized to  $n$  (not necessarily identically distributed but independent) normal random variables. In fact, it holds that if  $X_1, X_2, \dots, X_n$  are independent normal variables with means  $\mu_1, \mu_2, \dots, \mu_n$  and variances  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ , then for any real numbers  $a_1, a_2, \dots, a_n$ , it holds that

$$(a_1 X_1 + a_2 X_2 + \dots + a_n X_n) \sim N\left(a_1 \mu_1 + a_2 \mu_2 + \dots + a_n \mu_n, a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2\right).$$

In general, it cannot be taken for granted that the sum of two random variables follows the same distribution as the two variables themselves. As an example, consider the sum of two independent uniform distributions with  $X_1 \sim U[0, 10]$  and  $X_2 \sim U[20, 30]$ . It holds that  $E(X_1 + X_2) = E(X_1) + E(X_2)$  and  $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$ , but  $X_1 + X_2$  is obviously not uniformly distributed.

### 8.2.3 Exponential Distribution

The exponential distribution is useful in many situations, for example when one is interested in the waiting time, or lifetime, until an event of interest occurs. If we assume that the future lifetime is independent of the lifetime that has already taken place (i.e. no “ageing” process is working), the waiting times can be considered to be exponentially distributed.

**Definition 8.2.3** A random variable  $X$  is said to follow an exponential distribution with parameter  $\lambda > 0$  if its PDF is given by

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (8.23)$$

We write  $X \sim Exp(\lambda)$ . The mean and variance of an exponentially distributed random variable  $X$  are

$$E(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2},$$

respectively. The CDF of the exponential distribution is given as

$$F(x) = \begin{cases} 1 - \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (8.24)$$

Note, that  $P(X > x) = 1 - F(x) = \exp(-\lambda x)$  ( $x \geq 0$ ). An interesting property of the exponential distribution is its **memorylessness**: if time  $t$  has already been reached, the probability of reaching a time greater than  $t + \Delta$  does not depend on  $t$ . This can be written as

$$P(X > t + \Delta | X > t) = P(X > \Delta) \quad t, \Delta > 0.$$

The result can be derived using basic probability rules as follows:

$$\begin{aligned} P(X > t + \Delta | X > t) &= \frac{P(X > t + \Delta \text{ and } X > t)}{P(X > t)} = \frac{P(X > t + \Delta)}{P(X > t)} \\ &= \frac{\exp[-\lambda(t + \Delta)]}{\exp[-\lambda t]} = \exp[-\lambda \Delta] \\ &= 1 - F(\Delta) = P(X > \Delta). \end{aligned}$$

For example, suppose someone stands in a supermarket queue for  $t$  minutes. Say the person forgot to buy milk, so she leaves the queue, gets the milk, and stands in the queue again. If we use the exponential distribution to model the waiting time, we say that it does not matter what time it is: the random variable “waiting time from standing in the queue until paying the bill” is not influenced by how much

time has elapsed already; it does not matter if we queued before or not. Please note that the memorylessness property is shared by the geometric and the exponential distributions.

There is also a relationship between the Poisson and the exponential distribution:

**Theorem 8.2.1** *The number of events  $Y$  occurring within a continuum of time is Poisson distributed with parameter  $\lambda$  if and only if the time between two events is exponentially distributed with parameter  $\lambda$ .*

The continuum of time depends on the problem at hand. It may be a second, a minute, 3 months, a year, or any other time period.

*Example 8.2.4* Let  $Y$  be the random variable which counts the “number of accesses per second for a search engine”. Assume that  $Y$  is Poisson distributed with parameter  $\lambda = 10$  ( $E(Y) = 10$ ,  $\text{Var}(Y) = 10$ ). The random variable  $X$ , “waiting time until the next access”, is then exponentially distributed with parameter  $\lambda = 10$ . We therefore get

$$E(X) = \frac{1}{10}, \quad \text{Var}(X) = \frac{1}{10^2}.$$

In this example, the continuum is 1 s. The expected number of accesses per second is therefore  $E(Y) = 10$ , and the expected waiting time between two accesses is  $E(X) = 1/10$  s. The probability of experiencing a waiting time of less than 0.1 s is

$$F(0.1) = 1 - \exp(-\lambda x) = 1 - \exp(-10 \cdot 0.1) \approx 0.63.$$

In R, we can obtain the same result as

```
pexp(0.1, 10)
[1] 0.6321206
```



## 8.3 Sampling Distributions

All the distributions introduced in this chapter up to now are motivated by practical applications. However, there are theoretical distributions which play an important role in the construction and development of various statistical tools such as those introduced in Chaps. 9–11. We call these distributions “sampling distributions”. Now, we discuss the  $\chi^2$ -,  $t$ -, and  $F$ -distributions.

### 8.3.1 $\chi^2$ -Distribution

**Definition 8.3.1** Let  $Z_1, Z_2 \dots, Z_n$  be  $n$  independent and identically  $N(0, 1)$ -distributed random variables. The sum of their squares,  $\sum_{i=1}^n Z_i^2$ , is then  $\chi^2$ -distributed with  $n$  degrees of freedom and is denoted as  $\chi_n^2$ . The PDF of the  $\chi^2$ -distribution is given in Eq. (C.7) in Appendix C.3.

The  $\chi^2$ -distribution is not symmetric. A  $\chi^2$ -distributed random variable can only realize values greater than or equal to zero. Figure 8.7a shows the  $\chi^2_1$ -,  $\chi^2_2$ -, and  $\chi^2_5$ -distributions. It can be seen that the “degrees of freedom” specify the shape of the distribution. Their interpretation and meaning will nevertheless become clearer in the following chapters. The quantiles of the CDF of different  $\chi^2$ -distributions can be obtained in *R* using the `qchisq(p, df)` command. They are also listed in Table C.3 for different values of  $n$ .

**Theorem 8.3.1** Consider two independent random variables which are  $\chi^2_m$ - and  $\chi^2_n$ -distributed, respectively. The sum of these two random variables is  $\chi^2_{n+m}$ -distributed.

An important example of a  $\chi^2$ -distributed random variable is the sample variance ( $S_X^2$ ) of an i.i.d. sample of size  $n$  from a normally distributed population, i.e.

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}. \quad (8.25)$$

### 8.3.2 *t*-Distribution

**Definition 8.3.2** Let  $X$  and  $Y$  be two independent random variables where  $X \sim N(0, 1)$  and  $Y \sim \chi^2_n$ . The ratio

$$\frac{X}{\sqrt{Y/n}} \sim t_n$$

follows a ***t*-distribution** (Student’s *t*-distribution) with  $n$  degrees of freedom. The PDF of the *t*-distribution is given in Eq. (C.8) in Appendix C.3.

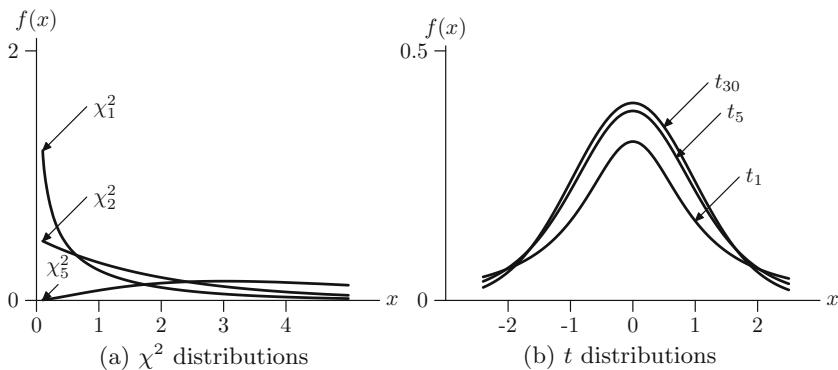
Figure 8.7b visualizes the  $t_1$ -,  $t_5$ -, and  $t_{30}$ -distributions. The quantiles of different *t*-distributions can be obtained in *R* using the `qt(p, df)` command. They are also listed in Table C.2 for different values of  $n$ .

An application of the *t*-distribution is the following: if we draw a sample of size  $n$  from a normal population  $N(\mu, \sigma^2)$  and calculate the arithmetic mean  $\bar{X}$  and the sample variance  $S_X^2$ , then the following theorem holds:

**Theorem 8.3.2** (Student’s theorem) Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  with  $X_i \stackrel{iid.}{\sim} N(\mu, \sigma^2)$ . The ratio

$$\frac{(\bar{X} - \mu)\sqrt{n}}{S_X} = \frac{(\bar{X} - \mu)\sqrt{n}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}} \sim t_{n-1} \quad (8.26)$$

is then *t*-distributed with  $n - 1$  degrees of freedom.



**Fig. 8.7** Probability density functions of  $\chi^2$  and  $t$  distributions\*

### 8.3.3 F-Distribution

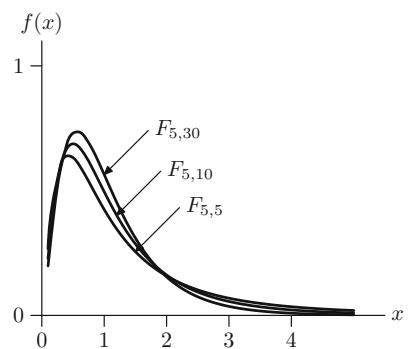
**Definition 8.3.3** Let  $X$  and  $Y$  be independent  $\chi_m^2$  and  $\chi_n^2$ -distributed random variables, then the distribution of the ratio

$$\frac{X/m}{Y/n} \sim F_{m,n} \quad (8.27)$$

follows the **Fisher F-distribution** with  $(m, n)$  degrees of freedom. The PDF of the  $F$ -distribution is given in Eq. (C.9) in Appendix C.3.

If  $X$  is a  $\chi_1^2$ -distributed random variable, then the ratio (8.27) is  $F_{1,n}$ -distributed. The square root of this ratio is  $t_n$ -distributed since the square root of a  $\chi_1^2$ -distributed random variable is  $N(0, 1)$ -distributed. If  $W$  is  $F$ -distributed,  $F_{m,n}$ , then  $1/W$  is  $F_{n,m}$ -distributed. Figure 8.8 visualizes the  $F_{5,5}$ ,  $F_{5,10}$  and  $F_{5,30}$  distributions. The

**Fig. 8.8** Probability density functions for different  $F$ -distributions\*



quantiles of different  $F$ -distributions can be obtained in  $R$  using the `qf(p, df1, df2)` command.

One application of the  $F$ -distribution relates to the ratio of two sample variances of two independent samples of size  $m$  and  $n$ , where each sample is an i.i.d. sample from a normal population, i.e.  $N(\mu_X, \sigma^2)$  and  $N(\mu_Y, \sigma^2)$ . For the sample variances  $S_X^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2$  and  $S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$  from the two populations, the ratio

$$\frac{S_X^2}{S_Y^2} \sim F_{m-1, n-1}$$

is  $F$ -distributed with  $(m - 1)$  degrees of freedom in the numerator and  $(n - 1)$  degrees of freedom in the denominator.

## 8.4 Key Points and Further Issues

### Note:

- ✓ Examples of different distributions are:

Distribution	Example
Uniform	Rolling a die (discrete) Waiting for a train (continuous)
Bernoulli	Any binary variable such as gender
Binomial	Number of “heads” when tossing a coin $n$ times
Poisson	Number of particles emitted by a radioactive source entering a small area in a given time interval
Multinomial	Categorical variables such as “party voted for”
Geometric	Number of raffle tickets until first ticket wins
Hypergeometric	National lotteries; Fisher’s test, see p. 428
Normal	Height or weight of women (men)
Exponential	Survival time of a PC
$\chi^2$	Sample variance; $\chi^2$ tests, see p. 235 ff
$t$	Confidence interval for the mean, see p. 197
$F$	Tests in the linear model, see p. 272

**Note:**

- ✓ One can use  $R$  to determine values of densities (PDF/PMF), cumulative probability distribution functions (CDF), quantiles of the CDF, and random numbers:

First letter	Function	Further letters	Example
d	Density	distribution name	dnorm
p	Probability	distribution name	pnorm
q	Quantiles	distribution name	qnorm
r	Random number	distribution name	rnorm

We encourage the use of  $R$  to obtain quantiles of sampling distributions, but Tables C.1–C.3 also list some of them.

- ✓ In this chapter, we assumed the parameters such as  $\mu$ ,  $\sigma$ ,  $\lambda$ , and others to be known. In Chap. 9, we will propose how to estimate these parameters from the data. In Chap. 10, we test statistical hypotheses about these parameters.
- ✓ For  $n$  i.i.d. random variables  $X_1, X_2, \dots, X_n$ , the arithmetic mean  $\bar{X}$  converges to a  $N(\mu, \sigma^2/n)$  distribution as  $n$  tends to infinity. See Appendix C.3 as well as Exercise 8.11 for the Theorem of Large Numbers and the Central Limit Theorem, respectively.

---

## 8.5 Exercises

*Exercise 8.1* A company producing cereals offers a toy in every sixth cereal package in celebration of their 50th anniversary. A father immediately buys 20 packages.

- What is the probability of finding 4 toys in the 20 packages?
- What is the probability of finding no toy at all?
- The packages contain three toys. What is the probability that among the 5 packages that are given to the family's youngest daughter, she finds two toys?

*Exercise 8.2* A study on breeding birds collects information such as the length of their eggs (in mm). Assume that the length is normally distributed with  $\mu = 42.1$  mm and  $\sigma^2 = 20.8^2$ . What is the probability of

- finding an egg with a length greater than 50 mm?
- finding an egg between 30 and 40 mm in length?

Calculate the results both manually and by using  $R$ .

*Exercise 8.3* A dodecahedron is a die with 12 sides. Suppose the numbers on the die are 1–12. Consider the random variable  $X$  which describes which number is shown after rolling the die once. What is the distribution of  $X$ ? Determine  $E(X)$  and  $\text{Var}(X)$ .

*Exercise 8.4* Felix states that he is able to distinguish a freshly ground coffee blend from an ordinary supermarket coffee. One of his friends asks him to taste 10 cups of coffee and tell him which coffee he has tasted. Suppose that Felix has actually no clue about coffee and simply guesses the brand. What is the probability of at least 8 correct guesses?

*Exercise 8.5* An advertising board is illuminated by several hundred bulbs. Some of the bulbs are fused or smashed regularly. If there are more than 5 fused bulbs on a day, the owner of the board replaces them, otherwise not. Consider the following data collected over a month which captures the number of days ( $n_i$ ) on which  $i$  bulbs were broken:

Fused bulbs	0	1	2	3	4	5
$n_i$	6	8	8	5	2	1

- (a) Suggest an appropriate distribution for  $X$ : “number of broken bulbs per day”.
- (b) What is the average number of broken bulbs per day? What is the variance?
- (c) Determine the probabilities  $P(X = x)$  using the distribution you chose in (a) and using the average number of broken bulbs you calculated in (b). Compare the probabilities with the proportions obtained from the data.
- (d) Calculate the probability that at least 6 bulbs are fused, which means they need to be replaced.
- (e) Consider the random variable  $Y$ : “time until next bulb breaks”. What is the distribution of  $Y$ ?
- (f) Calculate and interpret  $E(Y)$ .

*Exercise 8.6* Marco’s company organizes a raffle at an end-of-year function. There are 4000 raffle tickets to be sold, of which 500 win a prize. The price of each ticket is €1.50. The value of the prizes, which are mostly electrical appliances produced by the company, varies between €80 and €250, with an average value of €142.

- (a) Marco wants to have a 99 % guarantee of receiving three prizes. How much money does he need to spend? Use  $R$  to solve the question.
- (b) Use  $R$  to plot the function which describes the relationship between the number of tickets bought and the probability of winning at least three prizes.
- (c) Given the value of the prizes and the costs of the tickets, is it worth taking part in the raffle?

*Exercise 8.7* A country has a ratio between male and female births of 1.05 which means that 51.22 % of babies born are male.

- (a) What is the probability for a mother that the first girl is born during the first three births?
- (b) What is the probability of getting 2 girls among 4 babies?

*Exercise 8.8* A fisherman catches, on average, three fish in an hour. Let  $Y$  be a random variable denoting the number of fish caught in one hour and let  $X$  be the time interval between catching two fishes. We assume that  $X$  follows an exponential distribution.

- (a) What is the distribution of  $Y$ ?
- (b) Determine  $E(Y)$  and  $E(X)$ .
- (c) Calculate  $P(Y = 5)$  and  $P(Y < 1)$ .

*Exercise 8.9* A restaurant sells three different types of dessert: chocolate, brownies, yogurt with seasonal fruits, and lemon tart. Years of experience have shown that the probabilities with which the desserts are chosen are 0.2, 0.3, and 0.5, respectively.

- (a) What is the probability that out of 5 guests, 2 guests choose brownies, 1 guest chooses yogurt, and the remaining 2 guests choose lemon tart?
- (b) Suppose two out of the five guests are known to always choose lemon tart. What is the probability of the others choosing lemon tart as well?
- (c) Determine the expectation and variance assuming a group of 20 guests.

*Exercise 8.10* A reinsurance company works on a premium policy for natural disasters. Based on experience, it is known that  $W$  = “number of natural disasters from October to March” (winter) is Poisson distributed with  $\lambda_W = 4$ . Similarly, the random variable  $S$  = “number of natural disasters from April to September” (summer) is Poisson distributed with  $\lambda_S = 3$ . Determine the probability that there is at least 1 disaster during both summer and winter based on the assumption that the two random variables are independent.

*Exercise 8.11* Read Appendix C.3 to learn about the Theorem of Large Numbers and the Central Limit Theorem.

- (a) Draw 1000 realizations from a standard normal distribution using  $R$  and calculate the arithmetic mean. Repeat this process 1000 times. Evaluate the distribution of the arithmetic mean by drawing a kernel density plot and by calculating the mean and variance of it.

- (b) Repeat the procedure in (a) with an exponential distribution with  $\lambda = 1$ . Interpret your findings in the light of the Central Limit Theorem.
- (c) Repeat the procedure in (b) using 10,000 rather than 1000 realizations. How do the results change and why?

→ Solutions to all exercises in this chapter can be found on p. [375](#)

\*Source Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg

---

**Part III**

**Inductive Statistics**

---

## 9.1 Introduction

The first four chapters of this book illustrated how one can summarize a data set both numerically and graphically. The validity of interpretations made from such a descriptive analysis is valid only for the data set under consideration and cannot necessarily be generalized to other data. However, it is desirable to make conclusions about the entire population of interest and not only about the sample data. In this chapter, we describe the framework of **statistical inference** which allows us to infer from the sample data about the population of interest—at a given, prespecified uncertainty level—and knowledge about the random process generating the data.

Consider an example where the objective is to forecast an election outcome. This requires us to determine the proportion of votes that each of the  $k$  participating parties is going to receive, i.e. to calculate or estimate  $p_1, p_2, \dots, p_k$ . If it is possible to ask every voter about their party preference, then one can simply calculate the proportions  $p_1, p_2, \dots, p_k$  for each party. However, it is logically impossible to ask all eligible voters (which form the population in this case) about their preferred party. It seems more realistic to ask only a small fraction of voters and infer from their responses to the responses of the whole population. It is evident that there might be differences in responses between the sample and the population—but the more voters are asked, the closer we are to the population’s preference, i.e. the higher the precision of our estimates for  $p_1, p_2, \dots, p_k$  (the meaning of “precision” will become clearer later in this chapter). Also, it is intuitively clear that the sample must be a representative sample of the voters’ population to avoid any discrepancy or bias in the forecasting. When we speak of a representative sample, we mean that all the characteristics present in the population are contained in the sample too. There are many ways to get representative random samples. In fact, there is a branch of statistics, called sampling theory, which studies this subject [see, e.g. Groves et al. (2009) or Kauermann and Küchenhoff (2011) for more details]. A simple random sample is one where each voter has an equal probability of being selected in the sample and

each voter is independently chosen from the same population. In the following, we will assume that all samples are simple random samples. To further formalize the election forecast problem, assume that we are interested in the true proportions which each party receives on the election day. It is practically impossible to make a perfect prediction of these proportions because there are too many voters to interview, and moreover, a voter may possibly make their final decision possibly only when casting the vote and not before. The voter may change his/her opinion at any moment and may differ from what he/she claimed earlier. In statistics, we call these true proportions *parameters of the population*. The task is then to estimate these parameters on the basis of a sample. In the election example, the intuitive estimates for the proportions in the population are the proportions in the sample and we call them *sample estimates*. How to find good and precise estimates are some of the challenges that are addressed by the concept of *statistical inference*. Now, it is possible to describe the election forecast problem in a statistical and operational framework: estimate the parameters of a population by calculating the sample estimates. An important property of every good statistical inference procedure is that it provides not only estimates for the population parameters but also information about the precision of these estimates.

Consider another example in which we would like to study the distribution of weight of children in different age categories and get an understanding of the “normal” weight. Again, it is not possible to measure the weight of all the children of a specific age in the entire population of children in a particular country. Instead, we draw a random sample and use methods of statistical inference to estimate the weight of children in each age group. More specifically, we have several populations in this problem. We could consider all boys of a specific age and all girls of a specific age as two different populations. For example, all 3-year-old boys will form one possible population. Then, a random sample is drawn from this population. It is reasonable to assume that the distribution of the weight of  $k$ -year-old boys follows a normal distribution with some unknown parameters  $\mu_{kb}$  and  $\sigma_{kb}^2$ . Similarly, another population of  $k$ -year-old girls is assumed to follow a normal distribution with some unknown parameters  $\mu_{kg}$  and  $\sigma_{kg}^2$ . The indices  $kb$  and  $kg$  are used to emphasize that the parameters may vary by age and gender. The task is now to calculate the estimates of the unknown parameters (in the population) of the normal distributions from the samples. Using quantiles, a range of “normal” weights can then be specified, e.g. the interval from the 1 % quantile to the 99 % quantile of the estimated normal distribution or, alternatively, all weights which are not more than twice the standard deviation away from the mean. Children with weights outside this interval may be categorized as underweight or overweight. Note that we make a specific assumption for the distribution class; i.e. we assume a normal distribution for the weights and estimate its parameters. We call this a **parametric** estimation problem because it is based on distributional assumptions. Otherwise, if no distributional assumptions are made, we speak of a **nonparametric** estimation problem.

## 9.2 Properties of Point Estimators

As we discussed in the introduction, the primary goal in statistical inference is to find a good estimate of (a) population parameter(s). The parameters are associated with the probability distribution which is believed to characterize the population; e.g.  $\mu$  and  $\sigma^2$  are the parameters in a normal distribution  $N(\mu, \sigma^2)$ . If these parameters are known, then one can characterize the entire population. In practice, these parameters are unknown, so the objective is to estimate them. One can attempt to obtain them based on a function of the sample values. But what does this function look like; and if there is more than one such function, then which is the best one? What is the best approach to estimate the population parameters on the basis of a given sample of data? The answer is given by various statistical concepts such as bias, variability, consistency, efficiency, sufficiency, and completeness of the estimates. We are going to introduce them now.

Assume  $x = (x_1, x_2, \dots, x_n)$  are the observations of a random sample from a population of interest. The random sample represents the realized values of a random variable  $X$ . It can be said that  $x_1, x_2, \dots, x_n$  are the  $n$  observations collected on the random variable  $X$ . Any function of random variables is called a **statistic**. For example,  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $\max(X_1, X_2, \dots, X_n)$  etc. are functions of  $X_1, X_2, \dots, X_n$ , so they are a statistic. It follows that a statistic is also a random variable. Consider a statistic  $T(X)$  which is used to estimate a population parameter  $\theta$  (which may be either a scalar or a vector). We say  $T(X)$  is an **estimator** of  $\theta$ . To indicate that we estimate  $\theta$  using  $T(X)$ , we use the “hat” (^) symbol, i.e. we write  $\hat{\theta} = T(X)$ . When  $T$  is calculated from the sample values  $x_1, x_2, \dots, x_n$ , we write  $T(x)$  and call it an **estimate** of  $\theta$ . It becomes clear that  $T(X)$  is a random variable but  $T(x)$  is its observed value (dependent on the actual sample). For example,  $T(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  is an estimator and a statistic, but  $T(x) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  is its estimated value from the realized sample values  $x_1, x_2, \dots, x_n$ . Since the sample values are realizations from a random variable, each sample leads to a different value of the estimate of the population parameter. The population parameter is assumed to be a fixed value. Parameters can also be assumed to be random, for example in Bayesian statistics, but this is beyond the scope of this book.

### 9.2.1 Unbiasedness and Efficiency

**Definition 9.2.1** An estimator  $T(X)$  is called an *unbiased* estimator of  $\theta$  if

$$E_\theta(T(X)) = \theta . \quad (9.1)$$

The index  $\theta$  denotes that the expectation is calculated with respect to the distribution whose parameter is  $\theta$ .

The bias of an estimator  $T(X)$  is defined as

$$\text{Bias}_\theta(T(X)) = \mathbb{E}_\theta(T(X)) - \theta . \quad (9.2)$$

It follows that an estimator is said to be unbiased if its bias is zero.

**Definition 9.2.2** The variance of  $T(X)$  is defined as

$$\text{Var}_\theta(T(X)) = \mathbb{E} \{ [T(X) - \mathbb{E}(T(X))]^2 \} . \quad (9.3)$$

Both bias and variance are measures which characterize the properties of an estimator. In statistical theory, we search for “good” estimators in the sense that the bias and the variance are as small as possible and therefore the accuracy is as high as possible. Readers interested in a practical example may consult Examples 9.2.1 and 9.2.2, or the explanations for Fig. 9.1.

It turns out that we cannot minimize both measures simultaneously as there is always a so-called bias–variance tradeoff. A measure which combines bias and variance into one measure is the mean squared error.

**Definition 9.2.3** The mean squared error (MSE) of  $T(X)$  is defined as

$$\text{MSE}_\theta(T(X)) = \mathbb{E} \{ [T(X) - \theta]^2 \} . \quad (9.4)$$

The expression (9.4) can be partitioned into two parts: the variance and the squared bias, i.e.

$$\text{MSE}_\theta(T(X)) = \text{Var}_\theta(T(X)) + [\text{Bias}_\theta(T(X))]^2 . \quad (9.5)$$

This can be proven as follows:

$$\begin{aligned} \text{MSE}_\theta(T(X)) &= \mathbb{E}[T(X) - \theta]^2 \\ &= \mathbb{E}[(T(X) - \mathbb{E}_\theta(T(X)) + (\mathbb{E}_\theta(T(X)) - \theta))^2] \\ &= \mathbb{E}[T(X) - \mathbb{E}_\theta(T(X))]^2 + [\mathbb{E}_\theta(T(X)) - \theta]^2 \\ &= \text{Var}_\theta(T(X)) + [\text{Bias}_\theta(T(X))]^2 . \end{aligned}$$

Note that the calculation is based on the result that the cross product term is zero. The mean squared error can be used to compare different biased estimators.

**Definition 9.2.4** An estimator  $T_1(X)$  is said to be MSE-better than another estimator  $T_2(X)$  for estimating  $\theta$  if

$$\text{MSE}_\theta(T_1(X)) < \text{MSE}_\theta(T_2(X)) ,$$

where  $\theta \in \Theta$  and  $\Theta$  is the parameter space, i.e. the set of all possible values of  $\theta$ . Often,  $\Theta$  is  $\mathbb{R}$  or all positive real values  $\mathbb{R}_+$ . For example, for a normal distribution,  $N(\mu, \sigma^2)$ ,  $\mu$  can be any real value and  $\sigma^2$  has to be a number greater than zero.

Unfortunately, we cannot find an MSE-optimal estimator in the sense that an estimator is MSE-better than all other possible estimators for all possible values of  $\theta$ . This becomes clear if we define the constant estimator  $T(x) = c$  (independent of the actual sample): if  $\theta = c$ , i.e. if the constant value equals the true population parameter we want to estimate, then the MSE of this constant estimator is zero (but it will be greater than zero for all other values of  $\theta$ , and the bias increases more as we move  $c$  far away from the true  $\theta$ ). Usually, we can only find estimators which are locally best (in a certain subset of  $\Theta$ ). This is why classical statistical inference restricts the search for best estimators to the class of unbiased estimators. For unbiased estimators, the MSE is equal to the variance of an estimator. In this context, the following definition is used for comparing two (unbiased) estimators.

**Definition 9.2.5** An unbiased estimator  $T_1(X)$  is said to be more efficient than another unbiased estimator  $T_2(X)$  for estimating  $\theta$  if

$$\text{Var}_\theta(T_1(X)) \leq \text{Var}_\theta(T_2(X)), \quad \forall \theta \in \Theta,$$

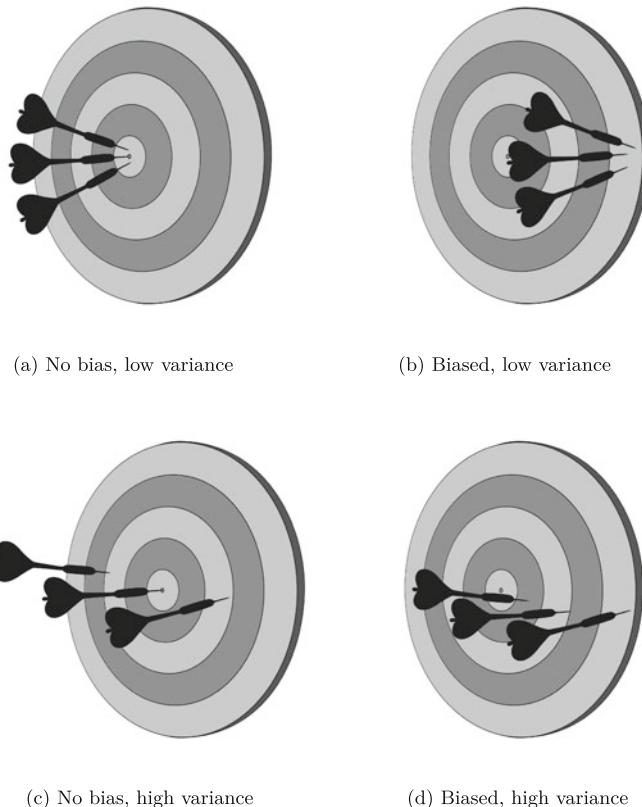
and

$$\text{Var}_\theta(T_1(X)) < \text{Var}_\theta(T_2(X))$$

for at least one  $\theta \in \Theta$ . It turns out that restricting our search of best estimators to unbiased estimators is sometimes a successful strategy; i.e. for many problems, a best or most efficient estimate can be found. If such an estimator exists, it is said to be UMVU (uniformly minimum variance unbiased). Uniformly means that it has the lowest variance among all other unbiased estimators for estimating the population parameter(s)  $\theta$ .

Consider the illustration in Fig. 9.1 to better understand the introduced concepts. Suppose we throw three darts at a target and the goal is to hit the centre of the target, i.e. the innermost circle of the dart board. The centre represents the population parameter  $\theta$ . The three darts play the role of three estimates  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$  (based on different realizations of the sample) of the population parameter  $\theta$ . Four possible situations are illustrated in Fig. 9.1. For example, in Fig. 9.1b, we illustrate the case of an estimator which is biased but has low variance: all three darts are “far” away from the centre of the target, but they are “close” together. If we look at Fig. 9.1a, c, we see that all three darts are symmetrically grouped around the centre of the target, meaning that there is no bias; however, in Fig. 9.1a there is much higher precision than in Fig. 9.1c. It is obvious that Fig. 9.1a presents an ideal situation: an estimator which is unbiased and has minimum variance.

**Theorem 9.2.1** Let  $X = (X_1, X_2, \dots, X_n)$  be an i.i.d. (random) sample of a random variable  $X$  with population mean  $E(X_i) = \mu$  and population variance  $\text{Var}(X_i) = \sigma^2$ , for all  $i = 1, 2, \dots, n$ . Then the arithmetic mean  $\bar{X} = \sum_{i=1}^n X_i$  is an unbiased estimator of  $\mu$  and the sample variance  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  is an unbiased estimator of  $\sigma^2$ .



**Fig. 9.1** Illustration of bias and variance

Note that the theorem holds, in general, for i.i.d. samples, irrespective of the choice of the distribution of the  $X_i$ 's. Note again that we are looking at the situation *before* we have any observations on  $X$ . Therefore, we again use capital letters to denote that the  $X_i$ 's are random variables which are not known beforehand (i.e. before we actually record the observations on our selected sampling units).

*Remark 9.2.1* The empirical variance  $\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  is a biased estimate of  $\sigma^2$  and its bias is  $-\frac{1}{n}\sigma^2$ .

*Example 9.2.1* Let  $X_1, X_2 \dots, X_n$  be identically and independently distributed variables whose population mean is  $\mu$  and population variance is  $\sigma^2$ . Then  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  is an unbiased estimator of  $\mu$ . This can be shown as follows:

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \stackrel{(7.29)}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu.$$

The variance of  $\bar{X}$  can be calculated as follows:

$$\begin{aligned}\text{Var}(\bar{X}) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i), \quad [\text{Cov}(X_i, X_j) = 0 \text{ using independence of } X_i\text{'s}] \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

We conclude that  $\bar{X}$  is an unbiased estimator of  $\mu$  and its variance is  $\frac{\sigma^2}{n}$  irrespective of the choice of the distribution of  $X$ . We have learned about the distribution of  $\bar{X}$  already in Chap. 8, see also Appendix C.3 for the Theorem of Large Numbers and the Central Limit Theorem; however, we would like to highlight the property of “unbiasedness” in the current context.

Now, we consider another example to illustrate that estimators may not always be unbiased but may have the same variance.

*Example 9.2.2* Let  $X_1, X_2, \dots, X_n$  be identically and independently distributed variables whose population mean is  $\mu$  and population variance is  $\sigma^2$ . Then  $\tilde{X} = \bar{X} + 1 = \frac{1}{n} \sum_{i=1}^n (X_i + 1)$  is a biased estimator of  $\mu$ . This can be shown as follows:

$$\begin{aligned}\text{E}(\tilde{X}) &\stackrel{(7.31)}{=} \text{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) + \text{E}\left(\frac{1}{n} \sum_{i=1}^n 1\right) \\ &\stackrel{(7.29)}{=} \frac{1}{n} \sum_{i=1}^n \text{E}(X_i) + \frac{1}{n} \cdot n = \frac{1}{n} \sum_{i=1}^n \mu + 1 \\ &= \mu + 1 \neq \mu.\end{aligned}$$

However, the variance of  $\tilde{X}$  is

$$\text{Var}(\tilde{X}) = \text{Var}(\bar{X} + 1) \stackrel{(7.34)}{=} \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

If we compare the two estimators  $\tilde{X} = \frac{1}{n} \sum_{i=1}^n (X_i + 1)$  and  $\bar{X} = \frac{1}{n} \sum_{i=1}^n (X_i)$ , we see that both have the same variance but the former ( $\tilde{X}$ ) is biased. The efficiency of both estimators is thus the same. It further follows that the mean squared error of  $\bar{X}$  is smaller than the mean squared error of  $\tilde{X}$  because the MSE consists of the sum of the variance and the squared bias. Therefore  $\bar{X}$  is MSE-better than  $\tilde{X}$ . The comparison of bias, variance and MSE tells us that we should prefer  $\bar{X}$  over  $\tilde{X}$  when estimating the population mean. This is intuitive, but the argument we make is a purely statistical one.

Theorem 9.2.1 contains the following special cases:

- The sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  based on an i.i.d. random sample  $X_1, X_2, \dots, X_n$  from a normally distributed population  $N(\mu, \sigma^2)$  is an unbiased point estimator of  $\mu$ .

- The sample variance  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  based on an i.i.d. random sample  $X_1, X_2, \dots, X_n$  from a normally distributed population  $N(\mu, \sigma^2)$  is an unbiased point estimator of  $\sigma^2$ . The sample variance  $\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  is a biased estimator for  $\sigma^2$ , but it is asymptotically unbiased in the sense that its bias tends to zero as the sample size  $n$  tends to infinity.
- The sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  based on an i.i.d. random sample  $X_1, X_2, \dots, X_n$  from a Bernoulli distributed population  $B(1, p)$  is an unbiased point estimator of the probability  $p$ .

For illustration, we show the validity of the third statement. Let us consider an i.i.d. random sample  $X_i, i = 1, 2, \dots, n$ , from a Bernoulli distribution, where  $X_i = 1$  if an event occurs and  $X_i = 0$  otherwise. Here,  $p$  is the probability of occurrence of an event in the population, i.e.  $p = P(X_i = 1)$ . Note that  $p$  is also the population mean:  $E(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p, i = 1, 2, \dots, n$ . The arithmetic mean (relative frequency) is an unbiased estimator of  $p$  because

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n p = p,$$

and thus, we can write the estimate of  $p$  as

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (9.6)$$

*Example 9.2.3* Suppose a random sample of size  $n = 20$  of the weight of 10-year-old children in a particular city is drawn. Let us assume that the children's weight in the population follows a normal distribution  $N(\mu, \sigma^2)$ . The sample provides the following values of weights (in kg):

$$40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3 \\ 34.6, 55.6, 50.9, 38.9, 37.8, 46.8, 43.6, 39.5, 49.9, 34.2$$

To obtain an estimate of the population mean  $\mu$ , we calculate the arithmetic mean of the observations as

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{20} (40.2 + 32.8 + \dots + 34.2) = 41.97,$$

because it is an unbiased estimator of  $\mu$ . Similarly, we use  $S^2$  to estimate  $\sigma^2$  because it is unbiased in comparison to  $\tilde{S}^2$ . Using  $s_x^2$  as an estimate for  $\sigma^2$  for the given observations, we get

$$\begin{aligned} \hat{\sigma}^2 &= s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{19} ((40.2 - 41.97)^2 + \dots + (34.2 - 41.97)^2) \approx 36.85. \end{aligned}$$

The square root of 36.85 is approximately 6.07 which is the standard deviation. Note that the standard deviation based on the sample values divided by the square root of the sample size, i.e.  $\hat{\sigma}/\sqrt{20}$ , is called the **standard error** of the mean  $\bar{X}$  (SEM). As already introduced in Chap. 3, we obtain these results in *R* using the `mean` and `var` commands.

*Example 9.2.4* A library draws a random sample of size  $n = 100$  members from the members' database to see how many members have to pay a penalty for returning books late, i.e.  $x_i = 1$ . It turns out that 39 members in the sample have to pay a penalty. Therefore, an unbiased estimator of the population proportion of all members of the library who return books late is

$$\hat{p} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{100} \cdot 39 = \frac{39}{100} = 0.39.$$

*Remark 9.2.2* Unbiasedness and efficiency can also be defined asymptotically: we say, for example, that an estimator is asymptotically unbiased, if the bias approaches zero when the sample size tends to infinity. The concept of asymptotic efficiency involves some mathematical knowledge which is beyond the intended scope of this book. Loosely speaking, an asymptotic efficient estimator is an estimator which achieves the lowest possible (asymptotic) variance under given distributional assumptions. The estimators introduced in Sect. 9.3.1, which are based on the maximum likelihood principle, have these properties (under certain mathematically defined regularity conditions).

Next, we illustrate the properties of consistency and sufficiency of an estimator.

## 9.2.2 Consistency of Estimators

For a good estimator, as the sample size increases, the values of the estimator should get closer to the parameter being estimated. This property of estimators is referred to as consistency.

**Definition 9.2.6** Let  $T_1, T_2, \dots, T_n$ , be a sequence of estimators for the parameter  $\theta$  where  $T_n = T_n(X_1, X_2, \dots, X_n)$  is a function of  $X_1, X_2, \dots, X_n$ . The sequence  $\{T_n\}$  is a **consistent** sequence of estimators for  $\theta$  if for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P[|T_n - \theta| < \epsilon] = 1$$

or equivalently

$$\lim_{n \rightarrow \infty} P[|T_n - \theta| \geq \epsilon] = 0.$$

This definition says that as the sample size  $n$  increases, the probability that  $T_n$  is getting closer to  $\theta$  is approaching 1. This means that the estimator  $T_n$  is getting closer to the parameter  $\theta$  as  $n$  grows larger. Note that there is no information on how *fast*  $T_n$  is converging to  $\theta$  in the sense of convergence defined above.

*Example 9.2.5* Let  $X_1, X_2 \dots, X_n$  be identically and independently distributed variables with expectation  $\mu$  and variance  $\sigma^2$ . Then for  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ , we have  $E(\bar{X}_n) = \mu$  and  $\text{Var}(\bar{X}_n) = \sigma^2/n$ . For any  $\epsilon > 0$ , we can write the following:

$$P[|\bar{X}_n - \mu| \geq \epsilon] = P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right]$$

where  $\epsilon = c\sigma/\sqrt{n}$ . Using Tschebyschev's inequality (Theorem 7.4.1, p. 139), we get  $\frac{1}{c^2} = \sigma^2/n\epsilon^2$ , and therefore

$$P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right] \leq \frac{1}{c^2} = \frac{\sigma^2}{n\epsilon^2}$$

and

$$\lim_{n \rightarrow \infty} P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right] \leq \lim_{n \rightarrow \infty} \frac{\sigma^2}{n\epsilon^2} = 0,$$

provided  $\sigma^2$  is finite. Hence  $\bar{X}_n, n = 1, 2, \dots$ , converges to  $\mu$  and therefore  $\bar{X}_n$  is a consistent estimator of  $\mu$ .

*Remark 9.2.3* We call this type of consistency *weak consistency*. Another definition is *MSE consistency*, which says that an estimator is MSE consistent if  $\text{MSE} \rightarrow 0$  as  $n \rightarrow \infty$ . If the estimator is unbiased, it is sufficient that  $\text{Var} \rightarrow 0$  as  $n \rightarrow \infty$ . If  $T_n(X)$  is MSE consistent, it is also weakly consistent. Therefore, it follows that an unbiased estimator with its variance approaching zero as the sample size approaches infinity is both MSE consistent and weakly consistent.

In Example 9.2.5, the variance of  $T_n(X) = \bar{X}_n$  is  $\sigma^2/n$  which goes to zero as  $n$  goes to  $\infty$  and therefore  $\bar{X}_n$  is both weakly consistent and MSE consistent.

### 9.2.3 Sufficiency of Estimators

Sufficiency is another criterion to judge the quality of an estimator. Before delving deeper into the subject matter, we first try to understand some basic concepts.

Consider two independent random variables  $X$  and  $Y$ , each following a  $N(\mu, 1)$  distribution. We conclude that both  $X$  and  $Y$  contain information about  $\mu$ . Consider two estimators of  $\mu$  as  $\hat{\mu}_1 = X + Y$  and  $\hat{\mu}_2 = X - Y$ . Suppose we want to know whether to use  $\hat{\mu}_1$  or  $\hat{\mu}_2$  to estimate  $\mu$ . We notice that  $E(\hat{\mu}_1) = E(X) + E(Y) = \mu + \mu = 2\mu$ ,  $E(\hat{\mu}_2) = E(X) - E(Y) = \mu - \mu = 0$ ,  $\text{Var}(\hat{\mu}_1) = \text{Var}(X) + \text{Var}(Y) = 1 + 1 = 2$  and  $\text{Var}(\hat{\mu}_2) = \text{Var}(X) + \text{Var}(Y) = 1 + 1 = 2$ . Using the additivity property of the normal distribution, which was introduced in Remark 8.2.2, we can say that  $\hat{\mu}_1 \sim N(2\mu, 2)$  and  $\hat{\mu}_2 \sim N(0, 2)$ . So  $\hat{\mu}_1$  contains information about  $\mu$ , whereas  $\hat{\mu}_2$

does not contain any information about  $\mu$ . In other words,  $\hat{\mu}_2$  loses the information about  $\mu$ . We call this property “loss of information”.

If we want to make conclusions about  $\mu$  using both  $X$  and  $Y$ , we need to acknowledge that the dimension of them is 2. On the other hand, if we use  $\hat{\mu}_1$  or equivalently  $\hat{\mu}_1/2 \sim N(\mu, \frac{1}{2})$ , then we need to concentrate only on one variable and we say that it has dimension 1. It follows that  $\hat{\mu}_1$  and  $\hat{\mu}_1/2$  provide the same information about  $\mu$  as provided by the entire sample on both  $X$  and  $Y$ . So we can say that either  $\hat{\mu}_1$  or  $\hat{\mu}_1/2$  is sufficient to provide the same information about  $\mu$  that can be obtained on the basis of the entire sample. This is the idea behind the concept of sufficiency and it results in the reduction of dimension. In general, we can say that if all the information about  $\mu$  contained in the sample of size  $n$  can be obtained, for example, through the sample mean then it is sufficient to use this one-dimensional summary statistic to make inference about  $\mu$ .

**Definition 9.2.7** Let  $X_1, X_2, \dots, X_n$  be a random sample from a probability density function (or probability mass function)  $f(x, \theta)$ . A statistic  $T$  is said to be sufficient for  $\theta$  if the conditional distribution of  $X_1, X_2, \dots, X_n$  given  $T = t$  is independent of  $\theta$ .

The Neyman–Fisher Factorization Theorem provides a practical way to find sufficient statistics.

**Theorem 9.2.2** (Neyman–Fisher Factorization Theorem (NFFT)) *Let  $X_1, X_2, \dots, X_n$  be a random sample from a probability density function (or probability mass function)  $f(x, \theta)$ . A statistic  $T = T(x_1, x_2, \dots, x_n)$  is said to be sufficient for  $\theta$  if and only if the joint density of  $X_1, X_2, \dots, X_n$  can be factorized as*

$$f(x_1, x_2, \dots, x_n; \theta) = g(t, \theta) \cdot h(x_1, x_2, \dots, x_n)$$

where  $h(x_1, x_2, \dots, x_n)$  is nonnegative and does not involve  $\theta$ ; and  $g(t, \theta)$  is a non-negative function of  $\theta$  which depends on  $x_1, x_2, \dots, x_n$  only through  $t$ , which is a particular value of  $T$ .

This theorem holds for discrete random variables too. Any one-to-one function of a sufficient statistic is also sufficient. A function  $f$  is called one-to-one if whenever  $f(a) = f(b)$  then  $a = b$ .

*Example 9.2.6* Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, 1)$  where  $\mu$  is unknown. We attempt to find a sufficient statistic for  $\mu$ . Consider the following function as the joint distribution of  $x_1, x_2, \dots, x_n$  (whose interpretation will become clearer in the next section):

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \mu) &= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left( -\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \right) \\ &= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left( -\frac{n\mu^2}{2} + \mu \sum_{i=1}^n x_i \right) \exp \left( -\frac{1}{2} \sum_{i=1}^n x_i^2 \right). \end{aligned}$$

Here

$$g(t, \mu) = \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left( -\frac{n\mu^2}{2} + \mu \sum_{i=1}^n x_i \right),$$

$$h(x_1, x_2, \dots, x_n) = \exp \left( -\frac{1}{2} \sum_{i=1}^n x_i^2 \right),$$

$$t = t(x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i.$$

Using the Neyman–Fisher Factorization Theorem, we conclude that  $T = T(X_1, X_2, \dots, X_n) = \sum_{i=1}^n X_i$  is a sufficient statistic for  $\mu$ . Also,  $T = T(X_1, X_2, \dots, X_n) = \bar{X}$  is sufficient for  $\mu$  as it is a one-to-one statistic of  $\sum_{i=1}^n X_i$ . On the other hand,  $T = \bar{X}^2$  is not sufficient for  $\mu$  as it is not a one-to-one function of  $\sum_{i=1}^n X_i$ . The important point here is that  $\bar{X}$  is a function of the sufficient statistic and hence a good estimator for  $\mu$ . It is thus summarizing the sample information about the parameter of interest in a complete yet parsimonious way. Another, multivariate, example of sufficiency is given in Appendix C.4.

## 9.3 Point Estimation

In the previous section, we introduced and discussed various properties of estimators. In this section, we want to show how one can find estimators with good properties. In the general case, properties such as unbiasedness and efficiency cannot be guaranteed for a finite sample. But often, the properties can be shown to hold asymptotically.

### 9.3.1 Maximum Likelihood Estimation

We have used several estimators throughout the book without stating explicitly that they are estimators. For example, we used the sample mean ( $\bar{X}$ ) to estimate  $\mu$  in a  $N(\mu, \sigma^2)$  distribution; we also used the sample proportion (relative frequency) to estimate  $p$  in a  $B(1, p)$  distribution, etc. The obvious question is how to obtain a good statistic to estimate an unknown parameter, for example how to determine that the sample mean can be used to estimate  $\mu$ . We need a general framework for parameter estimation. The method of **maximum likelihood** provides such an approach. For the purpose of illustration, we introduce the method of maximum likelihood estimation with an example using the Bernoulli distribution.

*Example 9.3.1* Consider an i.i.d. random sample  $X = (X_1, X_2, \dots, X_n)$  from a Bernoulli population with  $p = P(X_i = 1)$  and  $(1 - p) = P(X_i = 0)$ . The joint probability mass function for a given set of realizations  $x_1, x_2, \dots, x_n$  (i.e. the data) is

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | p) &= P(X_1 = x_1 | p) \cdot \dots \cdot P(X_n = x_n | p) \\ &= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}. \end{aligned} \quad (9.7)$$

This is a function of  $(x_1, x_2, \dots, x_n)$  given the parameter  $p$ . The product results from the fact that the draws are independent and the fact that  $p^{x_i} (1-p)^{1-x_i} = p$  if  $x_i = 1$  and  $p^{x_i} (1-p)^{1-x_i} = 1 - p$  if  $x_i = 0$ . That is, the term  $p^{x_i} (1-p)^{1-x_i}$  covers results from both possible outcomes. Now, consider a random sample where the values  $x = (x_1, x_2, \dots, x_n)$  are known, for example  $x = (0, 1, 0, 0, \dots, 1)$ . Then, (9.7) can be seen as a function of  $p$  because  $(x_1, x_2, \dots, x_n)$  is known. In this case, after obtaining a sample of data, the function is called the likelihood function and can be written as

$$L(x_1, x_2, \dots, x_n | p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}. \quad (9.8)$$

The joint density function of  $X_1, X_2, \dots, X_n$  is called the **likelihood function**. For better understanding, consider a sample of size 5 with  $x = (x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 1, x_5 = 0)$ . The likelihood (function) is

$$L(1, 1, 0, 1, 0 | p) = p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^3 (1-p)^2. \quad (9.9)$$

The maximum likelihood estimation principle now says that the estimator  $\hat{p}$  of  $p$  is the value of  $p$  which maximizes the likelihood (9.8) or (9.9). In other words, the maximum likelihood estimate is the value which maximizes the probability of observing the realized sample from the likelihood function. In general, i.e. for any sample, we have to maximize the likelihood function (9.9) with respect to  $p$ . We use the well-known principle of maxima–minima to maximize the likelihood function in this case. In principle, any other optimization procedure can also be used, for example numerical algorithms such as the Newton–Raphson algorithm. If the likelihood is differentiable, the first-order condition for the maximum is that the first derivative with respect to  $p$  is zero. For maximization, we can transform the likelihood by a strictly monotone increasing function. This guarantees that the potential maximum is taken at the same point as in the original likelihood. A good and highly common choice is the *natural logarithm* since it transforms products in sums and sums are easy to differentiate by differentiating each term in the sum. The log-likelihood in our example is therefore

$$l(1, 1, 0, 1, 0 | p) = \ln L(1, 1, 0, 1, 0 | p) = \ln \{p^3 (1-p)^2\} \quad (9.10)$$

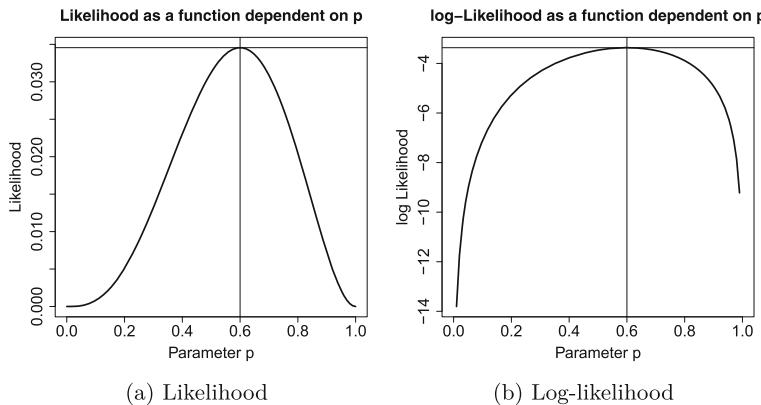
$$= 3 \ln(p) + 2 \ln(1-p) \quad (9.11)$$

where  $\ln$  denotes the natural logarithm function and we use the rules

$$\ln(a \cdot b) = \ln(a) + \ln(b), a > 0, b > 0$$

$$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b), a > 0, b > 0$$

$$\ln(a^b) = b \ln(a), a > 0.$$



**Fig. 9.2** Illustration of the likelihood and log-likelihood function of a binomial distribution

Taking the first derivative of (9.10) with respect to  $p$  results in

$$\frac{\partial l(1, 1, 0, 1, 0|p)}{\partial p} = \frac{3}{p} - \frac{2}{1-p}. \quad (9.12)$$

Setting (9.12) to zero and solving for  $p$  leads to

$$\begin{aligned} \frac{3}{p} - \frac{2}{1-p} &= 0 \\ \frac{3}{p} &= \frac{2}{1-p} \\ 3(1-p) &= 2p \\ 5p &= 3 \\ \hat{p}_{\text{ML}} &= \frac{3}{5} = \frac{1}{5}(1+1+0+1+0) = \bar{x}. \end{aligned}$$

The value of the second-order partial derivative of (9.9) with respect to  $p$  at  $p = \hat{p}_{\text{ML}}$  is negative which ensures that  $\hat{p}_{\text{ML}}$  maximizes the likelihood function. It follows from this example that the maximum likelihood estimate for  $p$  leads to the well-known arithmetic mean. Figure 9.2 shows the likelihood function and the log-likelihood function as functions of  $p$ , where  $p \in [0, 1]$ . The figures show that the likelihood function and the log-likelihood function have the same maxima at  $p = 3/5 = 0.6$ .

Maximum likelihood estimators have some important properties: they are usually consistent, asymptotically unbiased, asymptotically normally distributed, asymptotically efficient, and sufficient. Even if they are not, a function of a sufficient statistic can always be found which has such properties. This is the reason why maximum likelihood estimation is popular. By ‘asymptotically’ we mean that the properties hold as  $n$  tends to infinity, i.e. as the sample size increases. There might be other good estimators in a particular context, for example estimators that are efficient and not only asymptotically efficient; however, in general, the ML principle is a great

choice in many circumstances. We are going to use it in the following sections and chapters, for instance for general point estimation and in the linear regression model (Chap. 11).

*Remark 9.3.1* More examples of maximum likelihood estimators are given in Exercises 9.1–9.3.

### 9.3.2 Method of Moments

The **method of moments** is another well-known method to derive the estimators for population parameters. Below, we outline this principle briefly by way of example.

The idea is that the population parameters of interest can be related to the moments (e.g. expectation, variance) of the distribution of the considered random variables.

A simple case is the estimator for the expected value  $E(X) = \mu$  of a population using an i.i.d. random sample  $X = (X_1, \dots, X_n)$ . In this case,  $\hat{\mu} = \bar{X}$  is the natural moment estimator of  $\mu$ . Further, since  $E(X^2) = \sigma^2 + \mu^2$ , an estimator of  $\sigma^2 + \mu^2$  is  $\frac{1}{n} \sum_{i=1}^n X_i^2$ . Using  $\bar{X}^2$  as an estimator for  $\mu^2$ , this results in the biased, but asymptotically unbiased estimate

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

An extension of this method is the *generalized method of moments* (GMM). GMM estimators have interesting properties: under relatively weak conditions (not further discussed here), they are consistent and asymptotically normal, as well as efficient in the class of those estimators that do not use any additional information besides the information included in the moment conditions. Usually, they require a two-step estimating approach or an iterative estimating procedure.

The **least squares estimator** for a linear regression model with i.i.d. random errors, discussed in detail in Chap. 11, can be seen as a special case of a GMM estimator.

## 9.4 Interval Estimation

### 9.4.1 Introduction

Let us first consider an example to understand what we mean by interval estimation. Consider a situation in which a lady wants to know the time taken to travel from her home to the train station. Suppose she makes 20 trips and notes down the time taken. To get an estimate of the expected time, one can use the arithmetic mean. Let us say  $\bar{x} = 25$  min. This is the point estimate for the expected travelling time. It may not be appropriate to say that she will always take exactly 25 min to reach the train station.

Rather the time may vary by a few minutes each time. To take this into account, the time can be estimated in the form of an interval: it may then be found that the time varies mostly between 20 and 30 min. Such a statement is more informative. Both expectation and variation of the data are taken into account. The interval (20, 30 min) provides a range in which most of the values are expected to lie. We call this concept interval estimation.

A point estimate on its own does not take into account the precision of the estimate. The deviation between the point estimate and the true parameter (e.g.  $|\bar{x} - \mu|$ ) can be substantial, especially when the sample size is small. To incorporate the information about the precision of an estimate in the estimated value, a **confidence interval** can be constructed. It is a **random interval with lower and upper bounds**,  $I_l(\mathbf{X})$  and  $I_u(\mathbf{X})$ , such that the unknown parameter  $\theta$  is covered by a prespecified probability of at least  $1 - \alpha$ :

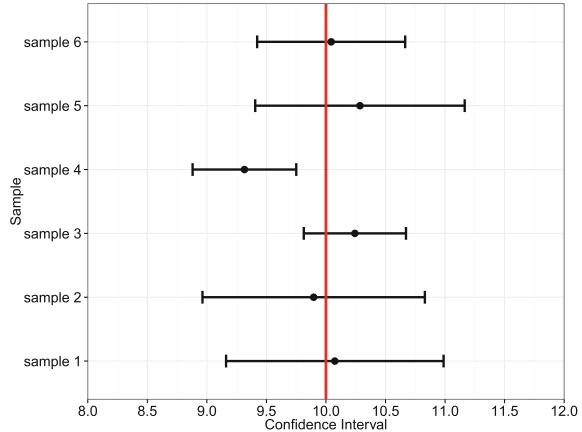
$$P_\theta(I_l(\mathbf{X}) \leq \theta \leq I_u(\mathbf{X})) \geq 1 - \alpha. \quad (9.13)$$

The probability  $1 - \alpha$  is called the **confidence level** or **confidence coefficient**,  $I_l(\mathbf{X})$  is called the **lower confidence bound** or **lower confidence limit** and  $I_u(\mathbf{X})$  is called the **upper confidence bound** or **upper confidence limit**. It is important to note that the bounds are random and the parameter is a fixed value. This is the reason why we say that the true parameter is covered by the interval with probability  $1 - \alpha$  and **not** that the probability that the interval contains the parameter is  $1 - \alpha$ . Please note that some software packages use the term “error bar” when referring to confidence intervals.

*Frequency interpretation of the confidence interval:* Suppose  $N$  independent samples  $\mathbf{X}^{(j)}$ ,  $j = 1, 2, \dots, N$ , of size  $n$  are sampled from the same population and  $N$  confidence intervals of the form  $[I_l(\mathbf{X}^{(j)}), I_u(\mathbf{X}^{(j)})]$  are calculated. If  $N$  is large enough, then on an average  $N(1 - \alpha)$  of the intervals (9.13) cover the true parameter.

*Example 9.4.1* Let a random variable follow a normal distribution with  $\mu = 10$  and  $\sigma^2 = 1$ . Suppose we draw a sample of  $n = 10$  observations repeatedly. The sample will differ in each draw, and hence, the mean and the confidence interval will also differ. The data sets are realizations from random variables. Have a look at Fig. 9.3 which illustrates the mean and the 95 % confidence intervals for 6 random samples. They vary with respect to the mean and the confidence interval width. Most of the means are close to  $\mu = 10$ , but not all. Similarly, most confidence intervals, but not all, include  $\mu$ . This is the idea of the frequency interpretation of the confidence interval: different samples will yield different point and interval estimates. Most of the times the interval will cover  $\mu$ , but not always. The coverage probability is specified by  $1 - \alpha$ , and the frequency interpretation means that we expect that (approximately)  $(1 - \alpha) \cdot 100\%$  of the intervals to cover the true parameter  $\mu$ . In that sense, the location of the interval will give us some idea about where the true but unknown population parameter  $\mu$  lies, while the length of the interval reflects our uncertainty about  $\mu$ : the wider the interval is, the higher is our uncertainty about the location of  $\mu$ .

**Fig. 9.3** Frequency interpretation of confidence intervals



We now introduce the following confidence intervals:

- Confidence interval for the mean  $\mu$  of a normal distribution.
- Confidence interval for the probability  $p$  of a binomial random variable.
- Confidence interval for the odds ratio.

#### 9.4.2 Confidence Interval for the Mean of a Normal Distribution

##### Confidence Interval for $\mu$ When $\sigma^2 = \sigma_0^2$ is Known.

Let  $X_1, X_2, \dots, X_n$  be an i.i.d. sample from a  $N(\mu, \sigma_0^2)$  distribution where  $\sigma_0^2$  is assumed to be known. We use the point estimate  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  to estimate  $\mu$  and construct a confidence interval around the mean  $\mu$ . Using the Central Limit Theorem (Appendix C.3, p. 426), it follows that  $\bar{X}$  follows a  $N(\mu, \sigma_0^2/n)$  distribution. Therefore  $\sqrt{n}(\bar{X} - \mu)/\sigma_0 \sim N(0, 1)$ , and it follows that

$$P_\mu \left( \left| \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0} \right| \leq z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha \quad (9.14)$$

where  $z_{1-\alpha/2}$  denotes the  $(1 - \alpha/2)$  quantile of the standard normal distribution  $N(0, 1)$ . We solve this inequality for the unknown  $\mu$  and get the desired confidence interval as follows:

$$P_\mu \left[ -z_{1-\frac{\alpha}{2}} \leq \left( \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0} \right) \leq z_{1-\frac{\alpha}{2}} \right] = 1 - \alpha$$

or

$$P_\mu \left[ \bar{X} - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \right] = 1 - \alpha .$$

The confidence interval for  $\mu$  is thus obtained as

$$[I_l(\mathbf{X}), I_u(\mathbf{X})] = \left[ \bar{X} - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \right]. \quad (9.15)$$

This is known as  $(1 - \alpha)\%$  confidence interval for  $\mu$  or the confidence interval for  $\mu$  with confidence coefficient  $\alpha$ .

We can use the *R* function `qnorm` or Table C.1 to obtain  $z_{1-\frac{\alpha}{2}}$ , see also Sects. 8.4, A.3, and C.7. For example, for  $\alpha = 0.05$  and  $\alpha = 0.01$  we get  $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$  and  $z_{1-\frac{\alpha}{2}} = z_{0.995} = 2.576$  using `qnorm(0.975)` and `qnorm(0.995)`. This gives us the quantiles we need to determine a 95 % and 99 % confidence interval, respectively.

*Example 9.4.2* We consider again Example 9.2.3 where we evaluated the weight of 10-year-old children. Assume that the variance is known to be 36; then the upper and lower limits of a 95 % confidence interval for the expected weight  $\mu$  can be calculated as follows:

$$\begin{aligned} I_l(X) &= \bar{X} - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} = 41.97 - 1.96 \frac{\sqrt{36}}{\sqrt{20}} \approx 39.34, \\ I_u(X) &= \bar{X} + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} = 41.97 + 1.96 \frac{\sqrt{36}}{\sqrt{20}} \approx 44.59. \end{aligned}$$

We get the confidence interval  $[I_u(X), I_l(X)] = [39.34, 44.59]$ . With 95 % confidence, the true parameter  $\mu$  is covered by the interval  $[39.34, 44.59]$ .

### Confidence Interval for $\mu$ When $\sigma^2$ is Unknown.

Let  $X_1, X_2, \dots, X_n$  be an i.i.d. sample from  $N(\mu, \sigma^2)$  where  $\sigma^2$  is assumed to be unknown and is being estimated by the sample variance  $S_X^2$ . We know from Sect. 8.3.1 that

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi_{n-1}^2.$$

It can be shown that  $\bar{X}$  and  $S_X^2$  are stochastically independent. Thus, we know that

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S_X} \sim t_{n-1}$$

follows a  $t$ -distribution with  $n - 1$  degrees of freedom. We can use this result to determine the confidence interval for  $\mu$  as

$$P_\mu \left[ -t_{1-\frac{\alpha}{2}, n-1} \leq \left( \frac{\sqrt{n}(\bar{X} - \mu)}{S_X} \right) \leq t_{1-\frac{\alpha}{2}, n-1} \right] = 1 - \alpha$$

or

$$P_\mu \left[ \bar{X} - t_{1-\frac{\alpha}{2}, n-1} \frac{S_X}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}, n-1} \frac{S_X}{\sqrt{n}} \right] = 1 - \alpha.$$

The confidence interval for  $\mu$  is thus obtained as

$$[I_l(\mathbf{X}), I_u(\mathbf{X})] = \left[ \bar{X} - t_{n-1; 1-\alpha/2} \cdot \frac{S_X}{\sqrt{n}}, \bar{X} + t_{n-1; 1-\alpha/2} \cdot \frac{S_X}{\sqrt{n}} \right] \quad (9.16)$$

which is the  $100(1 - \alpha)\%$  confidence interval for  $\mu$  or the confidence interval for  $\mu$  with confidence coefficient  $\alpha$ .

The interval (9.16) is, in general, wider than the interval (9.15) for identical  $\alpha$  and identical sample size  $n$ , since the unknown parameter  $\sigma^2$  is estimated by  $S_X^2$  which induces additional uncertainty. The quantiles for the  $t$ -distribution can be obtained using the *R* command `qt` or Table C.2.

*Example 9.4.3* Consider Example 9.4.2 where we evaluated the weight of 10-year-old children. We have already calculated the point estimate of  $\mu$  as  $\bar{x} = 41.97$ . With  $t_{19; 0.975} = 2.093$ , obtained via `qt(0.975, 19)` or Table C.2, the upper and lower limits of a 95% confidence interval for  $\mu$  are obtained as

$$\begin{aligned} I_u(X) &= \bar{x} - t_{19; 0.975} \cdot \frac{s_X}{\sqrt{n}} = 41.97 - 2.093 \cdot \frac{6.07}{\sqrt{20}} \approx 39.12, \\ I_o(X) &= \bar{x} + t_{19; 0.975} \cdot \frac{s_X}{\sqrt{n}} = 41.97 + 2.093 \cdot \frac{6.07}{\sqrt{20}} \approx 44.81. \end{aligned}$$

Therefore, the confidence interval is  $[I_l(X), I_u(X)] = [39.13, 44.81]$ . In *R*, we can use the `conf.int` value of the `t.test` command to get a confidence interval for the mean (see also Example 10.3.3 for more details on `t.test`). The default is a 95% confidence interval, but it can be changed easily if desired:

```
x <- c(40.2, 32.8, 38.2, 43.5, ..., 49.9, 34.2)
t.test(x, conf.level = 0.95)$conf.int
[1] 39.12384 44.80616
```



There is no unique best way to draw the calculated confidence intervals in *R*. Among many other options, one can simply work with the `plot` functionality or use `geom_errorbar` in conjunction with a `ggplot` object created with the library `ggplot2`, or use the `plotCI` command in the library `plotrix`.

### 9.4.3 Confidence Interval for a Binomial Probability

Let  $X_1, X_2, \dots, X_n$  be an i.i.d. sample from a Bernoulli distribution  $B(1, p)$ . Then  $Y = \sum_{i=1}^n X_i$  has a binomial distribution  $B(n, p)$ .

We have already introduced  $\hat{p}$  as an estimator for  $p$ :

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} Y.$$

From (8.8), we know that  $\text{Var}(Y) = np(1 - p)$ . Applying rule (7.33), the variance of the estimator  $\hat{p}$  is

$$\text{Var}(\hat{p}) = \frac{p(1 - p)}{n}$$

and it can be estimated by

$$S_{\hat{p}}^2 = \frac{\hat{p}(1 - \hat{p})}{n}.$$

Nowadays, the exact confidence intervals of the binomial distribution function can be easily calculated using computer implementations. Nevertheless, (i) for a sufficiently large sample size  $n$ , (ii) if  $p$  is not extremely low or high, and (iii) if the condition  $np(1 - p) \geq 9$  is fulfilled, we can use an approximation based on the normal distribution to calculate confidence intervals. To be more specific, one can show that

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} \xrightarrow{\text{approx.}} N(0, 1). \quad (9.17)$$

This gives us

$$P\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right] \approx 1 - \alpha, \quad (9.18)$$

and we get a confidence interval for  $p$  as

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right]. \quad (9.19)$$

*Example 9.4.4* We look again at Example 9.2.4 where we evaluated the proportion of members who had to pay a penalty. Out of all borrowers, 39 % brought back their books late and thus had to pay a fee. A 95 % confidence interval for the probability  $p$  of bringing back a book late can be constructed using the normal approximation, since  $n\hat{p}(1 - \hat{p}) = 100 \cdot 0.39 \cdot 0.61 = 23.79 > 9$ . With  $z_{1-\alpha/2} = z_{0.975} = 1.96$  and  $\hat{p} = 0.39$ , we get the 95 % confidence interval as

$$\left[0.39 - 1.96\sqrt{\frac{0.39 \cdot 0.61}{100}}, 0.39 + 1.96\sqrt{\frac{0.39 \cdot 0.61}{100}}\right] = [0.294, 0.486].$$

In R, an exact confidence interval can be found using the function `binom.test`:

```
binom.test(x=39,n=100)$conf.int
[1] 0.2940104 0.4926855
```



One can see that the exact and approximate confidence limits differ slightly due to the normal approximation which approximates the exact binomial probabilities.

### 9.4.4 Confidence Interval for the Odds Ratio

In Chap. 4, we introduced the odds ratio to determine the strength of association between two binary variables. One may be interested in the dispersion of the odds ratio and hence calculate a confidence interval for it. Recall the notation for  $2 \times 2$  contingency tables:

		$Y$		Total (row)
		$y_1$	$y_2$	
$X$	$x_1$	$a$	$b$	$a + b$
	$x_2$	$c$	$d$	$c + d$
Total (column)		$a + c$	$b + d$	$n$

In the spirit of the preceding sections, we can interpret the entries in this contingency table as population parameters. For example,  $a$  describes the absolute frequency of observations in the population for which  $Y = y_1$  and  $X = x_1$ . If we have a sample then we can estimate  $a$  by the number of *observed* observations  $n_{11}$  for which  $Y = y_1$  and  $X = x_1$ . We can thus view  $n_{11}$  to be an estimator for  $a$ ,  $n_{12}$  to be an estimator for  $b$ ,  $n_{21}$  to be an estimator for  $c$ , and  $n_{22}$  to be an estimator for  $d$ . It follows that

$$\widehat{\text{OR}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \quad (9.20)$$

serves as the point estimate for the population odds ratio  $\text{OR} = ad/bc$ . To construct a confidence interval for the odds ratio, we need to work on a log-scale. The log odds ratio,

$$\theta_0 = \ln \text{OR} = \ln a - \ln b - \ln c + \ln d, \quad (9.21)$$

takes the natural logarithm of the odds ratio. It is evident that it can be estimated using the observed absolute frequencies of the joint frequency distribution of  $X$  and  $Y$ :

$$\hat{\theta}_0 = \ln \widehat{\text{OR}} = \ln \frac{n_{11}n_{22}}{n_{12}n_{21}}. \quad (9.22)$$

It can be shown that  $\hat{\theta}_0$  follows approximately a normal distribution with expectation  $\theta_0$  and standard deviation

$$\hat{\sigma}_{\hat{\theta}_0} = \left( \frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} \right)^{\frac{1}{2}}. \quad (9.23)$$

Following the reasoning explained in the earlier section on confidence intervals for binomial probabilities, we can calculate the  $100(1 - \alpha)\%$  confidence interval for  $\theta_0$  under a normal approximation as follows:

$$\left[ \hat{\theta}_0 - z_{1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\theta}_0}, \hat{\theta}_0 + z_{1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\theta}_0} \right] = [I_u, I_o]. \quad (9.24)$$

Since we are interested in the confidence interval of the odds ratio, and not the log odds ratio, we need to transform back the lower and upper bound of the confidence interval as

$$[\exp(I_u), \exp(I_o)] . \quad (9.25)$$

*Example 9.4.5* Recall Example 4.2.5 from Chap. 4 where we were interested in the association of smoking with a particular disease. The data is summarized in the following  $2 \times 2$  contingency table:

		Smoking Yes No	Total (row)
Disease	Yes	34 66	100
	No	22 118	140
Total (column)		56 184	240

The odds ratio was estimated to be 2.76, and we therefore concluded that the chances of having the particular disease is 2.76 times higher for smokers compared with non-smokers. To calculate a 95 % confidence intervals, we need  $\hat{\theta}_0 = \ln(2.76)$ ,  $z_{1-\frac{\alpha}{2}} \approx 1.96$  and

$$\begin{aligned}\hat{\sigma}_{\hat{\theta}_0} &= \left( \frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} \right)^{\frac{1}{2}} \\ &= \left( \frac{1}{34} + \frac{1}{118} + \frac{1}{66} + \frac{1}{22} \right)^{\frac{1}{2}} \approx 0.314.\end{aligned}$$

The confidence interval for the log odds ratio is

$$[\ln(2.76) - 1.96 \cdot 0.314, \ln(2.76) + 1.96 \cdot 0.314] \approx [0.40, 1.63].$$

Exponentiation of the confidence interval bounds yields the 95 % confidence interval for the odds ratio as

$$[1.49, 5.11].$$

There are many ways to obtain the same results in R. One option is to use the `oddsratio` function of the library `epitools`. Note that we need to specify “wald” under the methods option to get confidence intervals which use the normal approximation as we did in this case.

```
library(epitools)
smd <- matrix(c(34,22,66,118),ncol=2,nrow=2) #data
oddsratio(smd,method='wald')
```



## 9.5 Sample Size Determinations

Confidence intervals help us estimating the precision of point estimates. What if we are required to adhere to a prespecified precision level? We know that the variance decreases as the sample size increases. In turn, confidence intervals become narrower. On the other hand, increasing the sample size has its own consequences. For example, the cost and time involved in setting up experiments, or conducting a survey, increases. In these situations it is important to find a balance between the variability of the estimates and the sample size. We cannot control the variability in the data in most of the situations, but it is possible to control the sample size and therefore the precision of our estimates. For example, we can control the number of people to be interviewed in a survey—given the resources which are available. We discuss how to determine the number of observations needed to get a particular precision (length) of the confidence interval. We find the answers to such questions using the formulae for confidence intervals.

### Sample Size Calculation for $\mu$ .

Let us consider the situation where we are interested in estimating the population mean  $\mu$ . The length of the confidence interval (9.15) for the point estimate  $\bar{X}$  is

$$2z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}. \quad (9.26)$$

We would now like to fix the width of the confidence interval and come up with a sample size which is required to achieve this width. Let us fix the length of the confidence interval as

$$\Delta = 2z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}. \quad (9.27)$$

Assume we have knowledge of  $\sigma_0$ . The knowledge about  $\sigma_0$  can be obtained, for example, through a pilot study or past experience with the experiment. We are interested in obtaining the value of  $n$  for which a confidence interval has a fixed confidence width of  $\Delta$  or less. Rearranging (9.27) gives us

$$n \geq \left[ 2 \frac{z_{1-\alpha/2} \sigma_0}{\Delta} \right]^2. \quad (9.28)$$

This means a minimum or optimum sample size is

$$n_{opt} = \left[ 2 \frac{z_{1-\alpha/2} \sigma_0}{\Delta} \right]^2. \quad (9.29)$$

The sample size  $n_{opt}$  ensures that the  $1 - \alpha$  confidence interval for  $\mu$  has at most length  $\Delta$ . But note that we have assumed that  $\sigma_0$  is known. If we do not know  $\sigma_0$  (which is more likely in practice), we have to make an assumption about it, e.g. by using an estimate from a former study, a pilot study, or other external information. Practically, (9.28) is used in the case of known and unknown  $\sigma_0^2$ .

*Example 9.5.1* A call centre is interested in determining the expected length of a telephone call as precisely as possible. The requirements are that the 95 % confidence interval for  $\mu$  should have a width of 1 min. Suppose that the call centre has developed a pilot study in which  $\sigma_0$  was estimated to be 5 min. The sample size  $n$  that is needed to estimate the expected length of the phone calls with the desired precision is:

$$n \geq \left[ \frac{2z_{1-\alpha/2}\sigma_0}{\Delta} \right]^2 = \left[ \frac{2 \times 1.96 \times 5}{1} \right]^2 \approx 384.$$

This means that at least 384 calls are required to get the desired confidence interval width.

### Sample Size Calculation for $p$ .

We can follow the earlier reasoning and determine the optimum sample size for a specific confidence interval width using the confidence interval definition (9.19). Since the width of the confidence interval is

$$\Delta = 2z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}},$$

we get

$$n \geq \left[ 2 \frac{z_{1-\alpha/2}}{\Delta} \right]^2 \hat{p}(1 - \hat{p}). \quad (9.30)$$

*Example 9.5.2* A factory may be interested in the probability of an error in an operating process. The length of the confidence interval should be  $\pm 2\%$ , i.e.  $\Delta = 0.04$ . Suppose it is speculated that the error probability is 10%; we may then use  $\hat{p} = 0.1$  as our prior judgment for the true value of  $p$ . This yields

$$n \geq \left[ 2 \frac{z_{1-\alpha/2}}{\Delta} \right]^2 \hat{p}(1 - \hat{p}) = \left[ 2 \times \frac{1.96}{0.04} \right]^2 0.1 \cdot (1 - 0.1) \approx 865. \quad (9.31)$$

This means we need a sample size of at least 865 to obtain the desired width of the confidence interval for  $p$ .

The above examples for both  $\mu$  and  $p$  have shown us that without external knowledge about the research question of interest, it is difficult to come up with an appropriate sample size. Results may vary considerably depending on what type of information is assumed to be known. With limited knowledge, it can be useful to report results for different widths of confidence intervals and hypothesized values of  $p$  or  $\sigma_0$ .

Sample size calculations can be highly complex in many practical situations and may not remain as simple as in the examples considered here. For example, Chap. 10 uses additional concepts in the context of hypothesis testing, such as the power, which can be taken into consideration when estimating sample sizes. However,

in this case, calculations and interpretations become more difficult and complex. A detailed overview of sample size calculations can be found in Chow et al. (2007) and Bock (1997).

## 9.6 Key Points and Further Issues

### Note:

- ✓ We have introduced important point estimates for the parameters of a normal and a binomial distribution:

$$\bar{x} \text{ for } \mu, \quad S^2 \text{ for } \sigma^2, \quad \bar{x} \text{ for } p.$$

In general, the choice of these point estimates is not arbitrary but follows some principles of statistical inference such as maximum likelihood estimation, or least squares estimation (introduced in Chap. 11).

- ✓ The maximum likelihood estimator is usually consistent, asymptotically unbiased, asymptotically normally distributed, and asymptotically efficient.
- ✓ The validity of all results in this chapter depends on the assumption that the data is complete and has no missing values. Incomplete data may yield different conclusions.
- ✓ A confidence interval is defined in terms of upper and lower confidence limits and covers the true target parameter with probability  $1 - \alpha$ . Confidence intervals are often constructed as follows:

$$\text{point estimate} \pm \text{quantile} \cdot \underbrace{\sqrt{\text{variance of point estimate}}}_{\text{standard error}}.$$

- ✓ More detailed introductions to inference are presented in Casella and Berger (2002) and Young and Smith (2005).

## 9.7 Exercises

*Exercise 9.1* Consider an i.i.d. sample of size  $n$  from a  $\text{Po}(\lambda)$  distributed random variable  $X$ .

- 
- (a) Determine the maximum likelihood estimate for  $\lambda$ .
  - (b) What does the log-likelihood function look like for the following realizations:  $x_1 = 4, x_2 = 3, x_3 = 8, x_4 = 6, x_5 = 6$ ? Plot the function using *R*. Hint: The `curve` command can be used to plot functions.
  - (c) Use the Neyman–Fisher Factorization Theorem to argue that the maximum likelihood estimate obtained in (a) is a sufficient statistic for  $\lambda$ .

*Exercise 9.2* Consider an i.i.d. sample of size  $n$  from a  $N(\mu, \sigma^2)$  distributed random variable  $X$ .

- (a) Determine the maximum likelihood estimator for  $\mu$  under the assumption that  $\sigma^2 = 1$ .
- (b) Now determine the maximum likelihood estimator for  $\mu$  for an arbitrary  $\sigma^2$ .
- (c) What is the maximum likelihood estimate for  $\sigma^2$ ?

*Exercise 9.3* Let  $X_1, X_2, \dots, X_n$  be  $n$  i.i.d. random variables which follow a uniform distribution,  $U(0, \theta)$ . Write down the likelihood function and argue, without differentiating the function, what the maximum likelihood estimate of  $\theta$  is.

*Exercise 9.4* Let  $X_1, X_2, \dots, X_n$  be  $n$  i.i.d. random variables which follow an exponential distribution. An intelligent statistician proposes to use the following two estimators to estimate  $\mu = 1/\lambda$ :

- (i)  $T_n(X) = nX_{\min}$  with  $X_{\min} = \min(X_1, \dots, X_n)$  and  $X_{\min} \sim \text{Exp}(n\lambda)$ ,
  - (ii)  $V_n(X) = n^{-1} \sum_{i=1}^n X_i$ .
- (a) Are both  $T_n(X)$  and  $V_n(X)$  (asymptotically) unbiased for  $\mu$ ?
  - (b) Calculate the mean squared error of both estimators. Which estimator is more efficient?
  - (c) Is  $V_n(X)$  MSE consistent, weakly consistent, both, or not consistent at all?

*Exercise 9.5* A national park in Namibia determines the weight (in kg) of a sample of common eland antelopes:

450 730 700 600 620 660 850 520 490 670 700 820  
910 770 760 620 550 520 590 490 620 660 940 790

Calculate

- (a) the point estimate of  $\mu$  and  $\sigma^2$  and
- (b) the confidence interval for  $\mu$  ( $\alpha = 0.05$ ).

under the assumption that the weight is normally distributed.

- (c) Use *R* to reproduce the results from (b).

*Exercise 9.6* We are interested in the heights of the players of the two basketball teams “Brose Baskets Bamberg” and “Bayer Giants Leverkusen” as well as the football team “SV Werder Bremen”. The following summary statistics are given:

	<i>N</i>	Minimum	Maximum	Mean	Std. dev.
Bamberg	16	185	211	199.06	7.047
Leverkusen	14	175	210	196.00	9.782
Bremen	23	178	195	187.52	5.239

Calculate a 95 % confidence interval for  $\mu$  for all three teams and interpret the results.

*Exercise 9.7* A married couple tosses a coin after each dinner to determine who has to wash the dishes. If the coin shows “head”, then the husband has to wash the dishes, and if the coin shows “tails”, then the wife has to wash the dishes. After 98 dinners, the wife notes that the coin has shown head 59 times.

- (a) Estimate the probability that the wife has to wash the dishes.
- (b) Calculate and interpret the 95 % confidence interval for  $p$ .
- (c) How many dinners are needed to estimate the true probability for the coin showing “head” with a precision of  $\pm 0.5\%$  under the assumption that the coin is fair?

*Exercise 9.8* Suppose 93 out of 104 pupils have passed the final examination at a certain school.

- (a) Calculate a 95 % confidence interval for the probability of failing the examination both by manual calculations and by using *R*, and compare the results.
- (b) At county level 3.2 % of pupils failed the examination. Are the school’s pupils worse than those in the whole county?

*Exercise 9.9* To estimate the audience rate for several TV stations, 3000 households are asked to allow a device, which records which TV station is watched, to be installed on their TVs. 2500 agreed to participate. Assume it is of interest to estimate the probability of someone switching on the TV and watching the show “Germany’s next top model”.

- (a) What is the precision with which the probability can be estimated?
- (b) What source of bias could potentially influence the estimates?

*Exercise 9.10* An Olympic decathlon athlete is interested in his performance compared with the performance of other athletes. He is a good runner and interested in his 100 m results compared with those of other athletes.

- He uses the decathlon data from this book (Appendix A.2) to come up with  $\hat{\sigma} = s = 0.233$ . What sample size does he need to calculate a 95 % confidence interval for the mean running time which is precise to  $\pm 0.1$  s?
- Calculate a 95 % confidence interval for the mean running time ( $\bar{x} = 10.93$ ) of the 30 athletes captured in the data set in Chap. A.2. Interpret the width of this interval compared with the width determined in a).
- The runner's own best time is 10.86 s. He wants to be among the best 10 % of all athletes. Calculate an appropriate confidence interval to compare his time with the 10 % best times.

*Exercise 9.11* Consider the pizza delivery data described in Chap. A.4. We distinguish between pizzas delivered on time (i.e. in less than 30 min) and not delivered on time (i.e. in more than 30 min). The contingency table for delivery time and operator looks as follows:

	Operator		Total
	Laura	Melissa	
<30 min	163	151	314
$\geq 30$ min	475	477	952
Total	638	628	1266

- Calculate and interpret the odds ratio and its 95 % confidence interval.
- Reproduce the results from (a) using *R*.

→ Solutions to all exercises in this chapter can be found on p. 384

---

## 10.1 Introduction

We introduced point and interval estimation of parameters in the previous chapter. Sometimes, the research question is less ambitious in the sense that we are not interested in precise estimates of a parameter, but we only want to examine whether a statement about a parameter of interest or the research hypothesis is true or not (although we will see later in this chapter that there is a connection between confidence intervals and statistical tests, called *duality*). Another related issue is that once an analyst estimates the parameters on the basis of a random sample, (s)he would like to infer something about the value of the parameter in the population. Statistical hypothesis tests facilitate the comparison of estimated values with hypothetical values.

*Example 10.1.1* As a simple example, consider the case where we want to find out whether the proportion of votes for a party  $P$  in an election will exceed 30 % or not. Typically, before the election, we will try to get representative data about the election proportions for different parties (e.g. by telephone interviews) and then make a statement like “yes”, we expect that  $P$  will get more than 30 % of the votes or “no”, we do not have enough evidence that  $P$  will get more than 30 % of the votes. In such a case, we will only know after the election whether our statement was right or wrong. Note that the term representative data only means that the sample is similar to the population with respect to the distributions of some key variables, e.g. age, gender, and education. Since we use one sample to compare it with a fixed value (30 %), we call it a **one-sample problem**.

*Example 10.1.2* Consider another example in which a clinical study is conducted to compare the effectiveness of a new drug ( $B$ ) to an established standard drug ( $A$ ) for a specific disease, for example too high blood pressure. Assume that, as a first step, we want to find out whether the new drug causes a higher reduction in blood

pressure than the already established older drug. A frequently used study design for this question is a randomized (i.e. patients are randomly allocated to one of the two treatments) controlled clinical trial (double blinded, i.e. neither the patient nor the doctor know which of the drugs a patient is receiving during the trial), conducted in a fixed time interval, say 3 months. A possible hypothesis is that the average change in the blood pressure in group  $B$  is higher than in group  $A$ , i.e.  $\delta_B > \delta_A$  where  $\delta_j = \mu_{j0} - \mu_{j3}$ ,  $j = A, B$  and  $\mu_{j0}$  is the average blood pressure at baseline before measuring the blood pressure again after 3 months ( $\mu_{j3}$ ). Note that we expect both the differences  $\delta_A$  and  $\delta_B$  to be positive, since otherwise we would have some doubt that either drug is effective at all. As a second step (after statistically proving our hypothesis), we are interested in whether the improvement of  $B$  compared to  $A$  is relevant in a medical or biological sense and is valid for the entire population or not. This will lead us again to the estimation problems of the previous chapter, i.e. quantifying an effect using point and interval estimation. Since we are comparing two drugs, we need to have two samples from each of the drugs; hence, we have a **two-sample problem**. Since the patients receiving  $A$  are different from those receiving  $B$  in this example, we refer to it as a “two-independent-samples problem”.

*Example 10.1.3* In another example, we consider an experiment in which a group of students receives extra mathematical tuition. Their ability to solve mathematical problems is evaluated before and after the extra tuition. We are interested in knowing whether the ability to solve mathematical problems increases after the tuition, or not. Since the same group of students is used in a pre–post experiment, this is called a “two-dependent-samples problem” or a “paired data problem”.

---

## 10.2 Basic Definitions

### 10.2.1 One- and Two-Sample Problems

In one-sample problems, the data is usually assumed to arise as *one* sample from a defined population. In two-sample problems, the data originates in the form of *two samples* possibly from two different populations. The heterogeneity is often modelled by assuming that the two populations only differ in some parameters or key quantities such as expectation (i.e. mean), median, or variance. As in our introductory example, the samples can either be independent (as in the drug Example 10.1.2) or dependent (as in the evaluation Example 10.1.3).

### 10.2.2 Hypotheses

A researcher may have a research question for which the truth about the population of interest is unknown. Suppose data can be obtained using a survey, observation, or

an experiment: if, given a prespecified uncertainty level, a statistical test based on the data supports the hypothesis about the population, we say that this hypothesis is statistically proven. Note that the research question has to be operationalized before it can be tested by a statistical test. Consider the drug Example 10.1.2: we want to examine whether the new drug  $B$  has a greater blood pressure lowering effect than the standard drug  $A$ . We have several options to operationalize this research question into a statistical set-up. One is to test whether the *average* reduction (from baseline to 3 months) of the blood pressure is higher (and positive) for drug  $B$  than drug  $A$ . We then state our hypotheses in terms of expected values (i.e.  $\mu$ ). Why do we have to use the expected values  $\mu$  and not simply compare the arithmetic means  $\bar{x}$ ? The reason is that the superiority of  $B$  shown in the sample will only be valid for this sample and not necessarily for another sample. We need to show the superiority of  $B$  in the entire population, and hence, our hypothesis needs to reflect this. Another option would be, for example, to use median changes in blood pressure values instead of mean changes in blood pressure values. An important point is that the research hypothesis which we want to prove has to be formulated as the statistical alternative hypothesis, often denoted by  $H_1$ . The reason for this will become clearer later in this chapter. The opposite of the research hypothesis has to be formulated as the statistical null hypothesis, denoted by  $H_0$ . In the drug example, the alternative and null hypotheses are, respectively,

$$H_1 : \delta_B > \delta_A$$

and

$$H_0 : \delta_B \leq \delta_A.$$

We note that the two hypotheses are disjoint and the union of them covers all possible differences of  $\delta_B$  and  $\delta_A$ . There is a boundary value ( $\delta_B = \delta_A$ ) which separates the two hypotheses. Since we want to show the superiority of  $B$ , the hypothesis was formulated as a one-sided hypothesis. Note that there are different ways to formulate two-sample hypotheses; for example,  $H_1 : \delta_B > \delta_A$  is equivalent to  $H_1 : \delta_B - \delta_A > 0$ . In fact, it is very common to formulate two-sample hypotheses as differences, which we will see later in this chapter.

### 10.2.3 One- and Two-Sided Tests

We distinguish between one-sided and two-sided hypotheses and tests. In the previous section, we gave an example of a one-sided test.

For an unknown population parameter  $\theta$  (e.g.  $\mu$ ) and a fixed value  $\theta_0$  (e.g. 5), the following three cases have to be distinguished:

Case	Null hypothesis	Alternative hypothesis	
(a)	$\theta = \theta_0$	$\theta \neq \theta_0$	Two-sided test problem
(b)	$\theta \geq \theta_0$	$\theta < \theta_0$	One-sided test problem
(c)	$\theta \leq \theta_0$	$\theta > \theta_0$	One-sided test problem

*Example 10.2.1* One-sample problems often test whether a target value is achieved or not. For example, consider the null hypothesis as

- $H_0$  : average filling weight of packages of flour = 1 kg
- $H_0$  : average body height (men) = 178 cm.

The alternative hypothesis  $H_1$  is formulated as deviation from the target value. If deviations in both directions are interesting, then  $H_1$  is formulated as a two-sided hypothesis,

- $H_1$  : average body height (men)  $\neq$  178 cm.

If deviations in a specific direction are the subject of interest, then  $H_1$  is formulated as a one-sided hypothesis, for example,

- $H_1$  : average filling weight of flour packages is lower than 1 kg.
- $H_1$  : average filling weight of flour packages is greater than 1 kg.

Two-sample problems often examine differences of two samples. Suppose the null hypothesis  $H_0$  is related to the average weight of flour packages filled by two machines, say 1 and 2. Then, the null hypothesis is

- $H_0$  : average weight of flour packages filled by machine 1 = average weight of flour packages filled by machine 2.

Then,  $H_1$  can be formulated as a one-sided or two-sided hypothesis. If we want to prove that machine 1 and machine 2 have different filling weights, then  $H_1$  would be formulated as a two-sided hypothesis

- $H_1$  : average filling weight of machine 1  $\neq$  average filling weight of machine 2.

If we want to prove that machine 1 has lower average filling weight than machine 2,  $H_1$  would be formulated as a one-sided hypothesis

- $H_1$  : average filling weight of machine 1 < average filling weight of machine 2.

If we want to prove that machine 2 has lower filling weight than machine 1,  $H_1$  would be formulated as a one-sided hypothesis

- $H_1$  : average filling weight of machine 1 > average filling weight of machine 2.

*Remark 10.2.1* Note that we have *not* considered the following situation:  $H_0 : \theta \neq \theta_0$ ,  $H_1 : \theta = \theta_0$ . In general, with the tests described in this chapter, we cannot prove the equality of a parameter to a predefined value and neither can we prove the equality of two parameters, as in  $H_0 : \theta_1 \neq \theta_2$ ,  $H_1 : \theta_1 = \theta_2$ . We can, for example,

not prove (statistically) that machines 1 and 2 in the previous example provide equal filling weight. This would lead to the more complex class of equivalence tests, which is a topic beyond the scope of this book.

### 10.2.4 Type I and Type II Error

If we undertake a statistical test, two types of error can occur.

- The hypothesis  $H_0$  is true but is rejected; this error is called **type I error**.
- The hypothesis  $H_0$  is not rejected although it is wrong; this is called **type II error**.

When a hypothesis is tested, then the following four situations are possible:

	$H_0$ is true	$H_0$ is not true
$H_0$ is not rejected	Correct decision	Type II error
$H_0$ is rejected	Type I error	Correct decision

The significance level is the probability of type I error,  $P(H_1|H_0) = \alpha$ , which is the probability of rejecting  $H_0$  (accepting  $H_1$ ) if  $H_0$  is true. If we construct a test, the significance level  $\alpha$  is prespecified, e.g.  $\alpha = 0.05$ . A significance test is constructed such that the probability of a type I error does not exceed  $\alpha$  while the probability of a type II error depends on the true but unknown parameter values in the population(s) and the sample size. Therefore, the two errors are not symmetrically treated in a significance test. In fact, the type II error  $\beta$ ,  $P(H_0|H_1) = \beta$  is not controlled by the construction of the test and can become very high, sometimes up to  $1 - \alpha$ . This is the reason why a test not rejecting  $H_0$  is not a (statistical) proof of  $H_0$ . In mathematical statistics, one searches for the best test which maintains  $\alpha$  and minimizes  $\beta$ . Minimization of both  $\alpha$  and  $\beta$  simultaneously is not possible. The reason is that when  $\alpha$  increases then  $\beta$  decreases and vice versa. So one of the errors needs to be fixed and the other error is minimized. Consequently, the error which is considered more serious is fixed and then the other error is minimized. The tests discussed in the below sections are obtained based on the assumption that the type I error is more serious than the type II error. So the test statistics are obtained by fixing  $\alpha$  and then minimizing  $\beta$ . In fact, the null hypothesis is framed in such a way that it implies that the type I error is more serious than the type II error. The probability  $1 - \beta = P(H_1|H_1)$  is called the **power** of the test. It is the probability of making a decision in favour of the research hypothesis  $H_1$ , if it is true, i.e. the probability of detecting a correct research hypothesis.

### 10.2.5 How to Conduct a Statistical Test

In general, we can follow the steps described below to test a hypothesis about a population parameter based on a sample of data.

- (1) Define the distributional assumptions for the random variables of interest, and specify them in terms of population parameters (e.g.  $\theta$  or  $\mu$  and  $\sigma$ ). This is necessary for parametric tests. There are other types of tests, so-called nonparametric tests, where the assumptions can be relaxed in the sense that we do not have to specify a particular distribution, see Sect. 10.6ff. Moreover, for some tests the distributional assumptions can be relaxed if the sample size is large.
- (2) Formulate the null hypothesis and the alternative hypothesis as described in Sects. 10.2.2 and 10.2.3.
- (3) Fix a significance value (often called type I error)  $\alpha$ , for example  $\alpha = 0.05$ , see also Sect. 10.2.4.
- (4) Construct a test statistic  $T(\mathbf{X}) = T(X_1, X_2, \dots, X_n)$ . The distribution of  $T$  has to be known under the null hypothesis  $H_0$ . We note again that  $(X_1, X_2, \dots, X_n)$  refers to the random variables before drawing the actual sample and  $x_1, x_2, \dots, x_n$  are the realized values (observations) in the sample.
- (5) Construct a critical region  $K$  for the statistic  $T$ , i.e. a region where—if  $T$  falls in this region— $H_0$  is rejected, such that

$$P_{H_0}(T(\mathbf{X}) \in K) \leq \alpha .$$

The notation  $P_{H_0}(\cdot)$  means that this inequality must hold for all parameter values  $\theta$  that belong to the null hypothesis  $H_0$ . Since we assume that we know the distribution of  $T(\mathbf{X})$  under  $H_0$ , the critical region is defined by those values of  $T(\mathbf{X})$  which are unlikely (i.e. with probability of less than  $\alpha$ ) to be observed under the null hypothesis. Note that although  $T(\mathbf{X})$  is a random variable,  $K$  is a well-defined region, see Fig. 10.1 for an example.

- (6) Calculate  $t(x) = T(x_1, x_2, \dots, x_n)$  based on the realized sample values  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ .
- (7) Decision rule: if  $t(x)$  falls into the critical region  $K$ , the null hypothesis  $H_0$  is rejected. The alternative hypothesis is then statistically proven. If  $t(x)$  falls outside the critical region,  $H_0$  is not rejected.

$t(x) \in K : H_0$  rejected  $\Rightarrow H_1$  is statistically significant,

$t(x) \notin K : H_0$  not rejected and therefore accepted.

The next two paragraphs show how to arrive at the test decisions from step 7 in a different way. Readers interested in an example of a statistical test may jump to Sect. 10.3.1 and possibly also Example 10.3.1.

### 10.2.6 Test Decisions Using the $p$ -Value

Statistical software usually does not show us all the steps of hypothesis testing as outlined in Sect. 10.2.5. It is common that instead of calculating and reporting the critical values, the test statistic is printed together with the so-called  $p$ -value. It is possible to use the  $p$ -value instead of critical regions for making test decisions. The  $p$ -value of the test statistic  $T(\mathbf{X})$  is defined as follows:

$$\text{two-sided case: } P_{H_0}(|T| \geq t(x)) = p\text{-value}$$

$$\text{one-sided case: } P_{H_0}(T \geq t(x)) = p\text{-value}$$

$$P_{H_0}(T \leq t(x)) = p\text{-value}$$

It can be interpreted as the probability of observing results equal to, or more extreme than those actually observed if the null hypothesis was true. Then, the decision rule is

$H_0$  is rejected if the  $p$ -value is smaller than the prespecified significance level  $\alpha$ .  
Otherwise,  $H_0$  cannot be rejected.

*Example 10.2.2* Assume that we are dealing with a two-sided test and assume further that the test statistic  $T(x)$  is  $N(0, 1)$ -distributed under  $H_0$ . The significance level is  $\alpha = 0.05$ . If we observe, for example,  $t = 3$ , then the  $p$ -value is  $P_{H_0}(|T| \geq 3)$ . This can be calculated in  $R$  as

`2*(1-pnorm(3))`



because `pnorm()` is used to calculate  $P(X \leq x)$ , and therefore, `1-pnorm()` can be used to calculate  $P(X > x)$ . We have to multiply with two because we are dealing with a two-sided hypothesis. The result is  $p = 0.002699796$ . Therefore,  $H_0$  is rejected. The one-sided  $p$ -value is half of the two-sided  $p$ -value, i.e.  $P(T \geq 3) = P(T \leq -3) = 0.001349898$ , and is not necessarily reported by  $R$ . It is therefore important to look carefully at the  $R$  output when dealing with one-sided hypotheses.

The  $p$ -value is sometimes also called the *significance*, although we prefer the term  $p$ -value. We use the term *significance* only in the context of a test result: a test is (statistically) significant if (and only if)  $H_0$  can be rejected.

Unfortunately, the  $p$ -value is often over-interpreted: both a test and the  $p$ -value can only provide a yes/no decision: either  $H_0$  is rejected or not. Interpreting the  $p$ -value as the probability that the null hypothesis is true is wrong! It is also incorrect to say that the  $p$ -value is the probability of making an error during the test decision. In our (frequentist) context, hypotheses are true or false and no probability is assigned to them. It can also be misleading to speak of “highly significant” results if the  $p$ -value is very small. A last remark: the  $p$ -value itself is a random variable: under the null hypothesis, it follows a uniform distribution, i.e.  $p \sim U(0, 1)$ .

### 10.2.7 Test Decisions Using Confidence Intervals

There is an interesting and useful relationship between confidence intervals and hypothesis tests. If the null hypothesis  $H_0$  is rejected at the significance level  $\alpha$ , then there exists a  $100(1 - \alpha)\%$  confidence interval which yields the same conclusion as the test: if the appropriate confidence interval does not contain the value  $\theta_0$  targeted in the hypothesis, then  $H_0$  is rejected. We call this **duality**. For example, recall Example 10.1.2 where we were interested in whether the average change in blood pressure for drug  $B$  is higher than for drug  $A$ , i.e.  $H_1 : \delta_B > \delta_A$ . This hypothesis is equivalent to  $H_1 : \delta_B - \delta_A > \delta_0 = 0$ . In the following section, we develop tests to decide whether  $H_1$  is statistically significant or not. Alternatively, we could construct a  $100(1 - \alpha)\%$  confidence interval for the difference  $\delta_B - \delta_A$  and evaluate whether the interval contains  $\delta_0 = 0$  or not; if yes, we accept  $H_0$ ; otherwise, we reject it. For some of the tests introduced in following section, we refer to the confidence intervals which lead to the same results as the corresponding test.

## 10.3 Parametric Tests for Location Parameters

### 10.3.1 Test for the Mean When the Variance is Known (One-Sample Gauss Test)

We develop a hypothesis test to test whether the unknown mean (expectation)  $\mu$  of a  $N(\mu, \sigma^2)$ -distributed random variable  $X$  either differs from a specific value  $\mu = \mu_0$  or is smaller (or greater) than  $\mu_0$ . We assume that the variance  $\sigma^2 = \sigma_0^2$  is known. We apply the scheme of Sect. 10.2.5 step by step to develop the test procedure and then give an illustrative example.

1. *Distributional assumption:* The random variable  $X$  follows a  $N(\mu, \sigma_0^2)$ -distribution with known variance  $\sigma_0^2$ . We assume that an i.i.d. random sample is drawn from  $X_1, X_2, \dots, X_n$  where the  $X_i$ s follow the same distribution as  $X$ ,  $i = 1, 2, \dots, n$ .

2. Define any of the following set of hypotheses  $H_0$  and  $H_1$ :

$$\begin{aligned} H_0 : \mu = \mu_0 &\text{ versus } H_1 : \mu \neq \mu_0, && \text{(two-sided test)} \\ H_0 : \mu \leq \mu_0 &\text{ versus } H_1 : \mu > \mu_0, && \text{(one-sided test)} \\ H_0 : \mu \geq \mu_0 &\text{ versus } H_1 : \mu < \mu_0, && \text{(one-sided test).} \end{aligned}$$

3. *Specify the probability of a type I error  $\alpha$ :* Often  $\alpha = 0.05 = 5\%$  is chosen.

4. *Construct a test statistic:* The unknown mean, i.e. the expectation  $\mu$ , is usually estimated by the sample mean  $\bar{x}$ . We already know that if the  $X_i$ s are i.i.d., then the sample mean is normally distributed. Under the assumption that  $H_0$  is true,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \stackrel{H_0}{\sim} N(\mu_0, \sigma_0^2/n),$$

where  $\stackrel{H_0}{\sim}$  means the “distribution under  $H_0$ ”. If we standardize the mean under  $H_0$ , we get a  $N(0, 1)$ -distributed test statistic

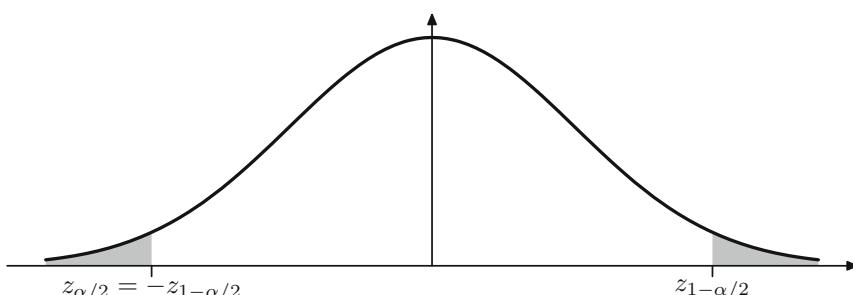
$$T(\mathbf{X}) = \frac{\bar{X} - \mu_0}{\sigma_0} \sqrt{n} \stackrel{H_0}{\sim} N(0, 1),$$

see also Theorem 7.3.2. Note that  $T(\mathbf{X})$  follows a normal distribution even if the  $X_i$ s are *not* normally distributed and if  $n$  is large enough which follows from the Central Limit Theorem (Appendix C.3). One can conclude that the distributional assumption from step 1 is thus particularly important for small samples, but not necessarily important for large samples. As a rule of thumb,  $n \geq 30$  is considered to be a large sample. This rule is based on the knowledge that a  $t$ -distribution with more than 30 degrees of freedom gets very close to a  $N(0, 1)$ -distribution.

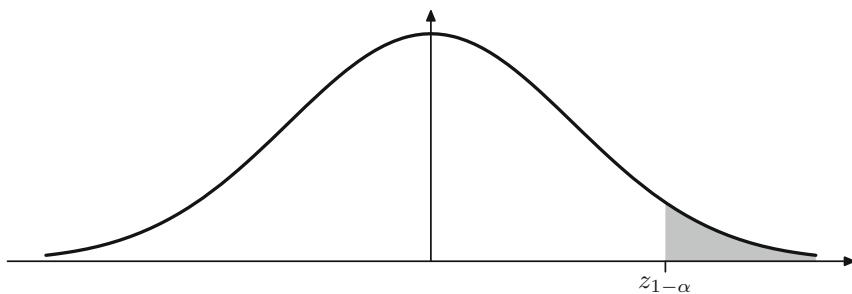
*5. Critical region:* Since the test statistic  $T(\mathbf{X})$  is  $N(0, 1)$ -distributed, we get the following critical regions, depending on the hypothesis:

Case	$H_0$	$H_1$	Critical region $K$
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (z_{1-\alpha}, \infty)$
(c)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, z_\alpha = -z_{1-\alpha})$

For case (a) with  $H_0: \mu = \mu_0$  and  $H_1: \mu \neq \mu_0$ , we are interested in extreme values of the test statistic on both tails: very small values and very large values of the test statistic give us evidence that  $H_0$  is wrong (because the statistic is mainly driven by the difference of the sample mean and the test value  $\mu_0$  for a fixed variance), see Fig. 10.1. In such a two-sided test, when the distribution of the test statistic is symmetric, we divide the critical region into two equal parts and assign each region of size  $\alpha/2$  to the left and right tails of the distribution. For  $\alpha = 0.05$ , 2.5 % of the most extreme values towards the right end of the distribution and 2.5 % of the most extreme values towards the left end of the distribution give us enough evidence that  $H_0$  is wrong and can be rejected and that  $H_1$  is accepted. It is also clear why  $\alpha$  is



**Fig. 10.1** Critical region of a two-sided one-sample Gauss-test  $H_0: \mu = \mu_0$  versus  $H_1: \mu \neq \mu_0$ . The critical region  $K = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$  has probability mass  $\alpha$  if  $H_0$  is true\*



**Fig. 10.2** Critical region of a one-sided one-sample Gauss test  $H_0: \mu \leq \mu_0$  versus  $H_1: \mu > \mu_0$ . The critical region  $K = (z_{1-\alpha}, \infty)$  has probability mass  $\alpha$  if  $H_0$  is true\*

the probability of a type I error: the most extreme values in the two tails together have 5 % probability and are just the probability that the test statistic falls into the critical region although  $H_0$  is true. Also, these areas are those which have the least probability of occurring if  $H_0$  is true. For  $\alpha = 0.05$ , we get  $z_{1-\frac{\alpha}{2}} = 1.96$ .

For case (b), only one direction is of interest. The critical region lies on the right tail of the distribution of the test statistic. A very large value of the test statistic has a low probability of occurrence if  $H_0$  is true. An illustration is given in Fig. 10.2: for  $\alpha = 0.05$ , we get  $z_{1-\alpha} = 1.64$  and any values greater than 1.64 are unlikely to be observed under  $H_0$ . Analogously, the critical region for case (c) is constructed. Here, the shaded area (critical region) is on the left-hand side. In this case, for  $\alpha = 0.05$ , we get  $z_\alpha = -z_{1-\alpha} = -1.64$ .

*6. Realization of the test statistic:* For an observed sample  $x_1, x_2, \dots, x_n$ , the arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

is used to calculate the realized (observed) test statistic  $t(x) = T(x_1, x_2, \dots, x_n)$  as

$$t(x) = \frac{\bar{x} - \mu_0}{\sigma_0} \sqrt{n}.$$

*7. Test decision:* If the realized test statistic from step 6 falls into the critical region,  $H_0$  is rejected (and therefore,  $H_1$  is statistically proven). Table 10.1 summarizes the test decisions depending on  $t(x)$  and the quantiles defining the appropriate critical regions.

*Example 10.3.1* A bakery supplies loaves of bread to supermarkets. The stated selling weight (and therefore the required minimum expected weight) is  $\mu = 2$  kg. However, not every package weighs exactly 2kg because there is variability in the weights. It is therefore important to find out if the average weight of the loaves

**Table 10.1** Rules to make test decisions for the one-sample Gauss test (and the two-sample Gauss test, the one-sample approximate binomial test, and the two-sample approximate binomial test—which are all discussed later in this chapter)

Case	$H_0$	$H_1$	Reject $H_0$ if
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$ t(x)  > z_{1-\alpha/2}$
(b)	$\mu \geq \mu_0$	$\mu < \mu_0$	$t(x) < z_\alpha$
(c)	$\mu \leq \mu_0$	$\mu > \mu_0$	$t(x) > z_{1-\alpha}$

is significantly smaller than 2 kg. The weight  $X$  (measured in kg) of the loaves is assumed to be normally distributed. We assume that the variance  $\sigma_0^2 = 0.1^2$  is known from experience. A supermarket draws a sample of  $n = 20$  loaves and weighs them. The average weight is calculated as  $\bar{x} = 1.97$  kg. Since the supermarket wants to be sure that the weights are, on average, not lower than 2 kg, a one-sided hypothesis is appropriate and is formulated as  $H_0: \mu \geq \mu_0 = 2$  kg versus  $H_1: \mu < \mu_0 = 2$  kg. The significance level is specified as  $\alpha = 0.05$ , and therefore,  $z_{1-\alpha} = 1.64$ . The test statistic is calculated as

$$t(x) = \frac{\bar{x} - \mu_0}{\sigma_0} \sqrt{n} = \frac{1.97 - 2}{0.1} \sqrt{20} = -1.34.$$

The null hypothesis is not rejected, since  $t(x) = -1.34 > -1.64 = -z_{1-0.05} = z_{0.05}$ .

*Interpretation:* The sample average  $\bar{x} = 1.97$  kg is below the target value of  $\mu = 2$  kg. But there is not enough evidence to reject the hypothesis that the sample comes from a  $N(2, 0.1^2)$ -distributed population. The probability to observe a sample of size  $n = 20$  with an average of at most 1.97 in a  $N(2, 0.1^2)$ -distributed population is greater than  $\alpha = 0.05 = 5\%$ . The difference between  $\bar{x} = 1.97$  kg and the target value  $\mu = 2$  kg is not statistically significant.

*Remark 10.3.1* The Gauss test assumes the variance to be known, which is often not the case in practice. The  $t$ -test (Sect. 10.3.2) assumes that the variance needs to be estimated. The  $t$ -test is therefore commonly employed when testing hypotheses about the mean. Its usage is outlined below. In R, the command `Gauss.test` from the library `compositions` offers an implementation of the Gauss test.

### 10.3.2 Test for the Mean When the Variance is Unknown (One-Sample $t$ -Test)

If the variance  $\sigma^2$  is unknown, hypotheses about the mean  $\mu$  of a normal random variable  $X \sim N(\mu, \sigma^2)$  can be tested in a similar way to the one-sample Gauss test. The difference is that the unknown variance is estimated from the sample. An

unbiased estimator of  $\sigma^2$  is the sample variance

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

The test statistic is therefore

$$T(\mathbf{X}) = \frac{\bar{X} - \mu_0}{S_X} \sqrt{n},$$

which follows a  $t$ -distribution with  $n-1$  degrees of freedom if  $H_0$  is true, as we know from Theorem 8.3.2.

### Critical regions and test decisions

Since  $T(\mathbf{X})$  follows a  $t$ -distribution under  $H_0$ , the critical regions refer to the regions of the  $t$ -distribution which are unlikely to be observed under  $H_0$ :

Case	$H_0$	$H_1$	Critical region $K$
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -t_{n-1;1-\alpha/2}) \cup (t_{n-1;1-\alpha/2}, \infty)$
(b)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, -t_{n-1;1-\alpha})$
(c)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (t_{n-1;1-\alpha}, \infty)$

The hypothesis  $H_0$  is rejected if the realized test statistic, i.e.

$$t(x) = \frac{\bar{x} - \mu_0}{S_X} \sqrt{n},$$

falls into the critical region. The critical regions are based on the appropriate quantiles of the  $t$ -distribution with  $(n-1)$  degrees of freedom, as outlined in Table 10.2.

*Example 10.3.2* We again consider Example 10.3.1. Now we assume that the variance of the loaves is unknown. Suppose a random sample of size  $n = 20$  has an arithmetic mean of  $\bar{x} = 1.9668$  and a sample variance of  $s^2 = 0.0927^2$ . We want to test whether this result contradicts the two-sided hypothesis  $H_0: \mu = 2$ , that is case (a). The significance level is fixed at  $\alpha = 0.05$ . For the realized test statistic  $t(x)$ , we calculate

$$t(x) = \frac{\bar{x} - \mu_0}{S_X} \sqrt{n} = \frac{1.9668 - 2}{0.0927} \sqrt{20} = -1.60.$$

**Table 10.2** Rules to make test decisions for the one-sample  $t$ -test (and the two-sample  $t$ -test, and the paired  $t$ -test, both explained below)

Case	$H_0$	$H_1$	Reject $H_0$ , if
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$ t(x)  > t_{n-1;1-\alpha/2}$
(b)	$\mu \geq \mu_0$	$\mu < \mu_0$	$t(x) < -t_{n-1;1-\alpha}$
(c)	$\mu \leq \mu_0$	$\mu > \mu_0$	$t(x) > t_{n-1;1-\alpha}$

$H_0$  is not rejected since  $|t| = 1.60 < 2.09 = t_{19,0.975}$ , where the quantiles  $\pm 2.09$  are defining the critical region (see Table C.2 or use R: `qt(0.975, 19)`). The same results can be obtained in R using the `t.test()` function, see Example 10.3.3 for more details. Or, we can directly calculate the (two-sided)  $p$ -value as

```
2*(1-pt(abs(1.6), df=19))
```



This yields a  $p$ -value of 0.1260951 which is not smaller than  $\alpha$ , and therefore,  $H_0$  is not rejected.

### 10.3.3 Comparing the Means of Two Independent Samples

In a two-sample problem, we may be interested in comparing the means of two *independent* samples. Assume that we have two samples of two normally distributed variables  $X \sim N(\mu_X, \sigma_X^2)$  and  $Y \sim N(\mu_Y, \sigma_Y^2)$  of size  $n_1$  and  $n_2$ , i.e.  $X_1, X_2, \dots, X_{n_1}$  are i.i.d. with the same distribution as  $X$  and  $Y_1, Y_2, \dots, Y_{n_2}$  are i.i.d. with the same distribution as  $Y$ . We can specify the following hypotheses:

Case	Null hypothesis	Alternative hypothesis	
(a)	$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	Two-sided test problem
(b)	$\mu_X \geq \mu_Y$	$\mu_X < \mu_Y$	One-sided test problem
(c)	$\mu_X \leq \mu_Y$	$\mu_X > \mu_Y$	One-sided test problem

We distinguish another three cases:

1.  $\sigma_X^2$  and  $\sigma_Y^2$  are known.
2.  $\sigma_X^2$  and  $\sigma_Y^2$  are unknown, but they are assumed to be equal, i.e.  $\sigma_X^2 = \sigma_Y^2$ .
3. Both  $\sigma_X^2$  and  $\sigma_Y^2$  are unknown and unequal ( $\sigma_X^2 \neq \sigma_Y^2$ ).

#### Case 1: The variances are known (two-sample Gauss test).

If the null hypothesis  $H_0: \mu_X = \mu_Y$  is true, then, using the usual rules for the normal distribution and the independence of the samples,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_1}\right),$$

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right),$$

and

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right).$$

It follows that the test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \quad (10.1)$$

follows a standard normal distribution,  $T(\mathbf{X}, \mathbf{Y}) \sim N(0, 1)$ . The realized test statistic is

$$t(x, y) = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}. \quad (10.2)$$

The test procedure is identical to the procedure of the one-sample Gauss test introduced in Sect. 10.3.1; that is, the test decision is based on Table 10.1.

### Case 2: The variances are unknown, but equal (two-sample $t$ -test).

We denote the unknown variance of both distributions as  $\sigma^2$  (i.e. both the populations are assumed to have variance  $\sigma^2$ ). We estimate  $\sigma^2$  by using the pooled sample variance where each sample is assigned weights relative to the sample size:

$$S^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}. \quad (10.3)$$

The test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \quad (10.4)$$

with  $S$  as in (10.3) follows a  $t$ -distribution with  $n_1 + n_2 - 2$  degrees of freedom if  $H_0$  is true. The realized test statistic is

$$t(x, y) = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}. \quad (10.5)$$

The test procedure is identical to the procedure of the one-sample  $t$ -test; that is, the test decision is based on Table 10.2.

### Case 3: The variances are unknown and unequal (Welch test).

We test  $H_0: \mu_X = \mu_Y$  versus  $H_1: \mu_X \neq \mu_Y$  given  $\sigma_X^2 \neq \sigma_Y^2$  and both  $\sigma_X^2$  and  $\sigma_Y^2$  are unknown. This problem is also known as the Behrens–Fisher problem and is the most frequently used test when comparing two means in practice. The test statistic can be written as

$$T(\mathbf{X}, \mathbf{Y}) = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}, \quad (10.6)$$

which is approximately  $t$ -distributed with  $v$  degrees of freedom:

$$v = \left( \frac{s_x^2}{n_1} + \frac{s_y^2}{n_2} \right)^2 / \left( \frac{(s_x^2/n_1)^2}{n_1 - 1} + \frac{(s_y^2/n_2)^2}{n_2 - 1} \right) \quad (10.7)$$

where  $s_x^2$  and  $s_y^2$  are the estimated values of  $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  and  $S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ , respectively. The test procedure, using the observed test statistic

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}}, \quad (10.8)$$

is identical to the procedure of the one-sample  $t$ -test; that is, the test decision is based on Table 10.2 except that the degrees of freedom are not  $n - 1$  but  $v$ . If  $v$  is not an integer, it can be rounded off to an integer value.

*Example 10.3.3* A small bakery sells cookies in packages of 500 g. The cookies are handmade and the packaging is either done by the baker himself or his wife. Some customers conjecture that the wife is more generous than the baker. One customer does an experiment: he buys packages of cookies packed by the baker and his wife on 16 different days and weighs the packages. He gets the following two samples (one for the baker, one for his wife).

Weight (wife) ( $X$ )	512	530	498	540	521	528	505	523
Weight (baker) ( $Y$ )	499	500	510	495	515	503	490	511

We want to test whether the complaint of the customers is justified. Let us start with the following simple hypotheses:

$$H_0 : \mu_x = \mu_y \text{ versus } H_1 : \mu_x \neq \mu_y,$$

i.e. we only want to test whether the weights are different, not that the wife is making heavier cookie packages. Since the variances are unknown, we assume that case 3 is the right choice. We calculate and obtain  $\bar{x} = 519.625$ ,  $\bar{y} = 502.875$ ,  $s_x^2 = 192.268$ , and  $s_y^2 = 73.554$ . The test statistic is:

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}} = \frac{|519.625 - 502.875|}{\sqrt{\frac{192.268}{8} + \frac{73.554}{8}}} \approx 2.91.$$

The degrees of freedom are:

$$v = \left( \frac{192.268}{8} + \frac{73.554}{8} \right)^2 / \left( \frac{(192.268/8)^2}{7} + \frac{(73.554/8)^2}{7} \right) \approx 11.67 \approx 12.$$

Since  $|t(x)| = 2.91 > 2.18 = t_{12;0.975}$ , it follows that  $H_0$  is rejected. Therefore,  $H_1$  is statistically significant. This means that the mean weight of the wife's packages is different from the mean weight of the baker's packages. Let us refine the hypothesis and try to find out whether the wife's packages have a higher mean weight. The hypotheses are now:

$$H_0 : \mu_x \leq \mu_y \text{ versus } H_1 : \mu_x > \mu_y.$$

The test statistic remains the same but the critical region and the degrees of freedom change. Thus,  $H_0$  is rejected if  $t(x, y) > t_{v;1-\alpha}$ . Using  $t_{v;1-\alpha} = t_{12;0.95} \approx 1.78$  and  $t(x, y) = 2.91$ , it follows that the null hypothesis can be rejected. The mean weight of the wife's packages is greater than the mean weight of the baker's packages.

In *R*, we would have obtained the same result using the `t.test` command:

```
x <- c(512,530,498,540,521,528,505,523)
y <- c(499,500,510,495,515,503,490,511)
t.test(x,y,alternative='greater')
```

**R**

Welch Two-Sample t-test

```
data: x and y
t = 2.9058, df = 11.672, p-value = 0.006762
alternative hypothesis: true difference in means is greater
than 0...
```

Note that we have to specify the *alternative* hypothesis under the option `alternative`. The output shows us the test statistic (2.9058), the degrees of freedom (11.672), the alternative hypothesis—but not the decision rule. We know that  $H_0$  is rejected if  $t(x, y) > t_{v;1-\alpha}$ , so the decision is easy in this case: we simply have to calculate  $t_{12;0.95}$  using `qt(0.95, 12)` in *R*. A simpler way to arrive at the same decision is to use the *p*-value. We know that  $H_0$  is rejected if  $p < \alpha$  which is the case in this example. It is also worthwhile mentioning that *R* displays the hypotheses slightly differently from ours: our alternative hypothesis is  $\mu_x > \mu_y$  which is identical to the statement  $\mu_x - \mu_y > 0$ , as shown by *R*, see also Sect. 10.2.2.

If we specify `two.sided` as an alternative (which is the default), a confidence interval for the mean *difference* is also part of the output:

```
t.test(x,y,alternative='two.sided')
```

**R**

```
...
95 % confidence interval:
4.151321 29.348679
```

It can be seen that the confidence interval of the difference does not cover the “0”. Therefore, the null hypothesis is rejected. This is the duality property referred to earlier in this chapter: the test decision is the same, no matter whether one evaluates (i) the confidence interval, (ii) the test statistic, or (iii) the *p*-value.

Any kind of *t*-test can be calculated with the `t.test` command: for example, the two-sample *t*-test requires to specify the option `var.equal=TRUE` while the Welch test is calculated when the (default) option `var.equal=FALSE` is set. We can also conduct a one-sample *t*-test. Suppose we are interested in whether the mean

weight of the wife's packages of cookies is greater than 500 g; then, we could test the hypotheses:

$$H_0 : \mu_x \leq 500 \text{ versus } H_1 : \mu_x > 500.$$

In R, we simply have to specify  $\mu_0$ :

```
t.test(x, mu=500, alternative='greater')
```



which gives us

```
One-Sample t-test
```

```
data: x
t = 4.0031, df = 7, p-value = 0.002585
alternative hypothesis: true mean is greater than 500
...
```

#### 10.3.4 Test for Comparing the Means of Two Dependent Samples (Paired $t$ -Test)

Suppose there are two dependent continuous random variables  $X$  and  $Y$  with  $E(X) = \mu_X$  and  $E(Y) = \mu_Y$ . They could be dependent because we measure the same variable twice on the same subjects at different times. Typically, this is the case in pre–post experiments, for example when we measure the weight of a person before starting a special diet and after finishing the diet; or when evaluating household expenditures on electronic appliances in two consecutive years. We then say that the samples are *paired*, or dependent. Since the same variable is measured twice on the same subject, it makes sense to calculate a difference between the two respective values. Let  $D = X - Y$  denote the random variable “difference of  $X$  and  $Y$ ”. If  $H_0: \mu_X = \mu_Y$  is true, then the expected difference is zero, and we get  $E(D) = \mu_D = 0$ . This means testing  $H_0: \mu_X = \mu_Y$  is identical to testing  $\mu_X - \mu_Y = \mu_D = 0$ . We further assume that  $D$  is normally distributed if  $H_0: \mu_X = \mu_Y$  is true (or equivalently if  $H_0: \mu_D = 0$  is true), i.e.  $D \sim N(0, \sigma_D^2)$ . For a random sample  $(D_1, D_2, \dots, D_n)$  of the differences, the test statistic

$$T(\mathbf{X}, \mathbf{Y}) = T(\mathbf{D}) = \frac{\bar{D}}{S_D} \sqrt{n} \quad (10.9)$$

is  $t$ -distributed with  $n - 1$  degrees of freedom. The sample mean is  $\bar{D} = \sum_{i=1}^n D_i / n$  and the sample variance is

$$S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}$$

which is an estimator of  $\sigma_D^2$ . The realized test statistic is thus

$$t(d) = \frac{\bar{d}}{s_d} \sqrt{n} \quad (10.10)$$

where  $\bar{d} = \sum_{i=1}^n d_i/n$  and  $s_d^2 = \sum_{i=1}^n (d_i - \bar{d})^2/n - 1$ .

The two-sided test  $H_0: \mu_D = 0$  versus  $H_1: \mu_D \neq 0$  and the one-sided tests  $H_0: \mu_D \leq 0$  versus  $H_1: \mu_D > 0$  or  $H_0: \mu_D \geq 0$  versus  $H_1: \mu_D < 0$  can be derived as in Sect. 10.3.2; that is, the test decision is based on Table 10.2. In fact, the paired  $t$ -test is a one-sample  $t$ -test on the differences of  $X$  and  $Y$ .

*Example 10.3.4* In an experiment,  $n = 10$  students have to solve different tasks before and after drinking a cup of coffee. Let  $Y$  and  $X$  denote the random variables “number of points before/after drinking a cup of coffee”. Assume that a higher number of points means that the student is performing better. Since the test is repeated on the same students, we have a paired sample. The data is given in the following table:

$i$	$y_i$ (before)	$x_i$ (after)	$d_i = x_i - y_i$	$(d_i - \bar{d})^2$
1	4	5	1	0
2	3	4	1	0
3	5	6	1	0
4	6	7	1	0
5	7	8	1	0
6	6	7	1	0
7	4	5	1	0
8	7	8	1	0
9	6	5	-1	4
10	2	5	3	4
Total			10	8

We calculate

$$\bar{d} = 1 \quad \text{and} \quad s_d^2 = \frac{8}{9} = 0.943^2,$$

respectively. For the realized test statistic  $t(d)$ , using  $\alpha = 0.05$ , we get

$$t(d) = \frac{1}{0.943} \sqrt{10} = 3.35 > t_{9;0.95} = 1.83,$$

such that  $H_0: \mu_X \leq \mu_Y$  is rejected and  $H_1: \mu_X > \mu_Y$  is accepted. We can conclude (for this example) that drinking coffee significantly increased the problem-solving capacity of the students.

In *R*, we would have obtained the same results using the `t.test` function and specifying the option `paired=TRUE`:

```
yp <- c(4,3,5,6,7,6,4,7,6,2)
xp <- c(5,4,6,7,8,7,5,8,5,5)
t.test(xp,yp,paired=TRUE)
```

```
Paired t-test

data: xp and yp
t = 3.3541, df = 9, p-value = 0.008468
alternative hypothesis: true difference in means != 0
95 % confidence interval:
0.325555 1.674445
sample estimates:
mean of the differences
1
```

We can make the test decision using the *R* output in three different ways:

- We compare the test statistic ( $t = -3.35$ ) with the critical value (1.83, obtained via `qt(0.95,9)`).
- We evaluate whether the *p*-value (0.008468) is smaller than the significance level  $\alpha = 0.05$ .
- We evaluate whether the confidence interval for the mean difference covers “0” or not.

## 10.4 Parametric Tests for Probabilities

### 10.4.1 One-Sample Binomial Test for the Probability $p$

#### Test construction and hypotheses.

Let  $X$  be a Bernoulli  $B(1; p)$  random variable with the two possible outcomes 1 and 0, which indicate occurrence and non-occurrence of an event of interest  $A$ . The probability for  $A$  in the population is  $p$ . From the sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  of independent  $B(1; p)$ -distributed random variables, we calculate the mean (relative frequency) as  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$  which is an unbiased estimate of  $p$ . The following hypotheses may thus be of interest:

Case	Null hypothesis	Alternative hypothesis	
(a)	$p = p_0$	$p \neq p_0$	Two-sided problem
(b)	$p \geq p_0$	$p < p_0$	One-sided problem
(c)	$p \leq p_0$	$p > p_0$	One-sided problem

In the following, we describe two possible solutions, one exact approach and an approximate solution. The approximate solution is based on the approximation of the binomial distribution by the normal distribution, which is appropriate if  $n$  is sufficiently large and the condition  $np(1 - p) \geq 9$  holds (i.e.  $p$  is neither too small nor too large). First, we present the approximate solution and then the exact one.

### Test statistic and test decisions.

- (a) **Approximate binomial test.** We define the standardized test statistic as

$$T(\mathbf{X}) = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n}. \quad (10.11)$$

It holds approximately that  $T(\mathbf{X}) \sim N(0, 1)$ , given that the conditions that (i)  $n$  is sufficiently large and (ii)  $np(1 - p) \geq 9$  are satisfied. The test can then be conducted along the lines of the Gauss test in Sect. 10.3.1; that is, the test decision is based on Table 10.1.

*Example 10.4.1* We return to Example 10.1.1. Let us assume that a representative sample of size  $n = 2000$  has been drawn from the population of eligible voters, from which 700 (35 %) have voted for the party of interest  $P$ . The research hypothesis (which has to be stated as  $H_1$ ) is that more than 30 % (i.e.  $p_0 = 0.3$ ) of the eligible voters cast their votes for party  $P$ . The sample is in favour of  $H_1$  because  $\hat{p} = 35\%$ , but to draw conclusions for the proportion of voters of party  $P$  in the population, we have to conduct a binomial test. Since  $n$  is large and  $np(1 - p) = 2000 \cdot 0.35 \cdot 0.65 = 455 \geq 9$ , the assumptions for the use of the test statistic (10.11) are satisfied.

We can write down the realized test statistic as

$$t(x) = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} = \frac{0.35 - 0.3}{\sqrt{0.3(1 - 0.3)}} \sqrt{2000} = 4.8795.$$

Using  $\alpha = 0.05$ , it follows that  $T(X) = 4.8795 > z_{1-\alpha} = 1.64$ , and thus, the null hypothesis  $H_0 : p \leq 0.3$  can be rejected. Therefore,  $H_1 : p > 0.3$  is statistically significant; that is, the proportion of votes for party  $P$  is greater than 30%.

- (b) The **exact binomial test** can be constructed using the knowledge that under  $H_0$ ,  $Y = \sum_{i=1}^n X_i$  (i.e. the number of successes) follows a binomial distribution. In fact, we can use  $Y$  directly as the test statistic:

$$T(\mathbf{X}) = Y \sim B(n, p_0).$$

The observed test statistic is  $t(x) = \sum_i x_i$ . For the two-sided case (a), the two critical numbers  $c_l$  and  $c_r$  ( $c_l < c_r$ ) which define the critical region, have to be found such that

$$P_{H_0}(Y \leq c_l) \leq \frac{\alpha}{2} \quad \text{and} \quad P_{H_0}(Y \geq c_r) \leq \frac{\alpha}{2}.$$

The null hypothesis is rejected if the test statistic, i.e.  $Y$ , is greater than or equal to  $c_r$  or less than or equal to  $c_l$ . For the one-sided case, a critical number  $c$  has to be found such that

$$P_{H_0}(Y \leq c) \leq \alpha$$

for hypotheses of type (b) and

$$P_{H_0}(Y \geq c) \leq \alpha$$

for hypotheses of type (c). If  $Y$  is less than the critical value  $c$  (for case (b)) or greater than the critical value (for case (c)), the null hypothesis is rejected.

*Example 10.4.2* We consider again Example 10.1.1 where we looked at the population of eligible voters, from which 700 (35 %) have voted for the party of interest  $P$ . The observed test statistic is  $t(x) = \sum_i x_i = 700$  and the alternative hypothesis is  $H_1 : p \geq 0.3$ , as in case (c). There are at least two ways in which we can obtain the results:

- (i) *Long way*: We can calculate the test statistic and compare it to the critical region. To get the critical region, we search  $c$  such that

$$P_{p=0.3}(Y \geq c) \leq 0.05 ,$$

which equates to

$$P_{p=0.3}(Y < c) \geq 0.95$$

and can be calculated in  $R$  as:

```
qbinom(p=0.95, prob=0.3, size=2000)
[1] 634
```



Since  $Y = 700 > c = 634$  we reject the null hypothesis. As in Example 10.4.1, we conclude that there is enough evidence that the proportion of votes for party  $P$  is greater than 30 %.

- (ii) *Short way*: The above result can be easily obtained in  $R$  using the `binom.test()` command. We need to specify the number of “successes” (here: 700), the number of “failures” ( $2000 - 700 = 1300$ ), and the alternative hypothesis:

```
binom.test(c(700,1300),p=0.3,alternative='greater')
```



```
data: c(700, 1300)
number of successes = 700, number of trials = 2000,
p-value = 8.395e-07
alternative hypothesis: true probability of success
is greater than 0.3
95 % confidence interval:
0.332378 1.000000
probability of success
0.35
```

Both the  $p$ -value (which is smaller than  $\alpha = 0.05$ ) and the confidence interval (for which we do not show the calculation) confirm the rejection of the null hypothesis.

Note that

```
binom.test(x=700,n=2000,p=0.3, alternative='greater')
```

returns the same result.

### 10.4.2 Two-Sample Binomial Test

#### Test construction and hypotheses.

We consider now the case of two independent i.i.d. samples from Bernoulli distributions with parameters  $p_1$  and  $p_2$ .

$$\begin{aligned}\mathbf{X} &= (X_1, X_2, \dots, X_{n_1}), \quad X_i \sim B(1; p_1) \\ \mathbf{Y} &= (Y_1, Y_2, \dots, Y_{n_2}), \quad Y_i \sim B(1; p_2).\end{aligned}$$

The sums

$$X = \sum_{i=1}^{n_1} X_i \sim B(n_1; p_1), \quad Y = \sum_{i=1}^{n_2} Y_i \sim B(n_2; p_2)$$

follow binomial distributions. One of the following hypotheses may be of interest:

Case	Null hypothesis	Alternative hypothesis	
(a)	$p_1 = p_2$	$p_1 \neq p_2$	Two-sided problem
(b)	$p_1 \geq p_2$	$p_1 < p_2$	One-sided problem
(c)	$p_1 \leq p_2$	$p_1 > p_2$	One-sided problem

Similar to the one-sample case, both exact and approximate tests exist. Here, we only present the approximate test. The **exact test of Fisher** is presented in Appendix C.5, p. 428. Let  $n_1$  and  $n_2$  denote the sample sizes. Then,  $X/n_1$  and  $Y/n_2$  are approximately normally distributed:

$$\begin{aligned}\frac{X}{n_1} &\stackrel{\text{approx.}}{\sim} N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \\ \frac{Y}{n_2} &\stackrel{\text{approx.}}{\sim} N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).\end{aligned}$$

Their difference  $D$

$$D \stackrel{\text{approx.}}{\sim} N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

is normally distributed too under  $H_0$  (given  $p = p_1 = p_2$  holds). Since the probabilities  $p_1$  and  $p_2$  are identical under  $H_0$ , we can pool the two samples and estimate  $p$  by

$$\hat{p} = \frac{X + Y}{n_1 + n_2}. \quad (10.12)$$

### Test statistic and test decision.

The test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{D}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}, \quad (10.13)$$

follows a  $N(0, 1)$ -distribution if  $n_1$  and  $n_2$  are sufficiently large and  $p$  is not near the boundaries 0 and 1 (one could use, for example, again the condition  $np(1 - p) > 9$  with  $n = n_1 + n_2$ ). The realized test statistic can be calculated using the observed difference  $\hat{d} = \hat{p}_1 - \hat{p}_2$ . The test can be conducted for the one-sided and the two-sided case as the Gauss test introduced in Sect. 10.3.1; that is, the decision rules from Table 10.1 can be applied.

*Example 10.4.3* Two competing lotteries claim that every fourth lottery ticket wins. Suppose we want to test whether the probabilities of winning are different for the two lotteries, i.e.  $H_0 : p_1 = p_2$  and  $H_1 : p_1 \neq p_2$ . We have the following data

	$n$	Winning	Not winning
Lottery A	63	14	49
Lottery B	45	13	32

We can estimate the probabilities of a winning ticket for each lottery, as well as the respective difference, as

$$\hat{p}_A = \frac{14}{63}, \quad \hat{p}_B = \frac{13}{45}, \quad \hat{d} = \hat{p}_A - \hat{p}_B = -\frac{1}{15}.$$

Under  $H_0$ , an estimate for  $p$  following (10.12) is

$$\hat{p} = \frac{14 + 13}{63 + 45} = \frac{27}{108} = 0.25.$$

The test statistic can be calculated as

$$t(x, y) = \frac{-\frac{1}{15}}{\sqrt{0.25(1 - 0.25) \left( \frac{1}{63} + \frac{1}{45} \right)}} = -0.79.$$

$H_0$  is not rejected since  $|t(x, y)| = 0.79 < 1.96 = z_{1-0.05/2}$ . Thus, there is no statistical evidence for different winning probabilities for the two lotteries. These hypotheses can be tested in *R* using the Test of Fisher, see Appendix C.5, p. 428, for more details.

## 10.5 Tests for Scale Parameters

There are various tests available to test hypotheses about scale parameters. Such tests are useful when one is interested in the dispersion of a variable, for example in quality control where the variability of a process may be of interest. One-sample tests of hypotheses for the variance of a normal distribution, e.g. hypotheses such as  $H_0 : \sigma^2 = \sigma_0^2$ , can be tested by the  $\chi^2$ -test for the variance, see Appendix C.5, p. 430. Two-sample problems can be addressed by the  $F$ -test (which is explained in Appendix C.5, p. 431); or by other tests such as the Levene test or Bartlett's test, which are also available in *R* (`leveneTest` in the package `car`, `bartlett` in the base distribution of *R*).

---

## 10.6 Wilcoxon–Mann–Whitney (WMW) U-Test

### Test construction and hypotheses.

The WMW  $U$ -test is often proposed as an alternative to the  $t$ -test because it also focuses on location but not on the expected value  $\mu$ . It is a *nonparametric* test and useful in situations where skewed distributions are compared with each other. We consider two independent random samples  $\mathbf{X} = (X_1, X_2, \dots, X_{n_1})$  and  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n_2})$  from two populations with observed values  $(x_1, x_2, \dots, x_{n_1})$  and  $(y_1, y_2, \dots, y_{n_2})$ , respectively. In this case, the null hypothesis  $H_0$  considering the location can be formulated as

$$H_0 : P(X > Y) = P(Y > X) = \frac{1}{2}.$$

The null hypothesis can be interpreted in the following way: the probability that a randomly drawn observation from the first population has a value  $x$  that is greater (or lower) than the value  $y$  of a randomly drawn subject from the second population is  $\frac{1}{2}$ . The alternative hypothesis  $H_1$  is then

$$H_1 : P(X > Y) \neq P(Y > X).$$

This means we are comparing the entire distribution of two variables. If there is a location shift in the sense that one distribution is shifted left (or right) compared with the other distribution, the null hypothesis will be rejected because this shift can be seen as part of the alternative hypothesis  $P(X > Y) \neq P(Y > X)$ . In fact, under some assumptions, the hypothesis can even be interpreted as comparing two medians, and this is what is often done in practice.

### Observed test statistic.

To construct the test statistic, it is necessary to merge  $(x_1, x_2, \dots, x_{n_1})$  and  $(y_1, y_2, \dots, y_{n_2})$  into one sorted sample, usually in ascending order, while keeping the information which value belongs to which sample. For now, we assume that all values of the two samples are distinct; that is, no ties are present. Then, each observation has

a rank between 1 and  $(n_1 + n_2)$ . Let  $R_{1+}$  be the sum of ranks of the  $x$ -sample and let  $R_{2+}$  be the sum of ranks of the  $y$ -sample. The test statistic is defined as  $U$ , where  $U$  is the minimum of the two values  $U_1, U_2$ ,  $U = \min(U_1, U_2)$  with

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_{1+}, \quad (10.14)$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2 + 1)}{2} - R_{2+}. \quad (10.15)$$

### Test decision.

$H_0$  is rejected if  $U < u_{n_1, n_2; \alpha}$ . Here,  $u_{n_1, n_2; \alpha}$  is the critical value derived from the distribution of  $U$  under the null hypothesis. The exact (complex) distribution can, for example, be derived computationally (in  $R$ ). We are presenting an approximate solution together with its implementation in  $R$ .

Since  $U_1 + U_2 = n_1 \cdot n_2$ , it is sufficient to compute only  $R_{i+}$  and  $U = \min\{U_i, n_1 n_2 - U_i\}$  ( $i = 1$  or  $i = 2$  are chosen such that  $R_{i+}$  is calculated for the sample with the lower sample size). For  $n_1, n_2 \geq 8$ , one can use the approximation

$$T(\mathbf{X}, \mathbf{Y}) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} \stackrel{\text{approx.}}{\sim} N(0, 1) \quad (10.16)$$

as the test statistic. For two-sided hypotheses,  $H_0$  is rejected if  $|t(x, y)| > z_{1-\alpha/2}$ ; for one-sided hypotheses  $H_0$  is rejected if  $|t(x, y)| > z_{1-\alpha}$ . In the case of ties, the denominator of the test statistic in (10.16) can be modified as

$$T(\mathbf{X}, \mathbf{Y}) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\left[ \frac{n_1 \cdot n_2}{n(n-1)} \right] \left[ \frac{n^3 - n}{12} - \sum_{j=1}^G \frac{t_j^3 - t_j}{12} \right]}} \stackrel{\text{approx.}}{\sim} N(0, 1),$$

where  $G$  is the number of different (groups of) ties and  $t_j$  denotes the number of tied ranks in tie group  $j$ .

*Example 10.6.1* In a study, the reaction times (in seconds) to a stimulus were measured for two groups. One group drank a strong coffee before the stimulus and the other group drank only the same amount of water. There were 9 study participants in the coffee group and 10 participants in the water group. The following reaction times were recorded:

Reaction time	1	2	3	4	5	6	7	8	9	10
Coffee group (C)	3.7	4.9	5.2	6.3	7.4	4.4	5.3	1.7	2.9	
Water group (W)	4.5	5.1	6.2	7.3	8.7	4.2	3.3	8.9	2.6	4.8

We test with the  $U$ -test whether there is a location difference between the two groups. First, the ranks of the combined sample are calculated as:

	1	2	3	4	5	6	7	8	9	10	Total
Value (C)	3.7	4.9	5.2	6.3	7.4	4.4	5.3	1.7	2.9		
Rank (C)	5	10	12	15	17	7	13	1	3		83
Value (W)	4.5	5.1	6.2	7.3	8.7	4.2	3.3	8.9	2.6	4.8	
Rank (W)	8	11	14	16	18	6	4	19	2	9	107

With  $R_{C+} = 83$  and  $R_{W+} = 107$ , we get

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_{C+} = 9 \cdot 10 + \frac{9 \cdot 10}{2} - 83 = 52,$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2 + 1)}{2} - R_{W+} = 9 \cdot 10 + \frac{10 \cdot 11}{2} - 107 = 38.$$

With  $n_1, n_2 \geq 8$  and  $U = U_2 = 38$ ,

$$t(x, y) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} = \frac{38 - \frac{9 \cdot 10}{2}}{\sqrt{\frac{9 \cdot 10 \cdot (9 + 10 + 1)}{12}}} \approx -0.572.$$

Since  $|t(x, y)| = 0.572 < z_{1-\alpha/2} = 1.96$ , the null hypothesis cannot be rejected; that is, there is no statistical evidence that the two groups have different reaction times.

In *R*, one can use the `wilcox.test` command to obtain the results:

```
coffee <- c(3.7, 4.9, 5.2, 6.3, ..., 2.9)
water <- c(4.5, 5.1, 6.2, ..., 4.8)
wilcox.test(coffee, water)
```



The output is

```
Wilcoxon rank sum test

data: coffee.sample and water.sample
W = 38, p-value = 0.6038
alternative hypothesis: true location shift is not equal to 0
```

We can see that the null hypothesis is not rejected because  $p = 0.6038 > \alpha = 0.05$ . The displayed test statistic is  $W$  which equates to our statistic  $U_2$ . The alternative hypothesis in *R* is framed as location shift, an interpretation which has already been given earlier in the chapter. Note that the test also suggests that the medians of the two samples are not statistically different.

## 10.7 $\chi^2$ -Goodness-of-Fit Test

### Test construction.

The  $\chi^2$ -goodness-of-fit test is one of the most popular tests for testing the goodness of fit of the observed data to a distribution. The construction principle is very general and can be used for variables of any scale. The test statistic is derived such that the *observed* absolute frequencies are compared with the *expected* absolute frequencies *under the null hypothesis*  $H_0$ .

*Example 10.7.1* Consider an experiment where a die is rolled  $n = 60$  times. Under the null hypothesis  $H_0$ , we assume that the die is fair, i.e.  $p_i = \frac{1}{6}$ ,  $i = 1, 2, \dots, 6$ , where  $p_i = P(X = i)$ . We could have also said that  $H_0$  is the hypothesis that the rolls are following a discrete uniform distribution. Thus, the expected absolute frequencies under  $H_0$  are  $np_i = 60 \cdot \frac{1}{6} = 10$ , while the observed frequencies in the sample are  $N_i$ ,  $i = 1, 2, \dots, 6$ . The  $N_i$  generally deviate from  $np_i$ . The  $\chi^2$ -statistic is based on the squared differences,  $\sum_{i=1}^6 (N_i - np_i)^2$ , and becomes large as the differences between the observed and the expected frequencies become larger. The  $\chi^2$ -test statistic is a modification of this sum by scaling each squared difference by the expected frequencies,  $np_i$ , and is explained below.

With a nominal variable, we can proceed as in Example 10.7.1. If the scale of the variable is ordinal or continuous, the number of different values can be large. Note that in the most extreme case, we can have as many different values as observations ( $n$ ), leading to  $N_i = 1$  for all  $i = 1, 2, \dots, n$ . Then, it is necessary to group the data into  $k$  intervals before applying the  $\chi^2$ -test. The reason is that the general theory of the  $\chi^2$ -test assumes that the number  $k$  (which was 6 in Example 10.7.1 above) is fixed and does not grow with the number of observations  $n$ ; that is, the theory says that the  $\chi^2$ -test only works properly if  $k$  is fixed and  $n$  is large. For this reason, we group the sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  into  $k$  classes as shown in Sect. 2.1.

Class	1	2	$\dots$	$k$	Total
Number of observations	$n_1$	$n_2$	$\dots$	$n_k$	$n$

The choice of the class intervals is somewhat arbitrary. As a rule of thumb  $np_i > 5$  should hold for most class intervals. The general hypotheses can be formulated in the form of distribution functions:

$$H_0 : F(x) = F_0(x) \text{ versus } H_1 : F(x) \neq F_0(x).$$

### Test statistic.

The test statistic is defined as

$$T(\mathbf{X}) = t(x) = \chi^2 = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}. \quad (10.17)$$

Here,

- $N_i$  ( $i = 1, 2, \dots, k$ ) are the absolute frequencies of observations of the sample  $\mathbf{X}$  in class  $i$ ,  $N_i$  is a random variable with realization  $n_i$  in the observed sample;
- $p_i$  ( $i = 1, 2, \dots, k$ ) are calculated from the distribution under  $H_0$ ,  $F_0(x)$ , and are the (hypothetical) probabilities that an observation of  $X$  falls in class  $i$ ;
- $np_i$  are the expected absolute frequencies in class  $i$  under  $H_0$ .

### Test decision.

For a significance level  $\alpha$ ,  $H_0$  is rejected if  $t(x)$  is greater than the  $(1 - \alpha)$ -quantile of the  $\chi^2$ -distribution with  $k - 1 - r$  degrees of freedom, i.e. if

$$t(x) = \chi^2 > c_{k-1-r, 1-\alpha}.$$

Note that  $r$  is the number of parameters of  $F_0(x)$ , if these parameters are estimated from the sample. The  $\chi^2$ -test statistic is only asymptotically  $\chi^2$ -distributed under  $H_0$ .

*Example 10.7.2* Let  $F_0(x)$  be the distribution function of the test distribution. If one specifies a normal distribution such as  $F_0(x) = N(3, 10)$ , or a discrete uniform distribution with  $p_i = 0.25$  ( $i = 1, 2, 3, 4$ ), then  $r = 0$ , since no parameters have to be estimated from the data. Otherwise, if we simply want to test whether the data is generated from a normal distribution  $N(\mu, \sigma^2)$  or the data follows a normal distribution  $N(\mu, \sigma^2)$ , then  $\mu$  and  $\sigma^2$  may be estimated from the sample by  $\bar{x}$  and  $s^2$ . Then,  $r = 2$  and the number of degrees of freedom is reduced.

*Example 10.7.3* Gregor Mendel (1822–1884) conducted crossing experiments with pea plants of different shape and colour. Let us look at the outcome of a pea crossing experiment with the following results:

Crossing result	Round Yellow	Round Green	Edged Yellow	Edged Green
Observations	315	108	101	32

Mendel had the hypothesis that the four different types occur in proportions of 9:3:3:1, that is

$$p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}.$$

The hypotheses are

$$H_0 : P(X = i) = p_i \quad \text{versus} \quad H_1 : P(X = i) \neq p_i, \quad i = 1, 2, 3, 4.$$

With  $n = 556$  observations, the test statistic can be calculated from the following observed and expected frequencies:

$i$	$N_i$	$p_i$	$n p_i$
1	315	$\frac{9}{16}$	312.75
2	108	$\frac{3}{16}$	104.25
3	101	$\frac{3}{16}$	104.25
4	32	$\frac{1}{16}$	34.75

The  $\chi^2$ -test statistic is calculated as

$$t(x) = \chi^2 = \frac{(315 - 312.75)^2}{312.75} + \cdots + \frac{(32 - 34.75)^2}{34.75} = 0.47.$$

Since  $\chi^2 = 0.47 < 7.815 = \chi^2_{0.95,3} = c_{0.95,3}$ , the null hypothesis is not rejected. Statistically, there is no evidence that Mendel was wrong with his 9:3:3:1 assumption. In R, the test can be conducted by applying the `chisq.test` command:

```
chisq.test(c(315, 108, 101, 32),
p=c(9/16,3/16,3/16,1/16))
qchisq(df=3, p=0.95)
```



which leads to the following output

```
Chi-squared test for given probabilities

data: c(315, 108, 101, 32)
X-squared = 0.47, df = 3, p-value = 0.9254
```

and the critical value is

```
[1] 7.814728
```

*Remark 10.7.1* In this example, the data was already summarized in a frequency table. For raw data, the `table` command can be used to preprocess the data, i.e. we can use `chisq.test(table(var1, var2))`.

Another popular goodness-of-fit test is the test of Kolmogorov–Smirnov. There are two different versions of this test, one for the one-sample scenario and one for the two-sample scenario. The null hypothesis for the latter is that the two independent samples come from the same distribution. In R, the command `ks.test()` can be used to perform Kolmogorov–Smirnov tests.

## 10.8 $\chi^2$ -Independence Test and Other $\chi^2$ -Tests

In Chap. 4, we introduced different methods to describe the association between two variables. Several association measures are possibly suitable if the variables are categorical, for example Cramer's  $V$ , Goodman's and Kruskal's  $\gamma$ , Spearman's rank correlation coefficient, and the odds ratio. If we are not interested in the strength of association but rather in finding out whether there is an association at all, one can use the  $\chi^2$ -independence test.

### Test construction.

In the following we assume that we observe a sample from a bivariate discrete distribution of two variables  $X$  and  $Y$  which can be summarized in a contingency table with absolute frequencies  $n_{ij}$ , ( $i = 1, 2, \dots, I$ ;  $j = 1, 2, \dots, J$ ):

		$Y$				
		1	2	...	$J$	
$X$	1	$n_{11}$	$n_{12}$	$\cdots$	$n_{1J}$	$n_{1+}$
	2	$n_{21}$	$n_{22}$	$\cdots$	$n_{2J}$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$I$	$n_{I1}$	$n_{I2}$	$\cdots$	$n_{IJ}$	$n_{I+}$	
	$n_{+1}$	$n_{+2}$	$\cdots$	$n_{+J}$	$n$	

Remember that

$n_{i+}$  is the  $i$ th row sum,

$n_{+j}$  is the  $j$ th column sum, and

$n$  is the total number of observations.

The hypotheses are  $H_0$ :  $X$  and  $Y$  are independent versus  $H_1$ :  $X$  and  $Y$  are not independent. If  $X$  and  $Y$  are independent, then the expected frequencies  $m_{ij}$  are

$$\hat{m}_{ij} = n\hat{\pi}_{ij} = \frac{n_{i+}n_{+j}}{n}. \quad (10.18)$$

### Test statistic.

Pearson's  $\chi^2$ -test statistic was introduced in Chap. 4, Eq. (4.6). It is

$$T(\mathbf{X}, \mathbf{Y}) = \chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - m_{ij})^2}{m_{ij}},$$

where  $m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$  (expected absolute cell frequencies under  $H_0$ ). Strictly speaking,  $m_{ij}$  are the true, unknown expected frequencies under  $H_0$  and are estimated by  $\hat{m}_{ij} = n\pi_{i+}\pi_{+j}$ , such that the realized test statistic equates to

$$t(x, y) = \chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \quad (10.19)$$

**Test decision.**

The number of degrees of freedom under  $H_0$  is  $(I - 1)(J - 1)$ , where  $I - 1$  are the parameters which have to be estimated for the marginal distribution of  $X$ , and  $J - 1$  are the number of parameters for the marginal distribution of  $Y$ . The test decision is:

$$\text{Reject } H_0, \text{ if } t(x, y) = \chi^2 > c_{(I-1)(J-1); 1-\alpha}.$$

Note that the alternative hypothesis  $H_1$  is very general. If  $H_0$  is rejected, nothing can be said about the structure of the dependence of  $X$  and  $Y$  from the  $\chi^2$ -value itself.

*Example 10.8.1* Consider the following contingency table. Here,  $X$  describes the educational level (1: primary, 2: secondary, 3: tertiary) and  $Y$  the preference for a specific political party (1: Party A, 2: Party B, 3: Party C). Our null hypothesis is that the two variables are independent, and we want to show the alternative hypothesis which says that there is a relationship between them.

		Y			Total
		1	2	3	
X	1	100	200	300	600
	2	100	100	100	300
	3	80	10	10	100
Total		280	310	410	1000

For the (estimated) expected frequencies  $\hat{m}_{ij} = \frac{n_i + n_j}{n}$ , we get

		Y		
		1	2	3
X	1	168	186	246
	2	84	93	123
	3	28	31	41

For example:  $\hat{m}_{11} = 600 \cdot 280 / 1000 = 168$ . The test statistic is

$$\begin{aligned} t(x, y) &= \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \\ &= \frac{(100 - 168)^2}{168} + \dots + \frac{(10 - 41)^2}{41} \approx 182.54. \end{aligned}$$

Since  $\chi^2_{4;0.95} = 9.49 < t(x, y) = 182.54$ ,  $H_0$  is rejected.

In *R*, either the summarized data (as shown below) can be used to calculate the test statistic or the raw data (summarized in a contingency table via `table(var1, var2)`):

```
ct <- matrix(nrow=3, ncol=3, byrow=T,
  data=c(100, 200, 300, 100, 100, 100, 80, 10, 10))
chisq.test(ct)
qchisq(df=(3-1)*(3-1), p=0.95)
```



The output is

```
Pearson's Chi-squared test

data: contingency.table
X-squared = 182.5428, df = 4, p-value < 2.2e-16
```

with the critical value

```
[1] 9.487729
```

which confirms our earlier manual calculations. The *p*-value is smaller than  $\alpha = 0.05$  which further confirms that the null hypothesis has to be rejected.

For a binary outcome, the  $\chi^2$ -test of independence can be formulated as a test for the null hypothesis that the proportions of the binary variable are equal in several ( $\geq 2$ ) groups, i.e. for a  $K \times 2$  (or  $2 \times K$ ) table. This test is called the  **$\chi^2$ -test of homogeneity**.

*Example 10.8.2* Consider two variables *X* and *Y*, where *X* is describing the rating of a coffee brand with the categories “bad taste” and “good taste” and *Y* denotes three age subgroups, e.g. “18–25”, “25–35”, and “35–45”. The observed data is

		<i>Y</i>			Total
		18–25	25–35	35–45	
<i>X</i>	Bad	10	30	65	105
	Good	90	70	35	195
Total	100	100	100	300	

Assume  $H_0$  is the hypothesis that the probabilities  $P(X = \text{'good'}|Y = \text{'18–25'})$ ,  $P(X = \text{'good'}|Y = \text{'25–35'})$ , and  $P(X = \text{'good'}|Y = \text{'35–45'})$  are all equal. Then, we can use the function either `prop.test` or `chisq.test` in *R* to test this hypothesis:

```
prop.test(x=rbind(c(10,30,65), c(90,70,35) ))
chisq.test(x=rbind(c(10,30,65), c(90,70,35) ))
```



This produces the following outputs:

```
3-sample test for equality of proportions

data: cbind(c(10, 30, 65), c(90, 70, 35))
X-squared = 68.1319, df = 2, p-value = 1.605e-15
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3
0.10    0.30    0.65
```

and

```
Pearson's Chi-squared test

data: cbind(c(10, 30, 65), c(90, 70, 35))
X-squared = 68.1319, df = 2, p-value = 1.605e-15
```

The results (test statistic,  $p$ -value) are identical and  $H_0$  is rejected. Note that `prop.test` strictly expects a  $K \times 2$  table (i.e. exactly 2 columns).

*Remark 10.8.1* For  $2 \times 2$ -tables with small sample sizes and therefore small cell frequencies, it is recommended to use the exact test of Fisher as described in Appendix C.5.

*Remark 10.8.2* The test described in Example 10.8.2 is a special case (since one variable is binary) of the general  $\chi^2$ -test of homogeneity. The  $\chi^2$ -test of homogeneity is valid for any  $K \times C$  table, where  $K$  is the number of subgroups of a variable  $Y$  and  $C$  is the number of values of the outcome  $X$  of interest. The null hypothesis  $H_0$  assumes that the conditional distributions of  $X$  given  $Y$  are identical in all subgroups, i.e.

$$P(X = x_c | Y = y_k) = P(X = x_c | Y = y_{k'})$$

forall  $c = 1, 2, \dots, C; k, k' = 1, 2, \dots, K, k \neq k'$ . Again, the usual  $\chi^2$ -test statistic can be used.

## 10.9 Key Points and Further Issues

### Note:

- ✓ A graphical summary on when to use the tests introduced in this chapter is given in Appendices D.2 and D.3.
- ✓ To arrive at a test decision, i.e. accept  $H_0$  or reject it, it does not matter whether one compares the test statistic to the critical region, one uses the  $p$ -value obtained from statistical software, or one evaluates the appropriate confidence interval. However, it is important not to misinterpret the  $p$ -value (see Sect. 10.2.6) and to choose the correct confidence interval.
- ✓ There is a difference between relevance and significance. A test might be significant, but the point estimate of the quantity of interest may not be relevant from a substantive point of view. Similarly, a test might not be significant, but the point and interval estimates may still yield relevant conclusions.
- ✓ The test statistic of the  $t$ -test (one-sample, two-sample, paired) is *asymptotically* normally distributed. This means that for relatively large  $n$  (as a rule of thumb  $> 30$  per group) the sample does not need to come from a normal distribution. However, the application of the  $t$ -test makes sense only when the expectation  $\mu$  can be interpreted meaningfully; this may not be the case for skewed distributions or distributions with outliers.

---

## 10.10 Exercises

*Exercise 10.1* Two people, A and B, are suspects for having committed a crime together. Both of them are interrogated in separate rooms. The jail sentence depends on who confesses to have committed the crime, and who does not:

	B does not confess	B does confess
A does not confess	Each serves 1 year	A: 3 years; B: goes free
A does confess	A: goes free; B: 3 years	Each serves 2 years

A has two hypotheses:

$$H_0 : \text{B does not confess} \quad \text{versus} \quad H_1 : \text{B does confess.}$$

Given the possible sentences he decides to not confess if  $H_0$  is true and to confess otherwise. Explain the concepts of type I error and type II error for this situation. Comment on the consequences if these errors are made.

*Exercise 10.2* A producer of chocolate bars hypothesizes that his production does not adhere to the weight standard of 100 g. As a measure of quality control, he weighs 15 bars and obtains the following results in grams:

96.40, 97.64, 98.48, 97.67, 100.11, 95.29, 99.80, 98.80, 100.53, 99.41, 97.64,  
101.11, 93.43, 96.99, 97.92

It is assumed that the production process is standardized in the sense that the variation is controlled to be  $\sigma = 2$ .

- (a) What are the hypotheses regarding the expected weight  $\mu$  for a two-sided test?
- (b) Which test should be used to test these hypotheses?
- (c) Conduct the test that was suggested to be used in (b). Use  $\alpha = 0.05$ .
- (d) The producer wants to show that the expected weight is smaller than 100 g. What are the appropriate hypotheses to use?
- (e) Conduct the test for the hypothesis in (d). Again use  $\alpha = 0.05$ .

*Exercise 10.3* Christian decides to purchase the new CD by Bruce Springsteen. His first thought is to buy it online, via an online auction. He discovers that he can also buy the CD immediately, without bidding at an auction, from the same online store. He also looks at the price at an internet book store which was recommended to him by a friend. He notes down the following prices (in €):

**Internet book store** 16.95

**Online store, no auction** 18.19, 16.98, 19.97, 16.98, 18.19, 15.99, 13.79, 15.90, 15.90, 15.90, 15.90, 19.97, 17.72

**Online store, auction** 10.50, 12.00, 9.54, 10.55, 11.99, 9.30, 10.59, 10.50, 10.01, 11.89, 11.03, 9.52, 15.49, 11.02

- (a) Calculate and interpret the arithmetic mean, variance, standard deviation, and coefficient of variation for the online store, both for the auction and non-auction offers.
- (b) Test the hypothesis that the mean price at the online store (no auction) is unequal to €16.95 ( $\alpha = 0.05$ ).
- (c) Calculate a confidence interval for the mean price at the online store (no auction) and interpret your findings in the light of the hypothesis in (b).
- (d) Test the hypothesis that the mean price at the online store (auction) is less than €16.95 ( $\alpha = 0.05$ ).
- (e) Test the hypothesis that the mean non-auction price is higher than the mean auction price. Assume that (i) the variances are equal in both samples and (ii) the variances are unequal ( $\alpha = 0.05$ ).
- (f) Test the hypothesis that the variance of the non-auction price is unequal to the variance of the auction price ( $\alpha = 0.05$ ).

- (g) Use the  $U$ -test to compare the location of the auction and non-auction prices.  
Compare the results with those of (e).
- (h) Calculate the results of (a)–(g) with  $R$ .

*Exercise 10.4* Ten of Leonard's best friends try a new diet: the “Banting” diet. Each of them weighs him/herself before and after the diet. The data is as follows:

Person ( $i$ )	1	2	3	4	5	6	7	8	9	10
Before diet ( $x_i$ )	80	95	70	82	71	70	120	105	111	90
After diet ( $y_i$ )	78	94	69	83	65	69	118	103	112	88

Choose a test and a confidence interval to test whether there is a difference between the mean weight before and after the diet ( $\alpha = 0.05$ ).

*Exercise 10.5* A company producing clothing often finds deficient T-shirts among its production.

- (a) The company's controlling department decides that the production is no longer profitable when there are more than 10 % deficient shirts. A sample of 230 shirts yields 30 shirts which contain deficiencies. Use the approximate binomial test to decide whether the T-shirt production is profitable or not ( $\alpha = 0.05$ ).
- (b) Test the same hypothesis as in (a) using the exact binomial test. You can use  $R$  to determine the quantiles needed for the calculation.
- (c) The company is offered a new cutting machine. To test whether the change of machine helps to improve the production quality, 115 sample shirts are evaluated, 7 of which have deficiencies. Use the two-sample binomial test to decide whether the new machine yields improvement or not ( $\alpha = 0.05$ ).
- (d) Test the same hypothesis as in (c) using the test of Fisher in  $R$ .

*Exercise 10.6* Two friends play a computer game and each of them repeats the same level 10 times. The scores obtained are:

	1	2	3	4	5	6	7	8	9	10
Player 1	91	101	112	99	108	88	99	105	111	104
Player 2	261	47	40	29	64	6	87	47	98	351

- (a) Player 2 insists that he is the better player and suggests to compare their mean performance. Use an appropriate test ( $\alpha = 0.05$ ) to test this hypothesis.
- (b) Player 1 insists that he is the better player. He proposes to not focus on the mean and to use the  $U$ -test for comparison. What are the advantages and disadvantages of using this test compared with (a)? What are the results ( $\alpha = 0.05$ )?

*Exercise 10.7* Otto loves gummy bears and buys 10 packets at a factory store. He opens all packets and sorts them by their colour. He counts 222 white gummy bears, 279 red gummy bears, 251 orange gummy bears, 232 yellow gummy bears, and 266 green ones. He is disappointed since white (pineapple flavour) is his favourite flavour. He hypothesizes that the producer of the bears does not uniformly distribute the bears into the packets. Choose an appropriate test to find out whether Otto's speculation could be true.

*Exercise 10.8* We consider Exercise 4.4 where we evaluated which of the passengers from the *Titanic* were rescued. The data was summarized as follows:

	1. Class	2. Class	3. Class	Staff	Total
Rescued	202	125	180	211	718
Not rescued	135	160	541	674	1510

- (a) The hypothesis derived from the descriptive analysis was that travel class and rescue status are not independent. Test this hypothesis.
- (b) Interpret the following *R* output:

```
4-sample test for equality of proportions
data: titanic
X-squared = 182.06, df = 3, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
prop 1    prop 2    prop 3    prop 4
0.5994065 0.4385965 0.2496533 0.2384181
```

- (c) Summarize the data in a  $2 \times 2$  table: passengers from the first and second class should be grouped together, and third class passengers and staff should be grouped together as well. Is the probability of being rescued higher in the first and second class? Provide an answer using the following three tests: exact test of Fisher,  $\chi^2$ -independence test, and  $\chi^2$ -homogeneity test. You can use *R* to conduct the test of Fisher.

*Exercise 10.9* We are interested in understanding how well the *t*-test can detect differences with respect to the mean. We use *R* to draw 3 samples each of 20 observations from three different normal distributions:  $X \sim N(5, 2^2)$ ,  $Y_1 \sim N(4, 2^2)$ , and  $Y_2 \sim N(3.5, 2^2)$ . The summary statistics of this experiment are as follows:

- $\bar{x} = 4.97$ ,  $s_x^2 = 2.94$ ,
- $\bar{y}_1 = 4.55$ ,  $s_{y_1}^2 = 2.46$ ,
- $\bar{y}_2 = 3.27$ ,  $s_{y_2}^2 = 3.44$ .

- (a) Use the  $t$ -test to compare the means of  $X$  and  $Y_1$ .
- (b) Use the  $t$ -test to compare the means of  $X$  and  $Y_2$ .
- (c) Interpret the results from (a) and (b).

*Exercise 10.10* Access the theatre data described in Appendix A.4. The data summarizes a survey conducted on visitors of a local Swiss theatre in terms of age, sex, annual income, general expenditure on cultural activities, expenditure on theatre visits, and the estimated expenditure on theatre visits in the year before the survey was done.

- (a) Compare the mean expenditure on cultural activities for men and women using the Welch test ( $\alpha = 0.05$ ).
- (b) Would the conclusions change if the two-sample  $t$ -test or the  $U$ -test were used for comparison?
- (c) Test the hypothesis that women spend on average more money on theatre visits than men ( $\alpha = 0.05$ ).
- (d) Compare the mean expenditure on theatre visits in the year of the survey and the preceding year ( $\alpha = 0.05$ ).

*Exercise 10.11* Use  $R$  to read in and analyse the pizza data described in Appendix A.4 (assume  $\alpha = 0.05$ ).

- (a) The manager's aim is to deliver pizzas in less than 30 min and with a temperature of greater than  $65^\circ\text{C}$ . Use an appropriate test to evaluate whether these aims have been reached on average.
- (b) If it takes longer than 40 min to deliver the pizza, then the customers are promised a free bottle of wine. This offer is only profitable if less than 15 % of deliveries are too late. Test the hypothesis  $p < 0.15$ .
- (c) The manager wonders whether there is any relationship between the operator taking the phone call and the pizza temperature. Assume that a hot pizza is defined to be one with a temperature greater  $65^\circ\text{C}$ . Use the test of Fisher, the  $\chi^2$ -independence test, and the  $\chi^2$ -test of homogeneity to test his hypothesis.
- (d) Each branch employs the same number of staff. It would thus be desirable if each branch receives the same number of orders. Use an appropriate test to investigate this hypothesis.
- (e) Is the proportion of calls taken by each operator the same in each branch?
- (f) Test whether there is a relationship between drivers and branches.

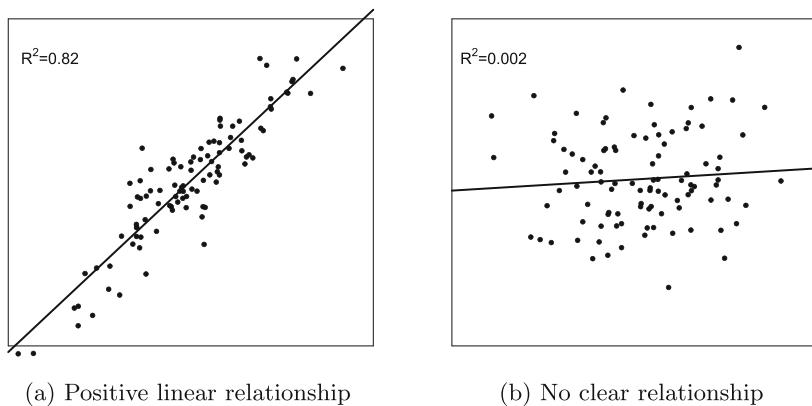
*Exercise 10.12* The authors of this book went to visit historical sites in India. None of them has a particularly strong interest in photography, and they speculated that each of them would take about the same number of pictures on their trip. After returning home, they counted 110, 118, and 105 pictures, respectively. Use an appropriate test to find out whether their speculation was correct ( $\alpha = 0.01$ ).

→ Solutions to all exercises in this chapter can be found on p. [393](#)

\*Source Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg

We learnt about various measures of association in Chap. 4. Such measures are used to understand the degree of *relationship* or *association* between two variables. The correlation coefficient of Bravais–Pearson, for example, can help us to quantify the degree of a linear relationship, but it does not tell us about the functional form of a relationship. Moreover, any association of interest may depend on more than one variable, and we would therefore like to explore *multivariate* relationships. For example, consider a situation where we measured the body weights and the long jump results (distances) of school children taking part in a competition. The correlation coefficient of Bravais–Pearson between the body weight and distance jumped can tell us how strong or weak the association between the two variables is. It can also tell us whether the distance jumped increases or decreases with the body weight in the sense that a negative correlation coefficient indicates shorter jumps for higher body weights, whereas a positive coefficient implies longer jumps for higher body weights. Suppose we want to know how far a child with known body weight, say 50 kg, will jump. Such questions cannot be answered from the value and sign of the correlation coefficient. Another question of interest may be to explore whether the relationship between the body weight and distance jumped is linear or not. One may also be interested in the joint effect of age and body weight on the distance jumped. Could it be that older children, who have on average a higher weight than younger children, perform better? What would be the association of weight and the long jump results of children of the same age? Such questions are addressed by linear regression analysis which we introduce in this chapter.

In many situations, the outcome does not depend on one variable but on several variables. For example, the recovery time of a patient after an operation depends on several factors such as weight, haemoglobin level, blood pressure, body temperature, diet control, rehabilitation, and others. Similarly, the weather depends on many factors such as temperature, pressure, humidity, speed of winds, and others. Multiple



**Fig. 11.1** Scatter plots

linear regression, introduced in Sect. 11.6, addresses the issue where the outcome depends on more than one variable.

To begin with, we consider only two quantitative variables  $X$  and  $Y$  in which the outcome  $Y$  depends on  $X$  and we explore the quantification of their relationship.

*Examples* Examples of associations in which we might be interested in are:

- body height ( $X$ ) and body weight ( $Y$ ) of persons,
- speed ( $X$ ) and braking distance ( $Y$ ) measured on cars,
- money invested (in €) in the marketing of a product ( $X$ ) and sales figures for this product (in €) ( $Y$ ) measured in various branches,
- amount of fertilizer used ( $X$ ) and yield of rice ( $Y$ ) measured on different acres, and
- temperature ( $X$ ) and hotel occupation ( $Y$ ) measured in cities.

## 11.1 The Linear Model

Consider the scatter plots from Fig. 4.2 on p. 80. Plotting  $X$ -values against  $Y$ -values enables us to visualize the relationship between two variables. Figure 11.1a reviews what a positive (linear) association between  $X$  and  $Y$  looks like: the higher the  $X$  values, the higher the  $Y$ -values (and vice versa). The plot indicates that there may be a linear relationship between  $X$  and  $Y$ . On the other hand, Fig. 11.1b displays a scatter plot which shows no clear relationship between  $X$  and  $Y$ . The  $R^2$  measure, shown in the two figures and explained in more detail later, equates to the squared correlation coefficient of Bravais–Pearson.

To summarize the observed association between  $X$  and  $Y$ , we can postulate the following linear relationship between them:

$$Y = \alpha + \beta X. \quad (11.1)$$

This Eq. (11.1) represents a straight line where  $\alpha$  is the **intercept** and  $\beta$  represents the **slope** of the line. The slope indicates the change in the  $Y$ -value when the  $X$ -value changes by one unit. If the sign of  $\beta$  is positive, it indicates that the value of  $Y$  increases as the value of  $X$  increases. If the sign of  $\beta$  is negative, it indicates that the value of  $Y$  decreases as the value of  $X$  increases. When  $X = 0$ , then  $Y = \alpha$ . If  $\alpha = 0$ , then  $Y = \beta X$  represents a line passing through the origin. Suppose the height and body weights in the example of school children are represented in Fig. 11.1a. This has the following interpretation: when the height of a child increases by 1 cm, then the body weight increases by  $\beta$  kilograms. The slope  $\beta$  in Fig. 11.1a would certainly be positive because we have a positive linear relationship between  $X$  and  $Y$ . Note that in this example, the intercept term has no particular meaning because when the height  $X = 0$ , the body weight  $Y = \alpha = 0$ . The scatter diagram in Fig. 11.1b does not exhibit any clear relationship between  $X$  and  $Y$ . Still, the slope of a possible line would likely be somewhere around 0.

It is obvious from Fig. 11.1a that by assuming a linear relationship between  $X$  and  $Y$ , any straight line will not exactly match the data points in the sense that it cannot pass through all the observations: the observations will lie above and below the line. The line represents a *model* to describe the process generating the data.

In our context, a model is a mathematical representation of the relationship between two or more variables. A model has two components—variables (e.g.  $X$ ,  $Y$ ) and parameters (e.g.  $\alpha$ ,  $\beta$ ). A model is said to be *linear* if it is linear in its parameters. A model is said to be *nonlinear* if it is nonlinear in its parameters (e.g.  $\beta^2$  instead of  $\beta$ ). Now assume that each observation potentially deviates by  $e_i$  from the line in Fig. 11.1a. The **linear model** in Eq. (11.1) can then be written as follows to take this into account:

$$Y = \alpha + \beta X + e. \quad (11.2)$$

Suppose we have  $n$  observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , then each observation satisfies

$$y_i = \alpha + \beta x_i + e_i. \quad (11.3)$$

Each deviation  $e_i$  is called an *error*. It represents the deviation of the data points  $(x_i, y_i)$  from the **regression line**. The line in Fig. 11.1a is the fitted (regression) line which we will discuss in detail later. We assume that the errors  $e_i$  are identically and independently distributed with mean 0 and constant variance  $\sigma^2$ , i.e.  $E(e_i) = 0$ ,  $\text{Var}(e_i) = \sigma^2$  for all  $i = 1, 2, \dots, n$ . We will discuss these assumptions in more detail in Sect. 11.7.

In the model (11.2),  $Y$  is called the **response**, **response variable**, **dependent variable** or **outcome**;  $X$  is called the **covariate**, **regressor** or **independent variable**. The scalars  $\alpha$  and  $\beta$  are the parameters of the model and are described as **regression coefficients** or **parameters** of the linear model. In particular,  $\alpha$  is called the **intercept term** and  $\beta$  is called the **slope parameter**.

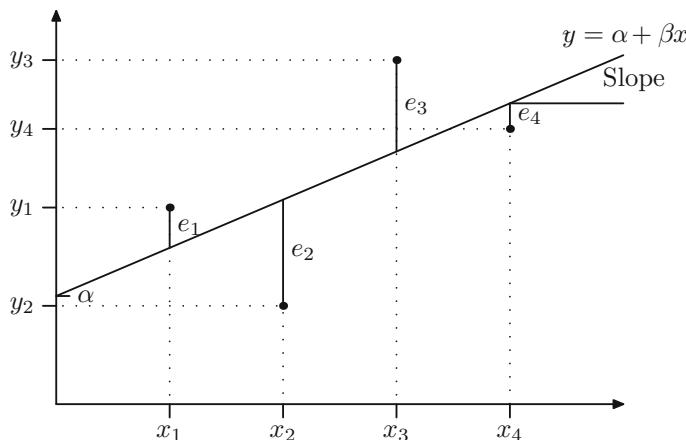
It may be noted that if the regression parameters  $\alpha$  and  $\beta$  are known, then the linear model is completely known. An important objective in identifying the model is to determine the values of the regression parameters using the available observations for  $X$  and  $Y$ . It can be understood from Fig. 11.1a that an ideal situation would be when all the data points lie exactly on the line or, in other words, the error  $e_i$  is zero for each observation. It seems meaningful to determine the values of the parameters in such a way that the errors are minimized.

There are several methods to estimate the values of the regression parameters. In the remainder of this chapter, we describe the methods of least squares and maximum likelihood estimation.

## 11.2 Method of Least Squares

Suppose  $n$  sets of observations  $P_i = (x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , are obtained on two variables  $P = (X, Y)$  and are plotted in a scatter plot. An example of four observations,  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $(x_3, y_3)$ , and  $(x_4, y_4)$ , is given in Fig. 11.2.

The **method of least squares** says that a line can be fitted to the given data set such that the errors are minimized. This implies that one can determine  $\alpha$  and  $\beta$  such that the sum of the squared distances between the data points and the line  $Y = \alpha + \beta X$  is minimized. For example, in Fig. 11.2, the first data point  $(x_1, y_1)$  does not lie on the plotted line and the deviation is  $e_1 = y_1 - (\alpha + \beta x_1)$ . Similarly, we obtain the



**Fig. 11.2** Regression line, observations, and errors  $e_i^*$

difference of the other three data points from the line:  $e_2, e_3, e_4$ . The error is zero if the point lies exactly on the line. The problem we would like to solve in this example is to minimize the sum of squares of  $e_1, e_2, e_3$ , and  $e_4$ , i.e.

$$\min_{\alpha, \beta} \sum_{i=1}^4 (y_i - \alpha - \beta x_i)^2. \quad (11.4)$$

We want the line to fit the data well. This can generally be achieved by choosing  $\alpha$  and  $\beta$  such that the squared errors of all the  $n$  observations are minimized:

$$\min_{\alpha, \beta} \sum_{i=1}^n e_i^2 = \min_{\alpha, \beta} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2. \quad (11.5)$$

If we solve this optimization problem by the principle of maxima and minima, we obtain estimates of  $\alpha$  and  $\beta$  as

$$\left. \begin{aligned} \hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \\ \hat{\alpha} &= \bar{y} - \hat{\beta} \bar{x} \end{aligned} \right\}, \quad (11.6)$$

see Appendix C.6 for a detailed derivation. Here,  $\hat{\alpha}$  and  $\hat{\beta}$  represent the estimates of the parameters  $\alpha$  and  $\beta$ , respectively, and are called the **least squares estimator** of  $\alpha$  and  $\beta$ , respectively. This gives us the model  $y = \hat{\alpha} + \hat{\beta}x$  which is called the *fitted model* or the *fitted regression line*. The literal meaning of “regression” is to move back. Since we are acquiring the data and then moving back to find the parameters of the model using the data, it is called a *regression model*. The fitted regression line  $y = \hat{\alpha} + \hat{\beta}x$  describes the postulated relationship between  $Y$  and  $X$ . The sign of  $\beta$  determines whether the relationship between  $X$  and  $Y$  is positive or negative. If the sign of  $\beta$  is positive, it indicates that if  $X$  increases, then  $Y$  increases too. On the other hand, if the sign of  $\beta$  is negative, it indicates that if  $X$  increases, then  $Y$  decreases. For any given value of  $X$ , say  $x_i$ , the *predicted* value  $\hat{y}_i$  is calculated by

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

and is called the *i*th *fitted value*.

If we compare the observed data point  $(x_i, y_i)$  with the point suggested (predicted, fitted) by the regression line,  $(x_i, \hat{y}_i)$ , the difference between  $y_i$  and  $\hat{y}_i$  is called the **residual** and is given as

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i). \quad (11.7)$$

This can be viewed as an estimator of the error  $e_i$ . Note that it is not really an estimator in the true statistical sense because  $e_i$  is random and so it cannot be estimated. However, since the  $e_i$  are unknown and  $\hat{e}_i$  measures the same difference between the estimated and true values of the  $y$ 's, see for example Fig. 11.2, it can be treated as estimating the error  $e_i$ .

*Example 11.2.1* A physiotherapist advises 12 of his patients, all of whom had the same knee surgery done, to regularly perform a set of exercises. He asks them to record how long they practise. He then summarizes the average time they practised ( $X$ , time in minutes) and how long it takes them to regain their full range of motion again ( $Y$ , time in days). The results are as follows:

$i$	1	2	3	4	5	6	7	8	9	10	11	12
$x_i$	24	35	64	20	33	27	42	41	22	50	36	31
$y_i$	90	65	30	60	60	80	45	45	80	35	50	45

To estimate the linear regression line  $y = \hat{\alpha} + \hat{\beta}x$ , we first calculate  $\bar{x} = 35.41$  and  $\bar{y} = 57.08$ . To obtain  $S_{xy}$  and  $S_{xx}$  we need the following table:

$i$	$x_i$	$y_i$	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	24	90	-11.41	32.92	-375.61	130.19
2	35	65	-0.41	7.92	-3.25	0.17
3	64	30	28.59	-27.08	-774.22	817.39
4	20	60	-15.41	2.92	-45.00	237.47
5	33	60	-2.41	2.92	-7.27	5.81
6	27	80	-8.41	22.92	-192.75	70.73
7	42	45	6.59	-12.08	-79.61	43.43
8	41	45	5.59	-12.08	-67.53	31.25
9	22	80	-13.41	22.92	-307.36	179.83
10	50	35	14.59	-22.08	-322.14	212.87
11	36	50	0.59	-7.08	-4.18	0.35
12	31	45	-4.41	-12.08	53.27	19.45
Total					-2125.65	1748.94

Using (11.6), we can now easily find the least squares estimates  $\hat{\alpha}$  and  $\hat{\beta}$  as

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{-2125.65}{1748.94} \approx -1.22,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 57.08 - (-1.215) \cdot 35.41 = 100.28.$$

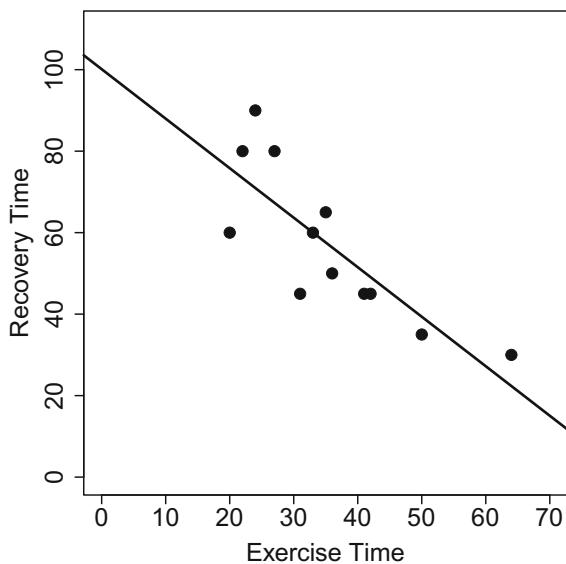
The fitted regression line is therefore

$$y = 100.28 - 1.22 \cdot x.$$

We can interpret the results as follows:

- For an increase of 1 min in exercising, the recovery time decreases by 1.22 days because  $\hat{\beta} = -1.22$ . The negative sign of  $\hat{\beta}$  indicates that the relationship between exercising time and recovery time is negative; i.e. as exercise time increases, the recovery time decreases.
- When comparing two patients with a difference in exercising time of 10 min, the linear model estimates a difference in recovery time of 12.2 days because  $10 \cdot 1.22 = 12.2$ .

**Fig. 11.3** Scatter plot and regression line for Example 11.2.1



- The model predicts an average recovery time of  $100.28 - 1.22 \cdot 38 = 53.92$  days for a patient who practises the set of exercises for 38 min.
- If a patient did not exercise at all, the recovery time would be predicted as  $\hat{a} = 100.28$  days by the model. However, see item (i) in Sect. 11.2.1 below about interpreting a regression line outside of the observed data range.

We can also obtain these results by using *R*. The command `lm(Y~X)` fits a linear model and provides the estimates of  $\hat{\alpha}$  and  $\hat{\beta}$ .

`lm(Y~X)`



We can draw the regression line onto a scatter plot using the command `abline`, see also Fig. 11.3.

```
plot(X, Y)
abline(a=100.28, b=-1.22)
```



### 11.2.1 Properties of the Linear Regression Line

There are a couple of interesting results related to the regression line and the least square estimates.

- As a rule of thumb, one should interpret the regression line  $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$  only in the interval  $[x_{(1)}, x_{(n)}]$ . For example, if  $X$  denotes “Year”, with observed values

from 1999 to 2015, and  $Y$  denotes the “annual volume of sales of a particular company”, then a prediction for the year 2030 may not be meaningful or valid because a linear relationship discovered in the past may not continue to hold in the future.

- (ii) For the points  $\hat{P}_i = (x_i, \hat{y}_i)$ , forming the regression line, we can write

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = \bar{y} + \hat{\beta}(x_i - \bar{x}). \quad (11.8)$$

- (iii) It follows for  $x_i = \bar{x}$  that  $\hat{y}_i = \bar{y}$ , i.e. the point  $(\bar{x}, \bar{y})$  always lies on the regression line. The linear regression line therefore always passes through  $(\bar{x}, \bar{y})$ .

- (iv) The sum of the residuals is zero. The  $i$ th residual is defined as

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i) = y_i - [\bar{y} + \hat{\beta}(x_i - \bar{x})].$$

The sum is therefore

$$\begin{aligned} \sum_{i=1}^n \hat{e}_i &= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} - \hat{\beta} \sum_{i=1}^n (x_i - \bar{x}) \\ &= n\bar{y} - n\bar{y} - \hat{\beta}(n\bar{x} - n\bar{x}) = 0. \end{aligned} \quad (11.9)$$

- (v) The arithmetic mean of  $\hat{y}$  is the same as the arithmetic mean of  $y$ :

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} (n\bar{y} + \hat{\beta}(n\bar{x} - n\bar{x})) = \bar{y}. \quad (11.10)$$

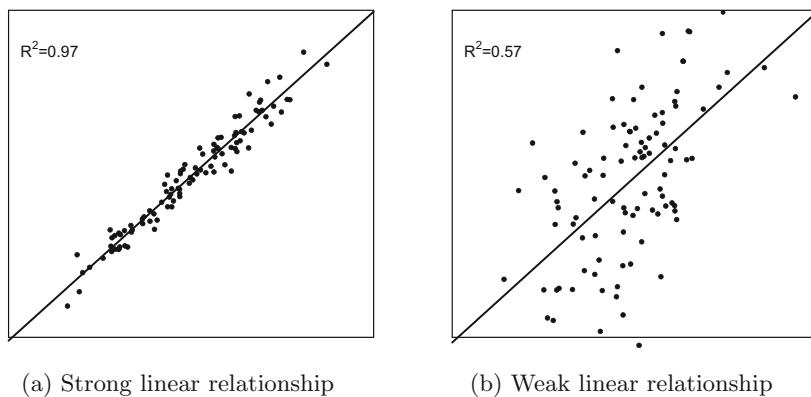
- (vi) The least squares estimate  $\hat{\beta}$  has a direct relationship with the correlation coefficient of Bravais–Pearson:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} = r \sqrt{\frac{S_{yy}}{S_{xx}}}. \quad (11.11)$$

The slope of the regression line is therefore proportional to the correlation coefficient  $r$ : a positive correlation coefficient implies a positive estimate of  $\beta$  and vice versa. However, a stronger correlation does not necessarily imply a steeper slope in the regression analysis because the slope depends upon  $\sqrt{S_{yy}/S_{xx}}$  as well.

### 11.3 Goodness of Fit

While one can easily fit a linear regression model, this does not mean that the model is necessarily good. Consider again Fig. 11.1: In Fig. 11.1a, the regression line is clearly capturing the linear trend seen in the scatter plot. The fit of the model to the data seems good. Figure 11.1b shows however that the data points vary considerably around the line. The quality of the model does not seem to be very good. If we would use the regression line to predict the data, we would likely obtain poor results. It is obvious from Fig. 11.1 that the model provides a good fit to the data when the observations are close to the line and capture a linear trend.



**Fig. 11.4** Different goodness of fit for different data

A look at the scatter plot provides a visual and qualitative approach to judging the quality of the fitted model. Consider Fig. 11.4a, b which both show a linear trend in the data, but the observations in Fig. 11.4a are closer to the regression line than those in Fig. 11.4b. Therefore, the goodness of fit is worse in the latter figure and any quantitative measure should capture this.

A quantitative measure for the goodness of fit can be derived by means of **variance decomposition** of the data. The total variation of  $y$  is partitioned into two components—sum of squares due to the fitted model and sum of squares due to random errors in the data:

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SQ_{\text{Total}}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SQ_{\text{Regression}}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SQ_{\text{Error}}}. \quad (11.12)$$

The proof of the above equation is given in Appendix C.6.

The left-hand side of (11.12) represents the total variability in  $y$  with respect to  $\bar{y}$ . It is proportional to the sample variance (3.21) and is called  $SQ_{\text{Total}}$  (total sum of squares). The first term on the right-hand side of the equation measures the variability which is explained by the fitted linear regression model ( $SQ_{\text{Regression}}$ , sum of squares due to regression); the second term relates to the error sum of squares ( $SQ_{\text{Error}}$ ) and reflects the variation due to random error involved in the fitted model. Larger values of  $SQ_{\text{Error}}$  indicate that the deviations of the observations from the regression line are large. This implies that the model provides a bad fit to the data and that the goodness of fit is low.

Obviously, one would like to have a fit in which the error sum of squares is as small as possible. To judge the goodness of fit, one can therefore look at the error sum of squares in relation to the total sum of squares: in an ideal situation, if the error sum of squares equals zero, the total sum of squares is equal to the sum of squares due to regression and the goodness of fit is optimal. On the other hand, if the sum of squares due to error is large, it will make the sum of squares due to regression smaller

and the goodness of fit should be bad. If the sum of squares due to regression is zero, it is evident that the model fit is the worst possible. These thoughts are reflected in the criterion for the goodness of fit, also known as  $R^2$ :

$$R^2 = \frac{SQ_{\text{Regression}}}{SQ_{\text{Total}}} = 1 - \frac{SQ_{\text{Error}}}{SQ_{\text{Total}}}. \quad (11.13)$$

It follows from the above definition that  $0 \leq R^2 \leq 1$ . The closer  $R^2$  is to 1, the better the fit because  $SQ_{\text{Error}}$  will then be small. The closer  $R^2$  is to 0, the worse the fit, because  $SQ_{\text{Error}}$  will then be large. If  $R^2$  takes any other value, say  $R^2 = 0.7$ , it means that only 70 % of the variation in data is explained by the fitted model, and hence, in simple language, the model is 70 % good. An important point to remember is that  $R^2$  is defined only when there is an intercept term in the model (an assumption we make throughout this chapter). So it is not used to measure the goodness of fit in models without an intercept term.

*Example 11.3.1* Consider again Fig. 11.1: In Fig. 11.1a, the line and data points fit well together. As a consequence  $R^2$  is high,  $R^2 = 0.82$ . Figure 11.1b shows data points with large deviations from the regression line; therefore,  $R^2$  is small, here  $R^2 = 0.002$ . Similarly, in Fig. 11.4a, an  $R^2$  of 0.97 relates to an almost perfect model fit, whereas in Fig. 11.4b, the model describes the data only moderately well ( $R^2 = 0.57$ ).

*Example 11.3.2* Consider again Example 11.2.1 where we analysed the relationship between exercise intensity and recovery time for patients convalescing from knee surgery. To calculate  $R^2$ , we need the following table:

$i$	$y_i$	$\hat{y}_i$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$
1	90	70.84	32.92	1083.73	13.76	189.34
2	65	57.42	7.92	62.73	0.34	0.12
3	30	22.04	-27.08	733.33	-35.04	1227.80
4	60	75.72	2.92	8.53	18.64	347.45
5	60	59.86	2.92	8.53	2.78	7.73
6	80	67.18	22.92	525.33	10.10	102.01
7	45	48.88	-12.08	145.93	-8.2	67.24
8	45	50.10	-12.08	145.93	-6.83	48.72
9	80	73.28	22.92	525.33	16.20	262.44
10	35	39.12	-22.08	487.53	-17.96	322.56
11	50	56.20	-7.08	50.13	-0.88	0.72
12	45	62.30	-12.08	145.93	5.22	27.25
Total				3922.96		2603.43

We calculate  $R^2$  with these results as

$$R^2 = \frac{SQ_{\text{Regression}}}{SQ_{\text{Total}}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{2603.43}{3922.96} = 0.66.$$

We conclude that the regression model provides a reasonable but not perfect fit to the data because 0.66 is not close to 0, but also not very close to 1. About 66 % of the variability in the data can be explained by the fitted linear regression model. The rest is random error: for example, individual variation in the recovery time of patients due to genetic and environmental factors, surgeon performance, and others.

We can also obtain this result in *R* by looking at the summary of the linear model:

```
summary(lm(Y~X))
```



Please note that we give a detailed explanation of the model summary in Sect. 11.7.

There is a direct relationship between  $R^2$  and the correlation coefficient of Bravais–Pearson  $r$ :

$$R^2 = r^2 = \left( \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \right)^2. \quad (11.14)$$

The proof is given in Appendix C.6.

*Example 11.3.3* Consider again Examples 11.3 and 11.5 where we analysed the association of exercising time and time to regain full range of motion after knee surgery. We calculated  $R^2 = 0.66$ . We therefore know that the correlation coefficient is  $r = \sqrt{R^2} = \sqrt{0.66} \approx 0.81$ .

## 11.4 Linear Regression with a Binary Covariate

Until now, we have assumed that the covariate  $X$  is continuous. It is however also straightforward to fit a linear regression model when  $X$  is binary, i.e. if  $X$  has two categories. In this case, the values of  $X$  in the first category are usually coded as 0 and the values of  $X$  in the second category are coded as 1. For example, if the binary variable is “gender”, we replace the word “male” with the number 0 and the word “female” with 1. We can then fit a linear regression model using these numbers, but the interpretation differs from the interpretations in case of a continuous variable. Recall the definition of the linear model,  $Y = \alpha + \beta X + e$  with  $E(e) = 0$ ; if  $X = 0$  (male) then  $E(Y|X = 0) = \alpha$  and if  $X = 1$  (female), then  $E(Y|X = 1) = \alpha + \beta \cdot 1 = \alpha + \beta$ . Thus,  $\alpha$  is the average value of  $Y$  for males, i.e.  $E(Y|X = 0)$ , and  $\beta = E(Y|X = 1) - E(Y|X = 0)$ . It follows that those subjects with  $X = 1$  (e.g. females) have on average  $Y$ -values which are  $\beta$  units higher than subjects with  $X = 0$  (e.g. males).

**Example 11.4.1** Recall Examples 11.2.1, 11.3.2, and 11.3.3 where we analysed the association of exercising time and recovery time after knee surgery. We keep the values of  $Y$  (recovery time, in days) and replace values of  $X$  (exercising time, in minutes) with 0 for patients exercising for less than 40 min and with 1 for patients exercising for 40 min or more. We have therefore a new variable  $X$  indicating whether a patient is exercising a lot ( $X = 1$ ) or not ( $X = 0$ ). To estimate the linear regression line  $\hat{y} = \hat{\alpha} + \hat{\beta}x$ , we first calculate  $\bar{x} = 4/12$  and  $\bar{y} = 57.08$ . To obtain  $S_{xy}$  and  $S_{xx}$ , we need the following table:

$i$	$x_i$	$y_i$	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	0	90	$-\frac{4}{12}$	32.92	-10.97	0.11
2	0	65	$-\frac{4}{12}$	7.92	-2.64	0.11
3	1	30	$\frac{8}{12}$	-27.08	-18.05	0.44
4	0	60	$-\frac{4}{12}$	2.92	-0.97	0.11
5	0	60	$-\frac{4}{12}$	2.92	-0.97	0.11
6	0	80	$-\frac{4}{12}$	22.92	-7.64	0.11
7	1	45	$\frac{8}{12}$	-12.08	-8.05	0.44
8	1	45	$\frac{8}{12}$	-12.08	-8.05	0.44
9	0	80	$-\frac{4}{12}$	22.92	-7.64	0.11
10	1	35	$\frac{8}{12}$	-22.08	-14.72	0.44
11	0	50	$-\frac{4}{12}$	-7.08	2.36	0.11
12	0	45	$-\frac{4}{12}$	-12.08	4.03	0.11
Total	Total				-72.34	2.64

We can now calculate the least squares estimates of  $\alpha$  and  $\beta$  using (11.6) as

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{-72.34}{2.64} \approx -27.4,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 57.08 - (-27.4) \cdot \frac{4}{12} = 66.2.$$

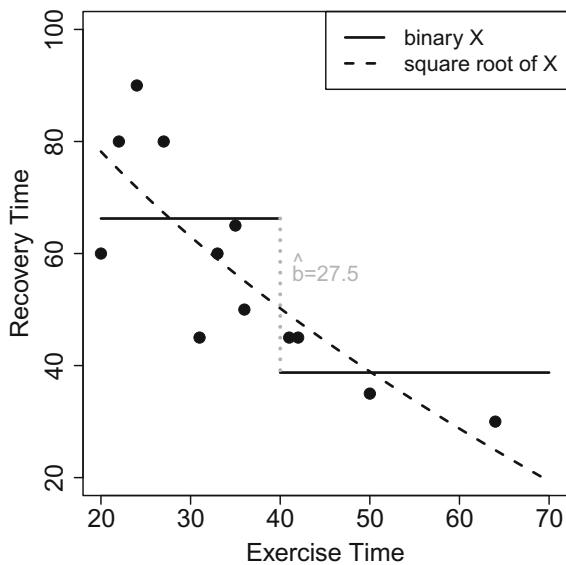
The fitted regression line is:

$$y = 66.2 - 27.4 \cdot x.$$

The interpretation is:

- The average recovery time for patients doing little exercise is  $66.2 - 27.4 \cdot 0 = 66.2$  days.
- Patients exercising heavily ( $x = 1$ ) have, on average, a 27.4 days shorter recovery period ( $\beta = -27.4$ ) than patients with a short exercise time ( $x = 0$ ).
- The average recovery time for patients exercising heavily is 38.8 days ( $66.2 - 27.4 \cdot 1$  for  $x = 1$ ).
- These interpretations are visualized in Fig. 11.5. Because “exercise time” ( $X$ ) is considered to be a nominal variable (two categories representing the intervals  $[0; 40)$  and  $(40; \infty)$ ), we can plot the regression line on a scatter plot as two parallel lines.

**Fig. 11.5** Scatter plot and regression lines for Examples 11.4.1 and 11.5.1



## 11.5 Linear Regression with a Transformed Covariate

Recall that a model is said to be linear when it is linear in its parameters. The definition of a linear model is not at all related to the linearity in the covariate. For example,

$$Y = \alpha + \beta^2 X + e$$

is *not* a linear model because the right-hand side of the equation is not a linear function in  $\beta$ . However,

$$Y = \alpha + \beta X^2 + e$$

is a linear model. This model can be fitted as for any other linear model: we obtain  $\hat{\alpha}$  and  $\hat{\beta}$  as usual and simply use the squared values of  $X$  instead of  $X$ . This can be justified by considering  $X^* = X^2$ , and then, the model is written as  $Y = \alpha + \beta X^* + e$  which is again a linear model. Only the interpretation changes: for each unit increase in the squared value of  $X$ , i.e.  $X^*$ ,  $Y$  increases by  $\beta$  units. Such an interpretation is often not even needed. One could simply plot the regression line of  $Y$  on  $X^*$  to visualize the functional form of the effect. This is also highlighted in the following example.

*Example 11.5.1* Recall Examples 11.2.1, 11.3.2, 11.3.3, and 11.4.1 where we analysed the association of exercising time and recovery time after knee surgery. We estimated  $\hat{\beta}$  as  $-1.22$  by using  $X$  as it is, as a continuous variable, see also Fig. 11.3. When using a binary  $X$ , based on a cut-off of 40 min, we obtained  $\hat{\beta} = -27.4$ , see also Fig. 11.5. If we now use  $\sqrt{X}$  rather than  $X$ , we obtain  $\hat{\beta} = -15.1$ . This means for

an increase of 1 unit of the square root of exercising time, the recovery time decreases by 15.1 days. Such an interpretation will be difficult to understand for many people. It is better to plot the linear regression line  $y = 145.8 - 15.1 \cdot \sqrt{x}$ , see Fig. 11.5. We can see that the new nonlinear line (obtained from a *linear* model) fits the data nicely and it may even be preferable to a straight regression line. Moreover, the value of  $\alpha$  substantially increased with this modelling approach, highlighting that no exercising at all may severely delay recovery from the surgery (which is biologically more meaningful). In R, we obtain these results by either creating a new variable  $\sqrt{X}$  or by using the `I()` command which allows specifying transformations in regression models.

```
newX <- sqrt(X)
lm(Y~newX)           #option 1
lm(Y~I(sqrt(X)))    #option 2
```



Generally the covariate  $X$  can be replaced by any transformation of it, such as using  $\log(X)$ ,  $\sqrt{X}$ ,  $\sin(X)$ , or  $X^2$ . The details are discussed in Sect. 11.6.3. It is also possible to compare the quality and goodness of fit of models with different transformations (see Sect. 11.8).

## 11.6 Linear Regression with Multiple Covariates

Up to now, we have only considered two variables— $Y$  and  $X$ . We therefore assume that  $Y$  is affected by only one covariate  $X$ . However, in many situations, there might be more than one covariate which affects  $Y$  and consequently all of them are relevant to the analysis (see also Sect. 11.10 on the difference between association and causation). For example, the yield of a crop depends on several factors such as quantity of fertilizer, irrigation, and temperature, among others. In another example, in the pizza data set (described in Appendix A.4), many variables could potentially be associated with delivery time—driver, amount of food ordered, operator handling the order, whether it is a weekend, and so on. A reasonable question to ask is: Do different drivers have (on average) a different delivery time—given they deal with the same operator, the amount of food ordered is the same, the day is the same, etc. These kind of questions can be answered with **multiple linear regression**. The model contains more than one, say  $p$ , covariates  $X_1, X_2, \dots, X_p$ . The linear model (11.2) can be extended as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e. \quad (11.15)$$

Note that the intercept term is denoted here by  $\beta_0$ . In comparison with (11.2),  $\alpha = \beta_0$  and  $\beta = \beta_1$ .

*Example 11.6.1* For the pizza delivery data, a particular linear model with multiple covariates could be specified as follows:

$$\begin{aligned}\text{Delivery Time} = & \beta_0 + \beta_1 \text{Number pizzas ordered} + \beta_2 \text{Weekend(0/1)} \\ & + \beta_3 \text{Operator (0/1)} + e.\end{aligned}$$

### 11.6.1 Matrix Notation

If we have a data set of  $n$  observations, then every set of observations satisfies model (11.15) and we can write

$$\begin{aligned}y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + e_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + e_2 \\ &\vdots & \vdots & \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + e_n\end{aligned}\tag{11.16}$$

It is possible to write the  $n$  equations in (11.16) in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}\tag{11.17}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

The letters  $\mathbf{y}$ ,  $\mathbf{X}$ ,  $\boldsymbol{\beta}$ ,  $\mathbf{e}$  are written in bold because they refer to vectors and matrices rather than scalars. The capital letter  $\mathbf{X}$  makes it clear that  $\mathbf{X}$  is a matrix of order  $n \times p$  representing the  $n$  observations on each of the covariates  $X_1, X_2, \dots, X_p$ . Similarly,  $\mathbf{y}$  is a  $n \times 1$  vector of  $n$  observations on  $Y$ ,  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of regression coefficients associated with  $X_1, X_2, \dots, X_p$ , and  $\mathbf{e}$  is a  $n \times 1$  vector of  $n$  errors. The lower case letter  $\mathbf{x}$  relates to a vector representing a variable which means we can denote the multiple linear model from now on also as

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p + \mathbf{e}\tag{11.18}$$

where  $\mathbf{1}$  is the  $n \times 1$  vector of 1's. We assume that  $E(\mathbf{e}) = 0$  and  $\text{Cov}(\mathbf{e}) = \sigma^2 I$  (see Sect. 11.9 for more details).

We would like to highlight that  $\mathbf{X}$  is not the data matrix. The matrix  $\mathbf{X}$  is called the **design matrix** and contains both a column of 1's denoting the presence of the intercept term and all explanatory variables which are relevant to the multiple linear model (including possible transformations and interactions, see Sects. 11.6.3 and 11.7.3). The errors  $\mathbf{e}$  reflect the deviations of the observations from the regression line and therefore the difference between the observed and fitted relationships. Such deviations may occur for various reasons. For example, the measured values can be

affected by variables which are not included in the model, the variables may not be accurately measured, there is unmeasured genetic variability in the study population, and all of which are covered in the error  $\mathbf{e}$ . The estimate of  $\beta$  is obtained by using the least squares principle by minimizing  $\sum_{i=1}^n e_i^2 = \mathbf{e}'\mathbf{e}$ . The **least squares estimate** of  $\beta$  is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (11.19)$$

The vector  $\hat{\beta}$  contains the estimates of  $(\beta_0, \beta_1, \dots, \beta_p)'$ . We can interpret it as earlier:  $\hat{\beta}_0$  is the estimate of the intercept which we obtain because we have added the column of 1's (and is identical to  $\alpha$  in (11.1)). The estimates  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  refer to the regression coefficients associated with the variables  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ , respectively. The interpretation of  $\hat{\beta}_j$  is that it represents the partial change in  $y_i$  when the value of  $x_i$  changes by one unit keeping all other covariates fixed.

A possible interpretation of the intercept term is that when all covariates equal zero then

$$E(\mathbf{y}) = \beta_0 + \beta_1 \cdot 0 + \dots + \beta_p \cdot 0 = \beta_0. \quad (11.20)$$

There are many situations in real life for which there is no meaningful interpretation of the intercept term because one or many covariates cannot take the value zero. For instance, in the pizza data set, the bill can never be €0, and thus, there is no need to describe the average delivery time for a given bill of €0. The intercept term here serves the purpose of improvement of the model fit, but it will not be interpreted.

In some situations, it may happen that the average value of  $y$  is zero when all covariates are zero. Then, the intercept term is zero as well and does not improve the model. For example, suppose the salary of a person depends on two factors—education level and type of qualification. If any person is completely illiterate, even then we observe in practice that his salary is never zero. So in this case, there is a benefit of the intercept term. In another example, consider that the velocity of a car depends on two variables—acceleration and quantity of petrol. If these two variables take values of zero, the velocity is zero on a plane surface. The intercept term will therefore be zero as well and yields no model improvement.

*Example 11.6.2* Consider the pizza data described in Appendix A.4. The data matrix  $\mathcal{X}$  is as follows:

$$\mathcal{X} = \begin{pmatrix} \text{Day} & \text{Date} & \text{Time} & \dots & \text{Discount} \\ \text{Thursday} & 1\text{-May-14} & 35.1 & \dots & 1 \\ \text{Thursday} & 1\text{-May-14} & 25.2 & \dots & 0 \\ \vdots & & & & \vdots \\ \text{Saturday} & 31\text{-May-14} & 35.7 & \dots & 0 \end{pmatrix}$$

Suppose the manager has the hypothesis that the operator and the overall bill (as a proxy for the amount ordered from the customer) influence the delivery time. We can postulate a linear model to describe this relationship as

$$\text{Delivery Time} = \beta_0 + \beta_1 \text{Bill} + \beta_2 \text{Operator} + e.$$

The model in matrix notation is as follows:

$$\underbrace{\begin{pmatrix} 35.1 \\ 25.2 \\ \vdots \\ 35.7 \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 58.4 & 1 \\ 1 & 26.4 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 42.7 & 0 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{1266} \end{pmatrix}}_{\mathbf{e}}$$

To understand the associations in the data, we need to estimate  $\boldsymbol{\beta}$  which we obtain by the least squares estimator  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . Rather than doing this tiresome task manually, we use  $R$ :

```
lm(time~bill+operator)
```



If we have more than one covariate in a linear regression model, we simply add all of them separated by the  $+$  sign when specifying the model formula in the `lm()` function. We obtain  $\hat{\beta}_0 = 23.1$ ,  $\hat{\beta}_1 = 0.26$ ,  $\hat{\beta}_2 = 0.16$ . The interpretation of these parameters is as follows:

- For each extra € that is spent, the delivery time increases by 0.26 min. Or, for each extra €10 spent, the delivery time increases on average by 2.6 min.
- The delivery time is on average 0.16 min longer for operator 1 compared with operator 0.
- For operator 0 and a bill of €0, the expected delivery time is  $\beta_0 = 23.1$  min. However, there is no bill of €0, and therefore, the intercept  $\beta_0$  cannot be interpreted meaningfully here and is included only for improvement of the model fit.

## 11.6.2 Categorical Covariates

We now consider the case when covariates are categorical rather than continuous.

*Examples*

- Region: East, West, South, North,
- Marital status: single, married, divorced, widowed,
- Day: Monday, Tuesday, . . . , Sunday.

We have already described how to treat a categorical variable which consists of two categories, i.e. a binary variable: one category is replaced by 1's and the other category is replaced by 0's subjects who belong to the category  $x = 1$  have on average  $y$ -values which are  $\hat{\beta}$  units higher than those subjects with  $x = 0$ .

Consider a variable  $x$  which has more than two categories, say  $k > 2$  categories. To include such a variable in a regression model, we create  $k - 1$  new variables  $x_i$ ,  $i = 1, 2, \dots, k - 1$ . Similar to how we treat binary variables, each of these variables is a **dummy variable** and consists of 1's for units which belong to the category of interest and 0's for the other category

$$x_i = \begin{cases} 1 & \text{for category } i, \\ 0 & \text{otherwise.} \end{cases} \quad (11.21)$$

The category for which we do not create a dummy variable is called the **reference category**, and the interpretation of the parameters of the dummy variables is with respect to this category. For example, for category  $i$ , the  $y$ -values are on average  $\beta_i$  higher than for the reference category. The concept of creating dummy variables and interpreting them is explained in the following example.

*Example 11.6.3* Consider again the pizza data set described in Appendix A.4. The manager may hypothesize that the delivery times vary with respect to the branch. There are  $k = 3$  branches: East, West, Centre. Instead of using the variable  $\mathbf{x} = \text{branch}$ , we create  $(k - 1)$ , i.e.  $(3 - 1) = 2$  new variables denoting  $\mathbf{x}_1 = \text{East}$  and  $\mathbf{x}_2 = \text{West}$ . We set  $\mathbf{x}_1 = 1$  for those deliveries which come from the branch in the East and set  $\mathbf{x}_1 = 0$  for other deliveries. Similarly, we set  $\mathbf{x}_2 = 1$  for those deliveries which come from the West and  $\mathbf{x}_2 = 0$  for other deliveries. The data then is as follows:

Delivery	$y$ (Delivery Time)	$\mathbf{x}$ (Branch)	$\mathbf{x}_1$ (East)	$\mathbf{x}_2$ (West)
1	35.1	East	1	0
2	25.2	East	1	0
3	45.6	West	0	1
4	29.4	East	1	0
5	30.0	West	0	1
6	40.3	Centre	0	0
:	:	:	:	:
1266	35.7	West	0	1

Deliveries which come from the East have  $\mathbf{x}_1 = 1$  and  $\mathbf{x}_2 = 0$ , deliveries which come from the West have  $\mathbf{x}_1 = 0$  and  $\mathbf{x}_2 = 1$ , and deliveries from the Centre have  $\mathbf{x}_1 = 0$  and  $\mathbf{x}_2 = 0$ . The regression model of interest is thus

$$y = \beta_0 + \beta_1 \text{East} + \beta_2 \text{West} + e.$$

We can calculate the least squares estimate  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  via R: either (i) we create the dummy variables ourselves or (ii) we ask R to create it for us. This requires that “branch” is a factor variable (which it is in our data, but if it was not, then we would have to define it in the model formula via `as.factor()`).

```
East <- as.numeric(branch=='East') R
West <- as.numeric(branch=='West')
lm(time~ East+West)           # option 1
lm(time~ branch)              # option 2a
lm(time~ as.factor(branch))   # option 2b
```

We obtain the following results:

$$Y = 36.3 - 5.2\text{East} - 1.1\text{West}.$$

The interpretations are as follows:

- The average delivery time for the branch in the Centre is 36.3 min, the delivery time for the branch in the East is  $36.3 - 5.2 = 31.1$  min, and the predicted delivery time for the West is  $36.3 - 1.1 = 35.2$  min.
- Therefore, deliveries arrive on average 5.2 min earlier in the East compared with the centre ( $\hat{\beta}_1 = -5.2$ ), and deliveries in the West arrive on average 1.1 min earlier than in the centre ( $\hat{\beta}_2 = -1.1$ ). In other words, it takes 5.2 min less to deliver a pizza in the East than in the Centre. The deliveries in the West take on average 1.1 min less than in the Centre.

Consider now a covariate with  $k = 3$  categories for which we create two new dummy variables,  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . The linear model is  $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + e$ .

- For  $\mathbf{x}_1 = 1$ , we obtain  $E(\mathbf{y}) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 = \beta_0 + \beta_1 \equiv E(\mathbf{y}|\mathbf{x}_1 = 1, \mathbf{x}_2 = 0)$ ;
- For  $\mathbf{x}_2 = 1$ , we obtain  $E(\mathbf{y}) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 1 = \beta_0 + \beta_2 \equiv E(\mathbf{y}|\mathbf{x}_1 = 0, \mathbf{x}_2 = 1)$ ; and
- For the reference category ( $\mathbf{x}_1 = \mathbf{x}_2 = 0$ ), we obtain  $\mathbf{y} = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 = \beta_0 \equiv E(\mathbf{y}|\mathbf{x}_1 = \mathbf{x}_2 = 0)$ .

We can thus conclude that the intercept  $\beta_0 = E(\mathbf{y}|\mathbf{x}_1 = 0, \mathbf{x}_2 = 0)$  describes the average  $\mathbf{y}$  for the reference category, that  $\beta_1 = E(\mathbf{y}|\mathbf{x}_1 = 1, \mathbf{x}_2 = 0) - E(\mathbf{y}|\mathbf{x}_1 = 0, \mathbf{x}_2 = 0)$  describes the difference between the first and the reference category, and that  $\beta_2 = E(\mathbf{y}|\mathbf{x}_1 = 0, \mathbf{x}_2 = 1) - E(\mathbf{y}|\mathbf{x}_1 = 0, \mathbf{x}_2 = 0)$  describes the difference between the second and the reference category.

*Remark 11.6.1* There are other ways of recoding categorical variables, such as effect coding. However, we do not describe them in this book.

### 11.6.3 Transformations

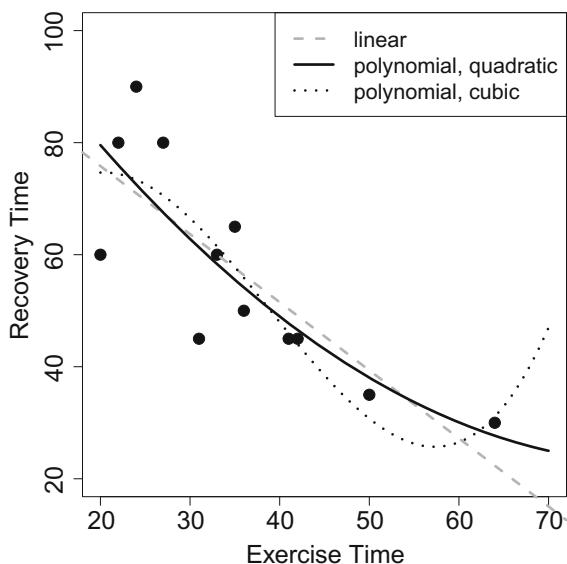
As we have already outlined in Sect. 11.5, if  $\mathbf{x}$  is transformed by a function  $T(\mathbf{x}) = (T(x_1), T(x_2), \dots, T(x_n))$  then

$$\mathbf{y} = f(\mathbf{x}) = \beta_0 + \beta_1 T(\mathbf{x}) + e \tag{11.22}$$

is still a linear model because the model is linear in its parameters  $\beta$ . Popular transformations  $T(\mathbf{x})$  are  $\log(\mathbf{x})$ ,  $\sqrt{\mathbf{x}}$ ,  $\exp(\mathbf{x})$ ,  $\mathbf{x}^p$ , among others. The choice of such a function is typically motivated by the application and data at hand. Alternatively, a relationship between  $\mathbf{y}$  and  $\mathbf{x}$  can be modelled via a **polynomial** as follows:

$$\mathbf{y} = f(\mathbf{x}) = \beta_0 + \beta_1 \mathbf{x} + \beta_2 \mathbf{x}^2 + \cdots + \beta_p \mathbf{x}^p + e. \tag{11.23}$$

**Fig. 11.6** Scatter plot and regression lines for Example 11.6.4



*Example 11.6.4* Consider again Examples 11.2.1, 11.3.2, 11.3.3, 11.4.1, and 11.5.1 where we analysed the association of intensity of exercising and recovery time after knee surgery. The linear regression line was estimated as

$$\text{Recovery Time} = 100.28 - 1.22 \text{ Exercising Time.}$$

One could question whether the association is indeed linear and fit the second- and third-order polynomials:

$$\text{Recovery Time} = \beta_0 - \beta_1 \text{Exercise} + \beta_2 \text{Exercise}^2,$$

$$\text{Recovery Time} = \beta_0 - \beta_1 \text{Exercise} + \beta_2 \text{Exercise}^2 + \beta_3 \text{Exercise}^3.$$

To obtain the estimates we can use *R*:

`lm(Y~X+I(X^2))`  
`lm(Y~X+I(X^2)+I(X^3))`



The results are  $\beta_0 = 121.8$ ,  $\beta_1 = -2.4$ ,  $\beta_2 = 0.014$  and  $\beta_0 = 12.9$ ,  $\beta_1 = 6.90$ ,  $\beta_2 = -0.23$ ,  $\beta_3 = 0.002$  for the two models respectively. While it is difficult to interpret these coefficients directly, we can simply plot the regression lines on a scatter plot (see Fig. 11.6).

We see that the regression based on the quadratic polynomial visually fits the data slightly better than the linear polynomial. It seems as if the relation between recovery time and exercise time is not exactly linear and the association is better modelled through a second-order polynomial. The regression line based on the cubic polynomial seems to be even closer to the measured points; however, the functional form of the association looks questionable. Driven by a single data point, we obtain a

regression line which suggests a heavily increased recovery time for exercising times greater than 65 min. While it is possible that too much exercising causes damage to the knee and delays recovery, the data does not seem to be clear enough to support the shape suggested by the model. This example illustrates the trade-off between fit (i.e. how good the regression line fits the data) and parsimony (i.e. how well a model can be understood if it becomes more complex). Section 11.8 explores this non-trivial issue in more detail.

**Transformations of the Outcome Variable.** It is also possible to apply transformations to the outcome variable. While in many situations, this makes interpretations quite difficult, a log transformation is quite common and easy to interpret. Consider the **log-linear model**

$$\log y = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p + \epsilon$$

where  $y$  is required to be greater than 0. Exponentiating on both sides leads to

$$y = e^{\beta_0 \mathbf{1}} \cdot e^{\beta_1 \mathbf{x}_1} \cdots e^{\beta_p \mathbf{x}_p} \cdot e^{\epsilon}.$$

It can be easily seen that a one unit increase in  $\mathbf{x}_j$  multiplies  $y$  by  $e^{\beta_j}$ . Therefore, if  $y$  is log-transformed, one simply interprets  $\exp(\beta)$  instead of  $\beta$ . For example, if  $y$  is the yearly income,  $\mathbf{x}$  denotes gender (1 = male, 0 = female), and  $\beta_1 = 0.2$  then one can say that a man's income is  $\exp(0.2) = 1.22$  times higher (i.e. 22 %) than a woman's.

## 11.7 The Inductive View of Linear Regression

So far we have introduced linear regression as a method which *describes* the relationship between dependent and independent variables through the best fit to the data. However, as we have highlighted in earlier chapters, often we are interested not only in describing the data but also in drawing conclusions from a sample about the population of interest. For instance, in an earlier example, we have estimated the association of branch and delivery time for the pizza delivery data. The linear model was estimated as delivery time = 36.3 – 5.2 East – 1.1 West; this indicates that the delivery time is on average 5.2 min shorter for the branch in the East of the town compared with the central branch and 1.1 min shorter for the branch in the West compared with the central branch. When considering all pizza deliveries (the population), not only those collected by us over the period of a month (the sample), we might ask ourselves whether 1.1 min is a real difference valid for the entire population or just caused by random error involved in the sample we chose. In particular, we would also like to know what the 95 % confidence interval for our parameter estimate is. Does the interval cover “zero”? If yes, we might conclude that we cannot be very sure that there is an association and do not reject the null hypothesis that  $\beta_{\text{West}} = 0$ . In reference to Chaps. 9 and 10, we now apply the concepts of statistical inference and testing in the context of regression analysis.

**Point and interval estimation in the linear model.** We now rewrite the model for the purpose of introducing statistical inference to linear regression as

$$\begin{aligned}\mathbf{y} &= \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p + \mathbf{e} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad \text{with } \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}).\end{aligned}\tag{11.24}$$

where  $\mathbf{y}$  is the  $n \times 1$  vector of outcomes and  $\mathbf{X}$  is the  $n \times (p+1)$  design matrix (including a column of 1's for the intercept). The identity matrix  $\mathbf{I}$  consists of 1's on the diagonal and 0's elsewhere, and the parameter vector is  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ . We would like to estimate  $\boldsymbol{\beta}$  to make conclusions about a relationship in the population. Similarly,  $\mathbf{e}$  reflects the random errors in the population. Most importantly, the linear model now contains assumptions about the errors. They are assumed to be normally distributed,  $N(0, \sigma^2 \mathbf{I})$ , which means that the expectation of the errors is 0,  $E(e_i) = 0$ , the variance is  $\text{Var}(e_i) = \sigma^2$  (and therefore the same for all  $e_i$ ), and it follows from  $\text{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}$  that  $\text{Cov}(e_i, e_{i'}) = 0$  for all  $i \neq i'$ . The assumption of a normal distribution is required to construct confidence intervals and test of hypotheses statistics. More details about these assumptions and the procedures to test their validity on the basis of a given sample of data are explained in Sect. 11.9.

The least squares estimate of  $\boldsymbol{\beta}$  is obtained by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}.\tag{11.25}$$

It can be shown that  $\hat{\boldsymbol{\beta}}$  follow a normal distribution with mean  $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$  and covariance matrix  $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{I}$  as

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1}).\tag{11.26}$$

Note that  $\hat{\boldsymbol{\beta}}$  is unbiased (since  $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ ); more details about (11.26) can be found in Appendix C.6. An unbiased estimator of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - (p + 1)} = \frac{\hat{\mathbf{e}}' \hat{\mathbf{e}}}{n - (p + 1)} = \frac{1}{n - (p + 1)} \sum_{i=1}^n \hat{e}_i^2.\tag{11.27}$$

The errors are estimated from the data as  $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$  and are called **residuals**.

Before giving a detailed data example, we would like to outline how to draw conclusions about the population of interest from the linear model. As we have seen, both  $\boldsymbol{\beta}$  and  $\sigma^2$  are unknown in the model and are estimated from the data using (11.25) and (11.27). These are our point estimates. We note that if  $\beta_j = 0$ , then  $\beta_j \mathbf{x}_j = 0$ , and then, the model will not contain the term  $\beta_j \mathbf{x}_j$ . This means that the covariate  $\mathbf{x}_j$  does not contribute to explaining the variation in  $\mathbf{y}$ . Testing the hypothesis  $\beta_j = 0$  is therefore equivalent to testing whether  $\mathbf{x}_j$  is associated with  $\mathbf{y}$  or not in the sense that it helps to explain the variations in  $\mathbf{y}$  or not. To test whether the point estimate is different from zero, or whether the deviations of estimates from zero can be explained by random variation, we have the following options:

1. The  $100(1 - \alpha)\%$  confidence interval for each  $\hat{\beta}_j$  is

$$\hat{\beta}_j \pm t_{n-p-1; 1-\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_j} \quad (11.28)$$

with  $\hat{\sigma}_{\hat{\beta}} = \sqrt{s_{jj}\hat{\sigma}^2}$  where  $s_{jj}$  is the  $j$ th diagonal element of the matrix  $(\mathbf{X}'\mathbf{X})^{-1}$ .

If the confidence interval does not overlap 0, then we can conclude that  $\beta$  is different from 0 and therefore  $\mathbf{x}_i$  is associated with  $\mathbf{y}$  and it is a relevant variable. If the confidence interval includes 0, we cannot conclude that there is an association between  $\mathbf{x}_i$  and  $\mathbf{y}$ .

2. We can formulate a formal test which tests if  $\hat{\beta}_j$  is different from 0. The hypotheses are:

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0.$$

The test statistic

$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_{\hat{\beta}_j}^2}}$$

follows a  $t_{n-p-1}$  distribution under  $H_0$ . If  $|T| > t_{n-p-1; 1-\alpha/2}$ , we can reject the null hypothesis (at  $\alpha$  level of significance); otherwise, we accept it.

3. The decisions we get from the confidence interval and the  $T$ -statistic from points 1 and 2 are identical to those obtained by checking whether the  $p$ -value (see also Sect. 10.2.6) is smaller than  $\alpha$ , in which case we also reject the null hypothesis.

*Example 11.7.1* Recall Examples 11.2.1, 11.3.2, 11.3.3, 11.4.1, 11.5.1, and 11.6.4 where we analysed the association of exercising time and time of recovery from knee surgery. We can use  $R$  to obtain a full inductive summary of the linear model:

`summary(lm(Y~X))`



Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	100.1244	10.3571	9.667	2.17e-06 ***
X	-1.2153	0.2768	-4.391	0.00135 **
<hr/>				
Signif. codes:	0 ***	0.001 **	0.01 * 0.05 . 0.1	1

Residual standard error: 11.58 on 10 degrees of freedom

Now, we explain the meaning and interpretation of different terms involved in the output.

- Under “Estimate”, the parameter estimates are listed and we read that the linear model is fitted as  $\mathbf{y}$  (recovery time) = 100.1244 – 1.2153 ·  $\mathbf{x}$  (exercise time). The subtle differences to Example 11.2.1 are due to rounding of numerical values.
- The variance is estimated as  $\hat{\sigma}^2 = 11.58^2$  (“Residual standard error”). This is easily calculated manually using the residual sum of squares:  $\sum \hat{e}_i^2 / (n - p - 1) = 1/10 \cdot \{(90 - 70.84)^2 + \dots + (45 - 62.30)^2\} = 11.58^2$ .

- The standard errors  $\hat{\sigma}_{\hat{\beta}}$  are listed under “Std. Error”. Given that  $n - p - 1 = 12 - 1 - 1 = 10$ , and therefore  $t_{10,0.975} = 2.28$ , we can construct a confidence interval for age:

$$-1.22 \pm 2.28 \cdot 0.28 = [-1.86; -0.58]$$

The interval does not include 0, and we therefore conclude that there is an association between exercising time and recovery time. The random error involved in the data is not sufficient to explain the deviation of  $\hat{\beta}_i = -1.22$  from 0 (given  $\alpha = 0.05$ ).

- We therefore reject the null hypothesis that  $\beta_j = 0$ . This can also be seen by comparing the test statistic (listed under “t value” and obtained by  $(\hat{\beta}_j - 0)/\sqrt{\hat{\sigma}_{\hat{\beta}}^2} = -1.22/0.277$ ) with  $t_{10,0.975}, |-4.39| > 2.28$ . Moreover,  $p = 0.001355 < \alpha = 0.05$ . We can say that there is a significant association between exercising and recovery time.

Sometimes, one is interested in whether a regression model is useful in the sense that all  $\beta_i$ 's are different from zero, and we therefore can conclude that there is an association between any  $\mathbf{x}_i$  and  $\mathbf{y}$ . The null hypothesis is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

and can be tested by the **overall F-test**

$$F = \frac{(\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}})/(p)}{(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})/(n - p - 1)} = \frac{n - p - 1}{p} \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i e_i^2}.$$

The null hypothesis is rejected if  $F > F_{1-\alpha; p, n-p-1}$ . Note that the null hypothesis in this case tests only the equality of slope parameters and does not include the intercept term.

*Example 11.7.2* In this chapter, we have already explored the associations between branch, operator, and bill with delivery time for the pizza data (Appendix A.4). If we fit a multiple linear model including all of the three variables, we obtain the following results:

```
summary(lm(time~branch+bill+operator))
```



Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	26.19138	0.78752	33.258	< 2e-16 ***
branchEast	-3.03606	0.42330	-7.172	1.25e-12 ***
branchWest	-0.50339	0.38907	-1.294	0.196
bill	0.21319	0.01535	13.885	< 2e-16 ***
operatorMelissa	0.15973	0.31784	0.503	0.615
---				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 . 0.1 1

```
Residual standard error: 5.653 on 1261 degrees of freedom
Multiple R-squared:  0.2369,    Adjusted R-squared:  0.2345
F-statistic: 97.87 on 4 and 1261 DF,  p-value: < 2.2e-16
```

By looking at the  $p$ -values in the last column, one can easily see (without calculating the confidence intervals or evaluating the  $t$ -statistic) that there is a significant association between the bill and delivery time; also, it is evident that the average delivery time in the branch in the East is significantly different ( $\approx 3$  min less) from the central branch (which is the reference category here). However, the estimated difference in delivery times for both the branches in the West and the operator was not found to be significantly different from zero. We conclude that some variables in the model are associated with delivery time, while for others, we could not show the existence of such an association. The last line of the output confirms this: the overall  $F$ -test has a test statistic of 97.87 which is larger than  $F_{1-\alpha; p, n-p-1} = F_{0.95; 4, 1261} = 2.37$ ; therefore, the  $p$ -value is smaller  $0.05 (2.2 \times 10^{-16})$  and the null hypothesis is rejected. The test suggests that there is at least one variable which is associated with delivery time.

### 11.7.1 Properties of Least Squares and Maximum Likelihood Estimators

The least squares estimator of  $\beta$  has several properties:

1. The least squares estimator of  $\beta$  is *unbiased*,  $E(\hat{\beta}) = \beta$  (see Appendix C.6 for the proof).
2. The estimator  $\hat{\sigma}^2$  as introduced in equation (11.27) is also *unbiased*, i.e.  $E(\hat{\sigma}^2) = \sigma^2$ .
3. The least squares estimator of  $\beta$  is *consistent*, i.e.  $\hat{\beta}$  converges to  $\beta$  as  $n$  approaches infinity.
4. The least squares estimator of  $\beta$  is *asymptotically normally distributed*.
5. The least squares estimator of  $\beta$  has the smallest variance among all linear and unbiased estimators (*best linear unbiased estimator*) of  $\beta$ .

We do not discuss the technical details of these properties in detail. It is more important to know that the least squares estimator has good features and that is why we choose it for fitting the model. Since we use a “good” estimator, we expect that the model will also be “good”. One may ask whether it is possible to use a different estimator. We have already made distributional assumptions in the model: we require the errors to be normally distributed, given that it is indeed possible to apply the maximum likelihood principle to obtain estimates for  $\beta$  and  $\sigma^2$  in our set-up.

**Theorem 11.7.1** *For the linear model (11.24), the least squares estimator and the maximum likelihood estimator for  $\beta$  are identical. However, the maximum likelihood estimator of  $\sigma^2$  is  $\hat{\sigma}_{ML}^2 = 1/n(\hat{\mathbf{e}}'\hat{\mathbf{e}})$  of  $\sigma^2$  which is a biased estimator of  $\sigma^2$ , but it is asymptotically unbiased.*

How to obtain the maximum likelihood estimator for the linear model is presented in Appendix C.6. The important message of Theorem 11.7.1 is that no matter whether we apply the least squares principle or the maximum likelihood principle, we always

obtain  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ ; this does not apply when estimating the variance, but it is an obvious choice to go for the unbiased estimator (11.27) in any given analysis.

### 11.7.2 The ANOVA Table

A table that is frequently shown by software packages when performing regression analysis is the analysis of variance (ANOVA) table. This table can have several meanings and interpretations and may look slightly different depending on the context in which it is used. We focus here on its interpretation i) as a way to summarize the effect of categorical variables and ii) as a hypothesis test to compare  $k$  means. This is illustrated in the following example.

*Example 11.7.3* Recall Example 11.6.3 where we established the relationship between branch and delivery time as

$$\text{Delivery Time} = 36.3 - 5.2\text{East} - 1.1\text{West}.$$

The *R* output for the respective linear model is as follows:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	36.3127	0.2957	122.819	< 2e-16 ***
branchEast	-5.2461	0.4209	-12.463	< 2e-16 ***
branchWest	-1.1182	0.4148	-2.696	0.00711 **

We see that the average delivery time of the branches in the East and the Centre (reference category) is different and that the average delivery time of the branches in the West and the Centre is different (because  $p < 0.05$ ). This is useful information, but it does not answer the question if branch as a whole influences the delivery time. It seems that this is the case, but the hypothesis we may have in mind may be

$$H_0 : \mu_{\text{East}} = \mu_{\text{West}} = \mu_{\text{Centre}}$$

which corresponds to

$$H_0 : \beta_{\text{East}} = \beta_{\text{West}} = \beta_{\text{Centre}}$$

in the context of the regression model. These are two identical hypotheses because in the regression set-up, we are essentially comparing three conditional means  $E(Y|X = x_1) = E(Y|X = x_2) = E(Y|X = x_3)$ . The ANOVA table summarizes the corresponding *F*-Test which tests this hypothesis:

```
m1 <- lm(time~branch)
anova(m1)
```



Response: time					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
branch	2	6334	3166.8	86.05	< 2.2e-16 ***
Residuals	1263	46481	36.8		

We see that the null hypothesis of 3 equal means is rejected because  $p$  is close to zero.

What does this table mean more generally? If we deal with linear regression with one (possibly categorical) covariate, the table will be as follows:

	df	Sum of squares	Mean squares	$F$ -statistic
Var	$p$	$SQ_{\text{Reg.}}$	$MSR = SQ_{\text{Reg.}}/p$	$MSR/MSE$
Res	$n - p - 1$	$SQ_{\text{Error}}$	$MSE = SQ_{\text{Error}}/(n - p - 1)$	

The table summarizes the sum of squares regression and residuals (see Sect. 11.3 for the definition), standardizes them by using the appropriate degrees of freedom (df), and uses the corresponding ratio as the  $F$ -statistic. Note that in the above example, this corresponds to the overall  $F$ -test introduced earlier. The overall  $F$ -test tests the hypothesis that any  $\beta_j$  is different from zero which is identical to the hypothesis above. Thus, if we fit a linear regression model with one variable, the ANOVA table will yield the same conclusions as the overall  $F$ -test which we obtain through the main summary. However, if we consider a multiple linear regression model, the ANOVA table may give us more information.

*Example 11.7.4* Suppose we are not only interested in the branch, but also in how the pizza delivery times are affected by operator and driver. We may for example hypothesize that the delivery time is identical for all drivers given they deliver for the same branch and speak to the same operator. In a standard regression output, we would get 4 coefficients for 5 drivers which would help us to compare the average delivery time of each driver to the reference driver; it would however not tell us if on an average, they all have the same delivery time. The overall  $F$ -test would not help us either because it would test if any  $\beta_j$  is different from zero which includes coefficients from branch and operator, not only driver. Using the `anova` command yields the results of the  $F$ -test of interest:

```
m2 <- lm(time~branch+operator+driver)
anova(m2)
```



```
Response: time
          Df Sum Sq Mean Sq F value    Pr(>F)
branch      2   6334   3166.8 88.6374 < 2.2e-16 ***
operator    1      16     16.3  0.4566   0.4994
driver      4   1519    379.7 10.6282 1.798e-08 ***
Residuals 1258  44946    35.7
```

We see that the null hypothesis of equal delivery times of the drivers is rejected. We can also test other hypotheses with this output: for instance, the null hypothesis of equal delivery times for each operator is not rejected because  $p \approx 0.5$ .

### 11.7.3 Interactions

It may be possible that the joint effect of some covariates affects the response. For example, drug concentrations may have a different effect on men, woman, and children; a fertilizer could work differently in different geographical areas; or a new education programme may show benefit only with certain teachers. It is fairly simple to target such questions in linear regression analysis by using **interactions**. Interactions are measured as the product of two (or more) variables. If either one or both variables are categorical, then one simply builds products for each dummy variable, thus creating  $(k - 1) \times (l - 1)$  new variables when dealing with two categorical variables (with  $k$  and  $l$  categories respectively). These product terms are called interactions and estimate how an association of one variable differs with respect to the values of the other variable. Now, we give examples for continuous–categorical, categorical–categorical, and continuous–continuous interactions for two variables  $\mathbf{x}_1$  and  $\mathbf{x}_2$ .

**Categorical–Continuous Interaction.** Suppose one variable  $\mathbf{x}_1$  is categorical with  $k$  categories, and the other variable  $\mathbf{x}_2$  is continuous. Then,  $k - 1$  new variables have to be created, each consisting of the product of the continuous variable and a dummy variable,  $\mathbf{x}_2 \times \mathbf{x}_{1i}$ ,  $i \in 1, \dots, k - 1$ . These variables are added to the regression model in addition to the *main effects* relating to  $\mathbf{x}_1$  and  $\mathbf{x}_2$  as follows:

$$\begin{aligned} \mathbf{y} = & \beta_0 + \beta_1 \mathbf{x}_{11} + \cdots + \beta_{k-1} \mathbf{x}_{1k-1} + \beta_k \mathbf{x}_2 \\ & + \beta_{k+1} \mathbf{x}_{11} \mathbf{x}_2 + \cdots + \beta_p \mathbf{x}_{1k-1} \mathbf{x}_2 + \mathbf{e}. \end{aligned}$$

It follows that for the reference category of  $\mathbf{x}_1$ , the effect of  $\mathbf{x}_2$  is just  $\beta_k$  (because each  $\mathbf{x}_2 \mathbf{x}_{1i}$  is zero since each  $\mathbf{x}_{1j}$  is zero). However, the effect for all other categories is  $\beta_2 + \beta_j$  where  $\beta_j$  refers to  $\mathbf{x}_{1j} \mathbf{x}_2$ . Therefore, the association between  $\mathbf{x}_2$  and the outcome  $\mathbf{y}$  differs by  $\beta_j$  between category  $j$  and the reference category. Testing  $H_0 : \beta_j = 0$  thus helps to identify whether there is an interaction effect with respect to category  $j$  or not.

*Example 11.7.5* Consider again the pizza data described in Appendix A.4. We may be interested in whether the association of delivery time and temperature varies with respect to branch. In addition to time and branch, we therefore need additional interaction variables. Since there are 3 branches, we need  $3 - 1 = 2$  interaction variables which are essentially the product of (i) time and branch “East” and (ii) time and branch “West”. This can be achieved in R by using either the “ $\star$ ” operator (which will create both the main and interaction effects) or the “ $:$ ” operator (which only creates the interaction term).

```
int.model.1 <- lm(temperature~time*branch)
int.model.2 <- lm(temperature~time+branch+time:branch)
summary(int.model.1)
summary(int.model.2)
```

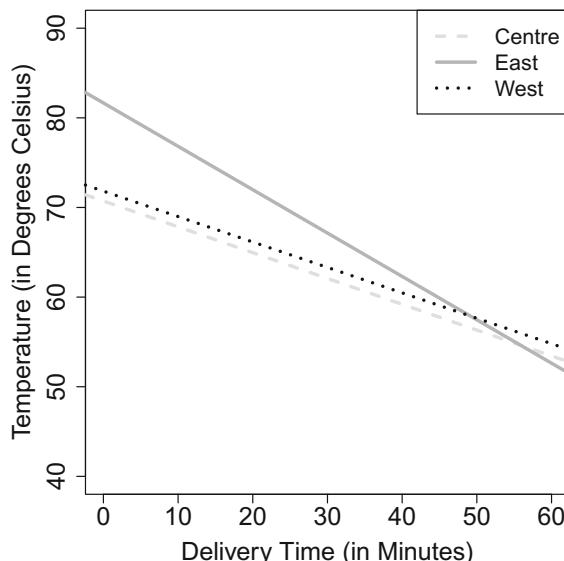


Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	70.718327	1.850918	38.207	< 2e-16	***
time	-0.288011	0.050342	-5.721	1.32e-08	***
branchEast	10.941411	2.320682	4.715	2.69e-06	***
branchWest	1.102597	2.566087	0.430	0.66750	
time:branchEast	-0.195885	0.066897	-2.928	0.00347	**
time:branchWest	0.004352	0.070844	0.061	0.95103	

The main effects of the model tell us that the temperature is almost 11 degrees higher for the eastern branch compared to the central branch (reference) and about 1 degree higher for the western branch. However, only the former difference is significantly different from 0 (since the  $p$ -value is smaller than  $\alpha = 0.05$ ). Moreover, the longer the delivery time, the lower the temperature (0.29 degrees for each minute). The parameter estimates related to the interaction imply that this association differs by branch: the estimate is indeed  $\beta_{\text{time}} = -0.29$  for the reference branch in the Centre, but the estimate for the branch in the East is  $-0.29 - 0.196 = -0.486$  and the estimate for the branch in the West is  $-0.29 + 0.004 = -0.294$ . However, the latter difference in the association of time and temperature is not significantly different from zero. We therefore conclude that the delivery time and pizza temperature are negatively associated and this is more strongly pronounced in the eastern branch compared to the other two branches. It is also possible to visualize this by means of a separate regression line for each branch, see Fig. 11.7.

**Fig. 11.7** Interaction of delivery time and branch in Example 11.7.5



- Centre: Temperature =  $70.7 - 0.29 \times \text{time}$ ,
- East: Temperature =  $70.7 + 10.9 - (0.29 + 0.195) \times \text{time}$ ,
- West: Temperature =  $70.7 + 1.1 - (0.29 - 0.004) \times \text{time}$ .

One can see that the pizzas delivered by the branch in the East are overall hotter but longer delivery times level that benefit off. One might speculate that the delivery boxes from the eastern branch are not properly closed and therefore—despite the overall good performance—the temperature falls more rapidly over time for this branch.

**Categorical–Categorical Interaction.** For two categorical variables  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , with  $k$  and  $l$  categories, respectively,  $(k-1) \times (l-1)$  new dummy variables  $\mathbf{x}_{1i} \times \mathbf{x}_{2j}$  need to be created as follows:

$$\begin{aligned} \mathbf{y} = & \beta_0 + \beta_1 \mathbf{x}_{11} + \cdots + \beta_{k-1} \mathbf{x}_{1k-1} + \beta_k \mathbf{x}_{21} + \cdots + \beta_{k+l-2} \mathbf{x}_{2l-1} \\ & + \beta_{k+l-1} \mathbf{x}_{11} \mathbf{x}_{21} + \cdots + \beta_p \mathbf{x}_{1k-1} \mathbf{x}_{2l-1} + \mathbf{e}. \end{aligned}$$

The interpretation of the regression parameters of interest is less complicated than it looks at first. If the interest is in  $\mathbf{x}_1$ , then the estimate referring to the category of interest (i.e.  $\mathbf{x}_{1i}$ ) is interpreted as usual—with the difference that it relates only to the reference category of  $\mathbf{x}_2$ . The effect of  $\mathbf{x}_{1i}$  may vary with respect to  $\mathbf{x}_2$ , and the sum of the respective main and interaction effects for each category of  $\mathbf{x}_2$  yields the respective estimates. These considerations are explained in more detail in the following example.

*Example 11.7.6* Consider again the pizza data. If we have the hypothesis that the delivery time depends on the operator (who receives the phone calls), but the effect is different for different branches, then a regression model with branch (3 categories, 2 dummy variables), operator (2 categories, one dummy variable), and their interaction (2 new variables) can be used.

`lm(time~operator*branch)`



Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	36.4203	0.4159	87.567	<2e-16	***
operatorMelissa	-0.2178	0.5917	-0.368	0.7129	
branchEast	-5.6685	0.5910	-9.591	<2e-16	***
branchWest	-1.3599	0.5861	-2.320	0.0205	*
operatorMelissa:branchEast	0.8599	0.8425	1.021	0.3076	
operatorMelissa:branchWest	0.4842	0.8300	0.583	0.5598	

The interaction terms can be interpreted as follows:

- If we are interested in the operator, we see that the delivery time is on average 0.21 min shorter for operator “Melissa”. When this operator deals with a branch other than the reference (Centre), the estimate changes to  $-0.21 + 0.86 = 0.64$  min longer delivery in the case of branch “East” and  $-0.21 + 0.48 = 0.26$  min for branch “West”.
- If we are interested in the branches, we observe that the delivery time is shortest for the eastern branch which has on average a 5.66 min shorter delivery time than the central branch. However, this is the estimate for the reference operator only; if operator “Melissa” is in charge, then the difference in delivery times for the two branches is only  $-5.66 + 0.86 = -4.8$  min. The same applies when comparing the western branch with the central branch: depending on the operator, the difference between these two branches is estimated as either  $-1.36$  or  $-1.36 + 0.48 = 0.88$  min, respectively.
- The interaction terms are not significantly different from zero. We therefore conclude that the hypothesis of different delivery times for the two operators, possibly varying by branch, could not be confirmed.

**Continuous–Continuous Interaction.** It is also possible to add an interaction of two continuous variables. This is done by adding the product of these two variables to the model. If  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are two continuous variables, then  $\mathbf{x}_1\mathbf{x}_2$  is the new variable added in the model as an interaction effect as

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_1 \mathbf{x}_2 + \mathbf{e}.$$

*Example 11.7.7* If we again consider the pizza data, with pizza temperature as an outcome, we may wish to test for an interaction of bill and time.

Coefficients:						
	Estimate	Std. Error	t value	Pr(> t )		
(Intercept)	92.555943	2.747414	33.688	< 2e-16	***	
bill	-0.454381	0.068322	-6.651	4.34e-11	***	
time	-0.679537	0.086081	-7.894	6.31e-15	***	
bill:time	0.008687	0.002023	4.294	1.89e-05	***	

The *R* output above reveals that there is a significant interaction effect. The interpretation is more difficult here. It is clear that a longer delivery time and a larger bill decrease the pizza’s temperature. However, for a large product of bill and time (i.e., when both are large), these associations are less pronounced because the negative coefficients become more and more outweighed by the positive interaction term. On the contrary, for a small bill and short delivery time, the temperature can be expected to be quite high.

**Combining Estimates.** As we have seen in the examples in this section, it can make sense to combine regression estimates when interpreting interaction effects. While it is simple to obtain the point estimates, it is certainly also important to report

their 95 % confidence intervals. As we know from Theorem 7.7.1, the variance of the combination of two random variables can be obtained by  $\text{Var}(X \pm Y) = \sigma_{XY}^2 = \text{Var}(X) + \text{Var}(Y) \pm 2 \text{Cov}(X, Y)$ . We can therefore estimate a confidence interval for the combined estimate  $\hat{\beta}_i + \hat{\beta}_j$  as

$$(\hat{\beta}_i + \hat{\beta}_j) \pm t_{n-p-1; 1-\alpha/2} \cdot \hat{\sigma}_{(\hat{\beta}_i + \hat{\beta}_j)} \quad (11.29)$$

where  $\hat{\sigma}_{(\hat{\beta}_i + \hat{\beta}_j)}$  is obtained from the estimated covariance matrix  $\widehat{\text{Cov}}(\hat{\beta})$  via

$$\hat{\sigma}_{(\hat{\beta}_i + \hat{\beta}_j)} = \sqrt{\widehat{\text{Var}}(\beta_i) + \widehat{\text{Var}}(\beta_j) + 2\widehat{\text{Cov}}(\beta_i, \beta_j)}.$$

*Example 11.7.8* Recall Example 11.7.5 where we estimated the association between pizza temperature, delivery time, and branch. There was a significant interaction effect for time and branch. Using R, we obtain the covariance matrix as follows:

```
mymodel <- lm(temperature~time*branch)
vcov(mymodel)
```



	(Int.)	time	East	West	time:East	time:West
(Int.)	3.4258	-0.09202	-3.4258	-3.4258	0.09202	0.09202
time	-0.0920	0.00253	0.0920	0.0920	-0.00253	-0.00253
branchEast	-3.4258	0.09202	5.3855	3.4258	-0.15232	-0.09202
branchWest	-3.4258	0.09202	3.4258	6.5848	-0.09202	-0.17946
time:East	0.0920	-0.00253	-0.1523	-0.0920	0.00447	0.00253
time:West	0.0920	-0.00253	-0.0920	-0.1794	0.00253	0.00501

The point estimate for the association between time and temperature in the eastern branch is  $-0.29 - 0.19$  (see Example 11.7.5). The standard error is calculated as  $\sqrt{0.00253 + 0.00447 - 2 \cdot 0.00253} = 0.044$ . The confidence interval is therefore  $-0.48 \pm 1.96 \cdot 0.044 = [-0.56; -0.39]$ .

*Remark 11.7.1* If there is more than one interaction term, then it is generally possible to test whether there is an overall interaction effect, such as  $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$  in the case of four interaction variables. These tests belong to the general class of “linear hypotheses”, and they are not explained in this book. It is also possible to create interactions between more than two variables. However, the interpretation then becomes difficult.

## 11.8 Comparing Different Models

There are many situations where different multiple linear models can be fitted to a given data set, but it is unclear which is the best one. This relates to the question of which variables should be included in a model and which ones can be removed.

For example, when we modelled the association of recovery time and exercising time by using different transformations, it remained unclear which of the proposed transformations was the best. In the pizza delivery example, we have about ten covariates: Does it make sense to include all of them in the model or do we understand the associations in the data better by restricting ourselves to the few most important variables? Is it necessary to include interactions or not?

There are various model selection criteria to compare the quality of different fitted models. They can be useful in many situations, but there are also a couple of limitations which make model selection a difficult and challenging task.

**Coefficient of Determination ( $R^2$ ).** A possible criterion to check the goodness of fit of a model is the coefficient of determination ( $R^2$ ). We discussed its development, interpretation, and philosophy in Sect. 11.3 (see p. 256 for more details).

**Adjusted  $R^2$ .** Although  $R^2$  is a reasonable measure for the goodness of fit of a model, it also has some limitations. One limitation is that if the number of covariates increases,  $R^2$  increases too (we omit the theorem and the proof for this statement). The added variables may not be relevant, but  $R^2$  will still increase, wrongly indicating an improvement in the fitted model. This criterion is therefore not a good measure for model selection. An adjusted version of  $R^2$  is defined as

$$R_{adj}^2 = 1 - \frac{SQ_{\text{Error}}/(n-p-1)}{SQ_{\text{Total}}/(n-1)}. \quad (11.30)$$

It can be used to compare the goodness of fit of models with a different number of covariates. Please note that both  $R^2$  and  $R_{adj}^2$  are only defined when there is an intercept term in the model (which we assume throughout the chapter).

*Example 11.8.1* In Fig. 11.6, the association of exercising time and recovery time after knee surgery is modelled linearly, quadratically, and cubically; in Fig. 11.5, this association is modelled by means of a square-root transformation. The model summary in R returns both  $R^2$  (under “Multiple R-squared”) and the adjusted  $R^2$  (under “adjusted R-squared”). The results are as follows:

	$R^2$	$R_{adj}^2$
Linear	0.6584	0.6243
Quadratic	0.6787	0.6074
Cubic	0.7151	0.6083
Square root	0.6694	0.6363

It can be seen that  $R^2$  is larger for the models with more variables; i.e. the cubic model (which includes three variables) has the largest  $R^2$ . The adjusted  $R^2$ , which takes the different model sizes into account, favours the model with the square-root transformation. This model provides therefore the best fit to the data among the four models considered using  $R^2$ .

**Akaike's Information Criterion** (AIC) is another criterion for model selection. The AIC is based on likelihood methodology and has its origins in information theory. We do not give a detailed motivation of this criterion here. It can be written for the linear model as

$$\text{AIC} = n \log\left(\frac{S Q_{\text{Error}}}{n}\right) + 2(p + 1). \quad (11.31)$$

The smaller the AIC, the better the model. The AIC takes not only the fit to the data via  $S Q_{\text{Error}}$  into account but also the parsimony of the model via the term  $2(p + 1)$ . It is in this sense a more mature criterion than  $R^2_{adj}$ , which considers only the fit to the data via  $S Q_{\text{Error}}$ . Akaike's Information Criterion can be calculated in R via the `extractAIC()` command. There are also other commands, but the results differ slightly because the different formulae use different constant terms. However, no matter what formula is used, the results regarding the model choice are always the same.

*Example 11.8.2* Consider again Example 11.8.1.  $R^2_{adj}$  preferred the model which includes exercising time via a square-root transformation over the other options. The AIC value for the model where exercise time is modelled linearly is 60.59, when modelled quadratically 61.84, when modelled cubically 62.4, and 60.19 for a square-root transformation. Thus, in line with  $R^2_{adj}$ , the AIC also prefers the square-root transformation.

**Backward selection.** Two models, which differ only by one variable  $x_j$ , can be compared by simply looking at the test result for  $\beta_j = 0$ : if the null hypothesis is rejected, the variable is kept in the model; otherwise, the other model is chosen. If there are more than two models, then it is better to consider a systematic approach to comparing them. For example, suppose we have 10 potentially relevant variables and we are not sure which of them to include in the final model. There are  $2^{10} = 1024$  possible different combinations of variables and in turn so many choices of models! The inclusion or deletion of variables can be done systematically with various procedures, for example with **backward selection** (also known as **backward elimination**) as follows:

1. Start with the full model which contains all variables of interest,  $\gamma = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\}$ .
2. Remove the variable  $\mathbf{x}_i \in \gamma$  which optimizes a criterion, i.e. leads to the smallest AIC (the highest  $R^2_{adj}$ , the highest test statistic, or the smallest significance) among all possible choices.
3. Repeat step 2. until a stop criterion is fulfilled, i.e. until no improvement regarding AIC,  $R^2$ , or the  $p$ -value can be achieved.

There are several other approaches. Instead of moving in the backward direction, we can also move in the forward direction. **Forward selection** starts with a model with no covariates and adds variables as long as the model continues to improve with respect to a particular criterion. The **stepwise selection** approach combines forward selection and backward selection: either one does backward selection and checks in between whether adding variables yields improvement, or one does forward selection and continuously checks whether removing the already added variables improves the model.

*Example 11.8.3* Consider the pizza data: if delivery time is the outcome, and branch, bill, operator, driver, temperature, and number of ordered pizzas are potential covariates, we may decide to include only the relevant variables in a final model. Using the `stepAIC` function of the library (MASS) allows us implementing backward selection with *R*.

```
library(MASS)
ms <- lm(time~branch+bill+operator+driver
+temperature+pizzas)
stepAIC(ms, direction='back')
```



At the first step, the *R* output states that the AIC of the full model is 4277.56. Then, the AIC values are displayed when variables are removed: 4275.9 if operator is removed, 4279.2 if driver is removed, and so on. One can see that the AIC is minimized if operator is excluded.

```
Start: AIC=4277.56
time ~ branch+bill+operator+driver+temperature+pizzas

          Df Sum of Sq    RSS      AIC
- operator     1      9.16 36508  4275.9
<none>                   36499  4277.6
- driver       4     279.71 36779  4279.2
- branch       2     532.42 37032  4291.9
- pizzas       1     701.57 37201  4299.7
- temperature   1    1931.50 38431  4340.8
- bill          1    2244.28 38743  4351.1
```

Now *R* fits the model without operator. The AIC is 4275.88. Excluding further variables does not improve the model with respect to the AIC.

```
Step: AIC=4275.88
time ~ branch + bill + driver + temperature + pizzas
```

	Df	Sum of Sq	RSS	AIC
<none>		36508	4275.9	
- driver	4	288.45	36797	4277.8
- branch	2	534.67	37043	4290.3
- pizzas	1	705.53	37214	4298.1
- temperature	1	1923.92	38432	4338.9
- bill	1	2249.60	38758	4349.6

We therefore conclude that the final “best” model includes all variables considered, except operator. Using stepwise selection (with option both instead of back) yields the same results. We could now interpret the summary of the chosen model.

**Limitations and further practical considerations.** The above discussions of variable selection approaches makes it clear that there are a variety of options to compare fitted models, but there is no unique best way to do so. Different criteria are based on different philosophies and may lead to different outcomes. There are a few considerations that should however always be taken into account:

- If categorical variables are included in the model, then all respective dummy variables need to be considered as a whole: all dummy variables of this variable should either be removed or kept in the model. For example, a variable with 3 categories is represented by two dummy variables in a regression model. In the process of model selection, both of these variables should either be kept in the model or removed, but not just one of them.
- A similar consideration refers to interactions. If an interaction is added to the model, then the main effects should be kept in the model as well. This enables straightforward interpretations as outlined in Sect. 11.7.3.
- The same applies to polynomials. If adding a cubic transformation of  $x_i$ , the squared transformation as well as the linear effect should be kept in the model.
- When applying backward, forward, or stepwise regression techniques, the results may vary considerably depending on the method chosen! This may give different models for the same data set. Depending on the strategy used to add and remove variables, the choice of the criterion (e.g. AIC versus  $p$ -values), and the detailed set-up (e.g. if the strategy is to add/remove a variable if the  $p$ -value is smaller than  $\alpha$ , then the choice of  $\alpha$  is important), one may obtain different results.
- All the above considerations show that model selection is a non-trivial task which should always be complemented by substantial knowledge of the subject matter. If there are not too many variables to choose from, simply reporting the full model can be meaningful too.

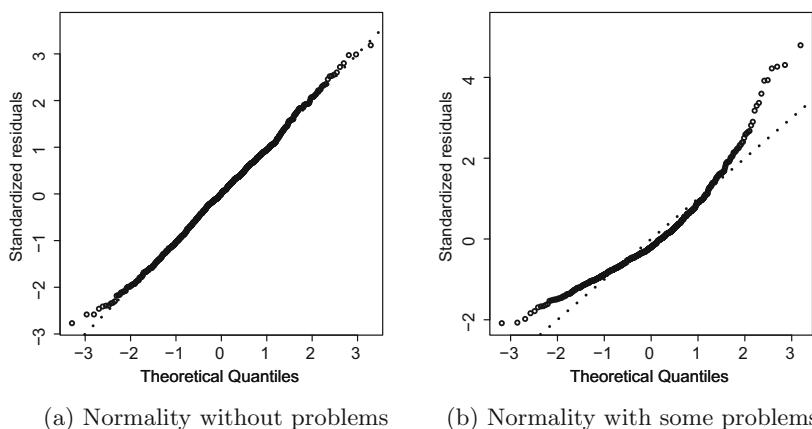
## 11.9 Checking Model Assumptions

The assumptions made for the linear model mostly relate to the error terms:  $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ . This implies a couple of things: i) the difference between  $\mathbf{y}$  and  $\mathbf{X}\beta$  (which is  $\mathbf{e}$ ) needs to follow a normal distribution, ii) the variance for each error  $e_i$  is constant as  $\sigma^2$  and therefore does not depend on  $i$ , and iii) there is no correlation between the error terms,  $\text{Cov}(e_i, e_{i'}) = 0$  for all  $i, i'$ . We also assume that  $\mathbf{X}\beta$  is adequate in the sense that we included all relevant variables and interactions in the model and that the estimator  $\hat{\beta}$  is stable and adequate due to influential observations or highly correlated variables. Now, we demonstrate how to check assumptions i) and ii).

**Normality assumption.** The errors  $\mathbf{e}$  are assumed to follow a normal distribution. We never know the errors themselves. The only possibility is to estimate the errors by means of the residuals  $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}$ . To ensure that the estimated residuals are comparable to the scale of a *standard* normal distribution, they can be standardized via

$$\hat{e}_i^* = \frac{\hat{e}_i}{\hat{\sigma} \sqrt{1 - P_{ii}}}$$

where  $P_{ii}$  is the  $i$ th diagonal element of the hat matrix  $P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Generally, the estimated **standardized residuals** are used to check the normality assumption. To obtain a first impression about the error distribution, one can simply plot a histogram of the standardized residuals. The normality assumption is checked with a QQ-plot where the theoretical quantiles of a standard normal distribution are plotted against the quantiles from the standardized residuals. If the data approximately matches the bisecting line, then we have evidence for fulfilment of the normality assumption (Fig. 11.8a), otherwise we do not (Fig. 11.8b).



**Fig. 11.8** Checking the normality assumption with a QQ-plot

Small deviations from the normality assumption are not a major problem as the least squares estimator of  $\beta$  remains unbiased. However, confidence intervals and tests of hypothesis rely on this assumption, and particularly for small sample sizes and/or stronger violations of the normality assumptions, these intervals and conclusions from tests of hypothesis no longer remain valid. In some cases, these problems can be fixed by transforming  $y$ .

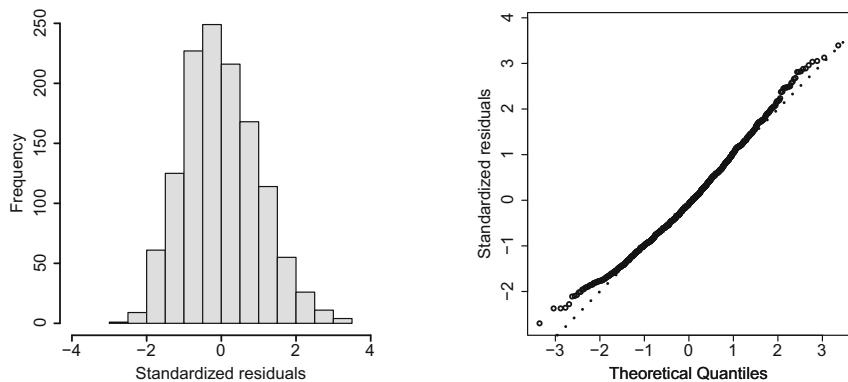
**Heteroscedasticity.** If errors have a constant variance, we say that the errors are homoscedastic and we call this phenomenon **homoscedasticity**. When the variance of errors is not constant, we say that the errors are heteroscedastic. This phenomenon is also known as **heteroscedasticity**. If the variance  $\sigma^2$  depends on  $i$ , then the variability of the  $e_i$  will be different for different groups of observations. For example, the daily expenditure on food ( $y$ ) may vary more among persons with a high income ( $x_1$ ), so fitting a linear model yields stronger variability of the  $e_i$  among higher income groups. Plotting the fitted values  $\hat{y}_i$  (or alternatively  $x_i$ ) against the standardized residuals (or a transformation thereof) can help to detect whether or not problems exist. If there is no pattern in the plot (random plot), then there is likely no violation of the assumption (see Fig. 11.10a). However, if there is a systematic trend, i.e. higher/lower variability for higher/lower  $\hat{y}_i$ , then this indicates heteroscedasticity (Fig. 11.10b, trumpet plot). The consequences are again that confidence intervals and tests may no longer be correct; again, in some situations, a transformation of  $y$  can help.

*Example 11.9.1* Recall Example 11.8.3 where we identified a good model to describe the delivery time for the pizza data. Branch, bill, temperature, number of pizzas ordered, and driver were found to be associated with delivery time. To explore whether the normality assumption is fulfilled for this model, we can create a histogram for the standardized residuals (using the R function `rstandard()`). A QQ-plot is contained in the various model diagnostic plots of the `plot()` function.

```
fm <- lm(time~branch+bill+driver+temperature
+ pizzas)
hist(rstandard(fm))
plot(fm, which=2)
```



Figure 11.9 shows a reasonably symmetric distribution of the residuals, maybe with a slightly longer tail to the right. As expected for a standard normal distribution, not many observations are larger than 2 or smaller than  $-2$  (which are close to the 2.5 and 97.5 % quantiles). The QQ-plot also points towards a normal error distribution

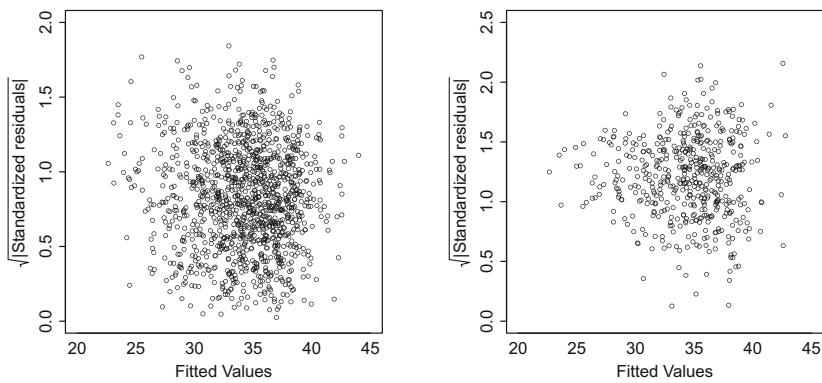


(a) Via the histogram of the standardized residuals      (b) Via a QQ-plot of the standardized residuals

**Fig. 11.9** Checking the normality assumption in Example 11.9.1

because the observed quantiles lie approximately on the bisecting line. It seems that the lowest residuals deviate a bit from the expected normal distribution, but not extremely. The normality assumption does not seem to be violated.

Plotting the fitted values  $\hat{y}_i$  against the square root of the absolute values of the standardized residuals ( $= \sqrt{|\hat{e}_i^*|}$ , used by *R* for stability reasons) yields a plot with no visible structure (see Fig. 11.10a). There is no indication of heteroscedasticity.



**Fig. 11.10** Checking the heteroscedasticity assumption

```
plot(fm, which=3)
```



Figure 11.10b shows an artificial example where heteroscedasticity occurs: the higher the fitted values, the higher is the variation of residuals. This is also called a **trumpet plot**.

---

## 11.10 Association Versus Causation

There is a difference between association and causation: association says that higher values of  $X$  relate to higher values of  $Y$  (or vice versa), whereas causation says that *because* values of  $X$  are higher, values of  $Y$  are higher too. Regression analysis, as introduced in this chapter, establishes associations between  $X_i$ 's and  $Y$ , not causation, unless strong assumptions are made.

For example, recall the pizza delivery example from Appendix A.4. We have seen in several examples and exercises that the delivery time varies by driver. In fact, in a multiple linear regression model where  $Y$  is the delivery time, and the covariates are branch, bill, operator, driver, and temperature, we get significant differences of the mean delivery times of some drivers (e.g. Domenico and Bruno). Does this mean that *because* Domenico is the driver, the pizzas are delivered faster? Not necessarily. If Domenico drives the fastest scooter, then this may be the cause of his speedy deliveries. However, we have not measured the variable “type of scooter” and thus cannot take it into account. The result we obtain from the regression model is still useful because it provides the manager with useful hypotheses and predictions about his delivery times, but the interpretation that the driver’s abilities cause shorter delivery times may not necessarily be appropriate.

This example highlights that one of the most important assumptions to interpret regression parameters in a causal way is to have measured (and used) all variables  $X_i$  which affect both the outcome  $Y$  and the variable of interest  $A$  (e.g. the variable “driver” above). Another assumption is that the relationship between all  $X_i$ 's and  $Y$  is modelled correctly, for example non-linear if appropriate. Moreover, we need some *a priori* assumptions about how the variables relate to each other, and some technical assumptions need to be met as well. The interested reader is referred to Hernan and Robins (2017) for more details.

## 11.11 Key Points and Further Issues

### Note:

- ✓ If  $X$  is a continuous variable, the parameter  $\beta$  in the linear regression model  $Y = \alpha + \beta X + e$  can be interpreted as follows:

“For each unit of increase in  $X$ , there is an increase of  $\beta$  units in  $E(Y)$ .”

- ✓ If  $X$  can take a value of zero, then it can be said that

“For  $X = 0$ ,  $E(Y)$  equals  $\alpha$ .”

If  $X = 0$  is out of the data range or implausible, it does not make sense to interpret  $\alpha$ .

- ✓ A model is said to be linear when it is linear in its parameters.
- ✓ It is possible to include many different covariates in the regression model: they can be binary, log-transformed, or take any other form, and it is still a *linear* model. The interpretation of these variables is always conditional on the other variables; i.e. the reported association assumes that all other variables in the model are held fixed.
- ✓ To evaluate whether  $x_j$  is associated with  $y$ , it is useful to test whether  $\beta_j = 0$ . If the null hypothesis  $H_0 : \beta_j = 0$  is rejected (i.e. the  $p$ -value is smaller than  $\alpha$ ), one says that there is a *significant* association of  $x_j$  and  $y$ . However, note that a non-significant result does not necessarily mean there is no association, it just means we could not show an association.
- ✓ An important assumption in the multivariate linear model  $y = X\beta + e$  relates to the errors:

$$e \sim N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

To check whether the errors are (approximately) normally distributed, one can plot the standardized residuals in a histogram or a QQ-plot. The heteroscedasticity assumption (i.e.  $\sigma^2$  does not depend on  $i$ ) is tested by plotting the fitted values ( $\hat{y}_i$ ) against the standardized residuals ( $\hat{e}_i$ ). The plot should show no pattern (random plot).

- ✓ Different models can be compared by systematic hypothesis testing,  $R_{adj}^2$ , AIC, and other methods. Different methods may yield different results, and there is no unique best option.

## 11.12 Exercises

*Exercise 11.1* The body mass index (BMI) and the systolic blood pressure of 6 people were measured to study a cardiovascular disease. The data are as follows:

Body mass index	26	23	27	28	24	25
Systolic blood pressure	170	150	160	175	155	150

- (a) The research hypothesis is that a high BMI relates to a high blood pressure. Estimate the linear model where blood pressure is the outcome and BMI is the covariate. Interpret the coefficients.
- (b) Calculate  $R^2$  to judge the goodness of fit of the model.

*Exercise 11.2* A psychologist speculates that spending a lot of time on the internet has a negative effect on children's sleep. Consider the following data on hours of deep sleep ( $Y$ ) and hours spent on the internet ( $X$ ) where  $x_i$  and  $y_i$  are the observations on internet time and deep sleep time of the  $i$ th ( $i = 1, 2, \dots, 9$ ) child respectively:

Child $i$	1	2	3	4	5	6	7	8	9
Internet time $x_i$ (in h)	0.3	2.2	0.5	0.7	1.0	1.8	3.0	0.2	2.3
Sleep time $y_i$ (in h)	5.8	4.4	6.5	5.8	5.6	5.0	4.8	6.0	6.1

- (a) Estimate the linear regression model for the given data and interpret the coefficients.
- (b) Calculate  $R^2$  to judge the goodness of fit of the model.
- (c) Reproduce the results of a) and b) in  $R$  and plot the regression line.
- (d) Now assume that we only distinguish between spending more than 1 hour on the internet ( $X = 1$ ) and spending less than (or equal to) one hour on the internet ( $X = 0$ ). Estimate the linear model again and compare the results. How can  $\hat{\beta}$  now be interpreted? Describe how  $\hat{\beta}$  changes if those who spend more than one hour on the internet are coded as 0 and the others as 1.

*Exercise 11.3* Consider the following data on weight and height of 17 female students:

Student $i$	1	2	3	4	5	6	7	8	9
Weight $y_i$	68	58	53	60	59	60	55	62	58
Height $x_i$	174	164	164	165	170	168	167	166	160
Student $i$	10	11	12	13	14	15	16	17	
Weight $y$	53	53	50	64	77	60	63	69	
Height $x$	160	163	157	168	179	170	168	170	

- (a) Calculate the correlation coefficient of Bravais–Pearson (use  $\sum_{i=1}^n x_i y_i = 170,821$ ,  $\bar{x} = 166.65$ ,  $\bar{y} = 60.12$ ,  $\sum_{i=1}^n y_i^2 = 62,184$ ,  $\sum_{i=1}^n x_i^2 = 472,569$ ). What does this imply with respect to a linear regression of height on weight?

- (b) Now estimate and interpret the linear regression model where “weight” is the outcome.
- (c) Predict the weight of a student with a height 175 cm.
- (d) Now produce a scatter plot of the data (manually or by using *R*) and interpret it.
- (e) Add the following two points to the scatter plot  $(x_{18}, y_{18}) = (175, 55)$  and  $(x_{19}, y_{19}) = (150, 75)$ . Speculate how the linear regression estimate will change after adding these points.
- (f) Re-estimate the model using all 19 observations and  $\sum x_i y_i = 191,696$  and  $\sum x_i^2 = 525,694$ .
- (g) Given the results of the two regression models: What are the general implications with respect to the least squares estimator of  $\beta$ ?

*Exercise 11.4* To study the association of the monthly average temperature (in  $^{\circ}\text{C}$ ,  $X$ ) and hotel occupation (in %,  $Y$ ), we consider data from three cities: Polenca (Mallorca, Spain) as a summer holiday destination, Davos (Switzerland) as a winter skiing destination, and Basel (Switzerland) as a business destination.

Month	Davos		Polenca		Basel	
	$X$	$Y$	$X$	$Y$	$X$	$Y$
Jan	-6	91	10	13	1	23
Feb	-5	89	10	21	0	82
Mar	2	76	14	42	5	40
Apr	4	52	17	64	9	45
May	7	42	22	79	14	39
Jun	15	36	24	81	20	43
Jul	17	37	26	86	23	50
Aug	19	39	27	92	24	95
Sep	13	26	22	36	21	64
Oct	9	27	19	23	14	78
Nov	4	68	14	13	9	9
Dec	0	92	12	41	4	12

- (a) Interpret the following regression model output where the outcome is “hotel occupation” and “temperature” is the covariate.

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	50.33459	7.81792	6.438	2.34e-07 ***
X	0.07717	0.51966	0.149	0.883

- (b) Interpret the following output where “city” is treated as a covariate and “hotel occupation” is the outcome.

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	48.3333	7.9457	6.083	7.56e-07 ***
cityDavos	7.9167	11.2369	0.705	0.486
cityPolenca	0.9167	11.2369	0.082	0.935

- (c) Interpret the following output and compare it with the output from b):

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
city	2	450.1	225.03	0.297	0.745
Residuals	33	25001.2	757.61		

- (d) In the following multiple linear regression model, both “city” and “temperature” are treated as covariates. How can the coefficients be interpreted?

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	44.1731	10.9949	4.018	0.000333 ***
X	0.3467	0.6258	0.554	0.583453
cityDavos	9.7946	11.8520	0.826	0.414692
cityPolenca	-1.1924	11.9780	-0.100	0.921326

- (e) Now consider the regression model for hotel occupation and temperature fitted separately for each city: How can the results be interpreted and what are the implications with respect to the models estimated in (a)–(d)? How can the models be improved?

Davos:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	73.9397	4.9462	14.949	3.61e-08 ***
X	-2.6870	0.4806	-5.591	0.000231 ***

Polenca:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-22.6469	16.7849	-1.349	0.20701
X	3.9759	0.8831	4.502	0.00114 **

Basel:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	32.574	13.245	2.459	0.0337 *
X	1.313	0.910	1.443	0.1796

- (f) Describe what the design matrix will look like if city, temperature, and the interaction between them are included in a regression model.

- (g) If the model described in (f) is fitted the output is as follows:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	32.5741	10.0657	3.236	0.002950 **
X	1.3133	0.6916	1.899	0.067230 .
cityDavos	41.3656	12.4993	3.309	0.002439 **
cityPolenca	-55.2210	21.0616	-2.622	0.013603 *
X:cityDavos	-4.0003	0.9984	-4.007	0.000375 ***
X:cityPolenca	2.6626	1.1941	2.230	0.033388 *

Interpret the results.

- (h) Summarize the association of temperature and hotel occupation by city—including 95 % confidence intervals—using the interaction model. The covariance matrix is as follows:

	(Int.)	X	Davos	Polenca	X:Davos	X:Polenca
(Int.)	101.31	-5.73	-101.31	-101.31	5.73	5.73
X		-5.73	0.47	5.73	5.73	-0.47
Davos	-101.31	5.73	156.23	101.31	-9.15	-5.73
Polenca	-101.31	5.73	101.31	443.59	-5.73	-22.87
X:Davos		5.73	-0.47	-9.15	-5.73	0.99
X:Polenca		5.73	-0.47	-5.73	-22.87	0.47
						1.42

*Exercise 11.5* The theatre data (see Appendix A.4) describes the monthly expenditure on theatre visits of 699 residents of a Swiss city. It captures not only the expenditure on theatre visits (in SFR) but also age, gender, yearly income (in 1000 SFR), and expenditure on cultural activities in general as well as expenditure on theatre visits in the preceding year.

- (a) The summary of the multiple linear model where expenditure on theatre visits is the outcome is as follows:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-127.22271	19.15459	-6.642	6.26e-11 ***
Age	0.39757	0.19689	[1]	[2]
Sex	22.22059	5.22693	4.251	2.42e-05 ***
Income	1.34817	0.20947	6.436	2.29e-10 ***
Culture	0.53664	0.05053	10.620	<2e-16 ***
Theatre_ly	0.17191	0.11711	1.468	0.1426

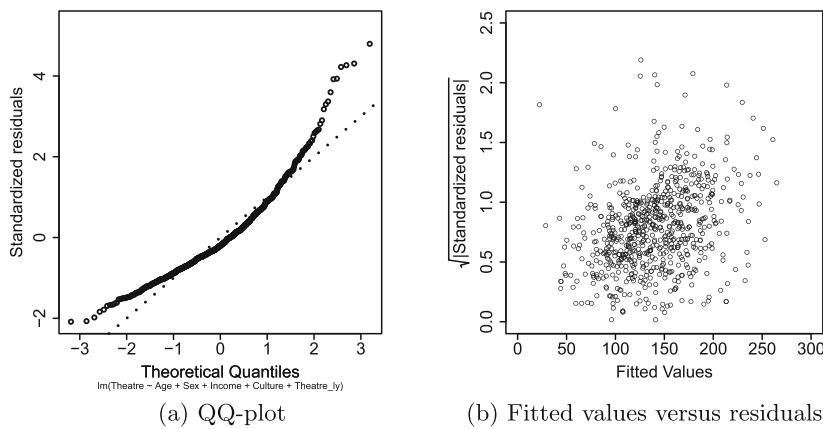
How can the missing values [1] and [2] be calculated?

- (b) Interpret the model diagnostics in Fig. 11.11.  
(c) Given the diagnostics in (b), how can the model be improved? Plot a histogram of theatre expenditure in *R* if you need further insight.  
(d) Consider the model where theatre expenditure is log-transformed:

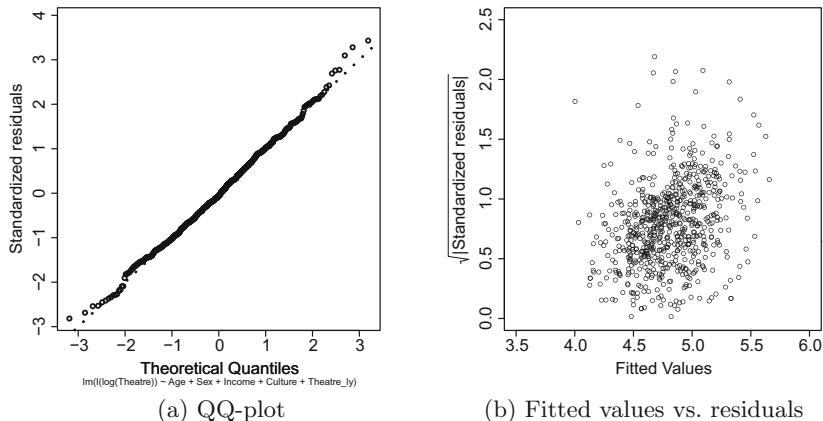
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.9541546	0.1266802	23.320	< 2e-16 ***
Age	0.0038690	0.0013022	2.971	0.00307 **
Sex	0.1794468	0.0345687	5.191	2.75e-07 ***
Income	0.0087906	0.0013853	6.346	4.00e-10 ***
Culture	0.0035360	0.0003342	10.581	< 2e-16 ***
Theatre_ly	0.0013492	0.0007745	1.742	0.08197 .

How can the coefficients be interpreted?



**Fig. 11.11** Checking the model assumptions



**Fig. 11.12** Checking the model assumptions

- (e) Judge the quality of the model from d) by means of Figs. 11.12a and 11.12b. What do they look like compared with those from b)?

*Exercise 11.6* Consider the pizza delivery data described in Appendix A.4.

- (a) Read the data into *R*. Fit a multiple linear regression model with delivery time as the outcome and temperature, branch, day, operator, driver, bill, number of ordered pizzas, and discount customer as covariates. Give a summary of the coefficients.

- (b) Use  $R$  to calculate the 95 % confidence intervals of all coefficients. Hint: the standard errors of the coefficients can be accessed either via the covariance matrix or the model summary.
- (c) Reproduce the least squares estimate of  $\sigma^2$ . Hint: use `residuals()` to obtain the residuals.
- (d) Now use  $R$  to estimate both  $R^2$  and  $R^2_{adj}$ . Compare the results with the model output from a).
- (e) Use backward selection by means of the `stepAIC` function from the library MASS to find the best model according to AIC.
- (f) Obtain  $R^2_{adj}$  from the model identified in e) and compare it to the full model from a).
- (g) Identify whether the model assumptions are satisfied or not.
- (h) Are all variables from the model in (e) causing the delivery time to be either delayed or improved?
- (i) Test whether it is useful to add a quadratic polynomial of temperature to the model.
- (j) Use the model identified in (e) to predict the delivery time of the last captured delivery (i.e. number 1266). Use the `predict()` command to ease the calculation of the prediction.

→ Solutions of all exercises in this chapter can be found on p. [409](#)

\*Source Toutenburg, H., Heumann, C., *Deskriptive Statistik*, 7th edition, 2009, Springer, Heidelberg

---

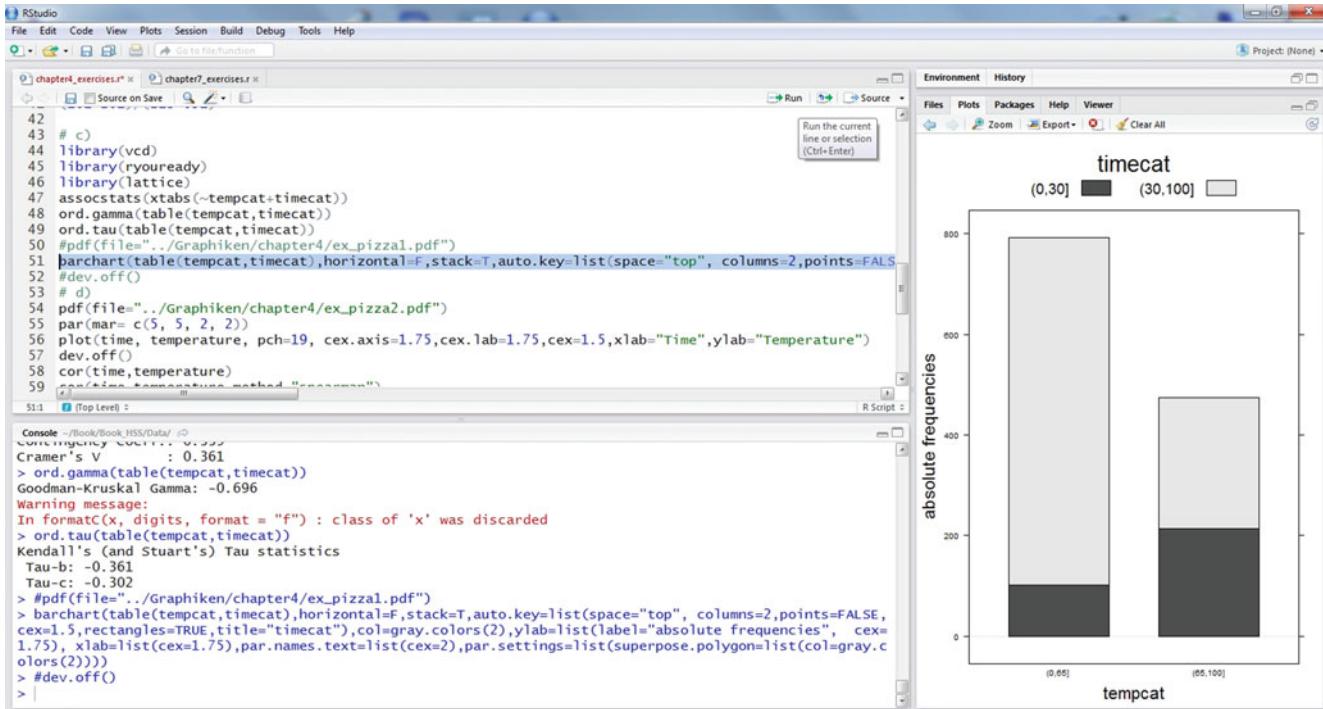
### Background

The open-source software *R* was developed as a free implementation of the language *S* which was designed as a language for statistical computation, statistical programming, and graphics. The main intention was to allow users to explore data in an easy and interactive way, supported by meaningful graphical representations. The statistical software *R* was originally created by Ross Ihaka and Robert Gentleman (University of Auckland, New Zealand).

---

### Installation and Basic Functionalities

- The “base” *R* version, i.e. the software with its most relevant commands, can be downloaded from <https://www.r-project.org/>. After installing *R*, it is recommended to install an editor too. An editor allows the user to conveniently save and display *R*-code, submit this code to the *R* console (i.e. the *R* software itself), and control the settings and the output. A popular choice of editor is *RStudio* (free of charge) which can be downloaded from <https://www.rstudio.com/> (see Fig. A.1 for a screenshot of the software). There are alternative good editors, for example “Tinn-R” (<http://sourceforge.net/projects/tinn-r/>).
- A lot of additional user-written packages are available online and can be installed within the *R* console or using the *R* menu. Within the console, the `install.packages("package to install")` function can be used. Please note that an internet connection is required.



**Fig.A.1** Screenshot of R Studio with the command window (*top left*), the output window (“console”, i.e. R itself, *bottom left*), and a plot window (*right*). Other windows (e.g. about the environment or package updates) are closed

Statistics has a close relationship to algebra: data sets can be seen as matrices, and variables as vectors. R makes use of these structures and this is why we first introduce data structure functionalities before explaining some of the most relevant basic statistical commands.

## R as a Calculator, Basic Data Structures and Arithmetic Operations

- The character # marks the beginning of a comment. All characters until the end of the *line* are ignored by R. We use # to comment on our R-code.
- If we know the name of a command we would like to use, and we want to learn about its functionality, typing ?command in the R command line prompt displays a help page, e.g.

```
?mean
```

displays a help page for the arithmetic mean function.

- Using

```
example(mean)
```

shows application examples of the respective function.

- The command c(1,2,3,4,5) combines the numbers 1, 2, 3, 4 and 5 into a vector (e.g. a variable).
- Vectors can be assigned to an “object”. For example,

```
X <- c(2,12,22,32)
```

assigns a numeric vector of length 4 to the object X. In general, the arrow sign (<-) is a very important concept to store data, summaries, and outputs in objects (i.e. the name in the front of the <- sign). Note that R is case sensitive; i.e. X and x are two different variable names. Instead of “<-”, one can also use “=”.

- Sequences of numbers can be created using the seq and rep commands. For example,

```
seq(1,10)
```

and

```
rep(1,10)
```

yield

```
[1] 1 2 3 4 5 6 7 8 9 10
```

and

```
[1] 1 1 1 1 1 1 1 1 1 1
```

respectively.

- Basic data structures are **vectors**, **matrices**, **arrays**, **lists**, and **data frames**. They can contain numeric values, logical values or even characters (strings). In the latter case, arithmetic operations are not allowed.

- A numeric vector of length 5 can be constructed by the command

```
x <- vector(mode="numeric", length=5)
```

The elements can be accessed by squared brackets: [ ]. For example,

```
x[3] <- 4
```

assigns the value 4 to the third element of the object x. Logical vectors containing the values TRUE and FALSE are also possible. In arithmetic operations, TRUE corresponds to 1 and FALSE corresponds to 0 (which is the default). Consider the following example:

```
x.log <- vector(mode="logical", length=4)
x.log[1] = x.log[3] = x.log[4] = TRUE
mean(x.log)
```

returns as output 0.75 because the mean of (1, 0, 1, 1) = (TRUE, FALSE, TRUE, TRUE) is 0.75.

- A matrix can be constructed by the **matrix()** command:

```
x <- matrix(nrow=4, ncol=2, data=1:8, byrow=T)
```

creates a  $4 \times 2$  matrix, where the data values are the natural numbers 1, 2, 3, ..., 8 which are stored row-wise in the matrix,

[,1]	[,2]
[1,]	1    2
[2,]	3    4
[3,]	5    6
[4,]	7    8

because of the parameter **byrow=T** (which is equivalent to **byrow=TRUE**). The default is **byrow=F** which would store the data column-wise.

- Arrays are more general data structures than vectors and matrices in the sense that they can have more than two dimensions. For instance,

```
x <- array(data=1:12, dim=c(3,2,2))
```

creates a three-dimensional array with  $3 \cdot 2 \cdot 2 = 12$  elements.

- A list can contain objects of different types. For example, a list element can be a vector or matrix. Lists can be initialized by the command **list** and can grow dynamically. It is important to understand that list elements should be accessed by the name of the entry via the dollar sign or using double brackets:

```

x <- list(one=c(1,2,3,4,5),two=c("Hello", "world", "!"))
x
$one
[1] 1 2 3 4 5

$two
[1] "Hello" "world" "!"

x[[2]]
[1] "Hello" "world" "!"
x$one
[1] 1 2 3 4 5

```

- A data frame is the standard data structure for storing a data set with rows as observations and columns as variables. Many statistical procedures in R (such as the `lm` function for linear models) expect a data frame. A data frame is conceptually not much different from a matrix and can either be initialized by reading a data set from an external file or by binding several column vectors. As an example, we consider three variables (age, favourite hobby, and favourite animal) each with five observations:

```

age     <- c(25,33,30,40,28)
hobby   <- c("Reading","Sports","Games","Reading","Games")
animal  <- c("Elephant", "Giraffe", NA, "Monkey", "Cat")
dat     <- data.frame(age,hobby,animal)
names(dat) <- c("Age","Favourite.hobby","Favourite.animal")
dat

```

The resulting output is

```

> dat
  Age Favourite.hobby Favourite.animal
1  25      Reading      Elephant
2  33      Sports       Giraffe
3  30      Games        <NA>
4  40      Reading      Monkey
5  28      Games        Cat

```

where `NA` denotes a missing value. With `write.table` or a specialized version thereof such as `write.csv` (for writing the data in a file using comma-separated fields), a data frame can be saved in a file. The command sequence

```

write.csv(x=dat,file="toy.csv",row.names=FALSE)
read.dat <- read.csv(file="toy.csv")
read.dat

```

---

saves the data frame as an external (comma-separated) file and then loads the data again using `read.csv`.

Individual elements of the data frame can be accessed using squared brackets, as for matrices. For example, `dat[1, 2]` returns the first observation of the second variable column for the data set `dat`. Individual columns (variables) can also be selected using the `$` sign:

```
dat$Age
```

returns the age column:

```
[1] 25 33 30 40 28
```

- The `factor` command is very useful to store nominal variables, and the command `ordered` is ideal for ordinal variables. Both commands are extremely important since factor variables with more than two categories are automatically expanded into several columns of dummy variables if necessary, e.g. if they are included as covariates in a linear model. In the previous paragraph, two factor variables have already been created. This can be confirmed by typing

```
is.factor(dat$Favourite.hobby)
is.factor(dat$Favourite.animal)
```

which return the value `TRUE`. Have a look at the following two factor variables:

```
sex <- factor("female", "male", "male", "female", "female")
grade <- ordered(c("low", "medium", "low", "high", "high"),
                 levels=c("low", "medium", "high"))
```

Please note that by default alphabetical order is used to order the categories (e.g. female is coded as 1 and male as 2). However, the mapping of integers to strings can be controlled by the user as seen for the variable “`grade`”:

```
grade
```

returns

```
[1] low     medium  low     high    high
Levels: low < medium < high
```

- Basic arithmetic operations can be applied directly to a numeric vector. Basic operations are addition `+`, subtraction `-`, multiplication `*` and division `/`, integer division `%/%`, modulo operation `%%`, and exponentiation with two possible notations: `**` or `^`. Examples are given as:

```

2^3           # command
[1] 8          # output
2**3          # command
[1] 8          # output
2^0.5          # command
[1] 1.414214   # output
c(2,3,5,7)^2  # command: application to a vector
[1] 4 9 25 49 # output
c(2,3,5,7)^c(2,3) # command: !!! ATTENTION!
[1] 4 27 25 343 # output
c(1,2,3,4,5,6)^c(2,3,4) # command
[1] 1 8 81 16 125 1296 #output
c(2,3,5,7)^c(2,3,4) # command: !!! WARNING MESSAGE!
[1] 4 27 625 49
Warning message:
longer object length
  is not a multiple of shorter object length
  in: c(2, 3, 5, 7)^c(2, 3, 4)

```

The last four commands show the “recycling property” of R. It tries to match the vectors with respect to the length if possible. In fact,

`c(2,3,5,7)^c(2,3)`

is expanded to

`c(2,3,5,7)^c(2,3,2,3)`

The last example shows that R gives a warning if the length of the shorter vector cannot be expanded to the length of the longer vector by a simple multiplication with a natural number (2, 3, 4, ...). Here

`c(2,3,5,7)^c(2,3,4)`

is expanded to

`c(2,3,5,7)^c(2,3,4,2)`

such that not all elements of

```
c(2,3,4)
```

are “recycled”.

## More on indexing

The standard ways of accessing/indexing elements in a vector, matrix, list, or data frame have already been introduced above, but *R* allows a lot more flexible accessing of elements.

1. Selecting elements using vectors of positive numbers (letters and LETTERS show the 26 letters of the alphabet)

```
letters[1:3]
letters[ c(2,4,6) ]

[1] "a" "b" "c"
[1] "b" "d" "f"
```

2. Selecting elements using logical vectors

```
x <- 1:10          # numbers 1 to 10
x[ (x>5) ]        # selecting any number >5
x[ (x%%2==0) ]     # numbers that are divisible by 2
x[ (x%%2==1) ]     # numbers that are not divisible by 2
x[5] <- NA         # 5th element of x is defined
                     # to be missing (NA)

x
y <- x[!is.na(x)] # all x which are not missing
y
```

returns the output

```
[1]  6  7  8  9 10
[1]  2  4  6  8 10
[1]  1  3  5  7  9
[1]  1  2  3  4 NA  6  7  8  9 10
[1]  1  2  3  4  6  7  8  9 10
```

### 3. Selecting (deleting) elements using negative numbers

```
x <- 1:10  
x[-(1:5)] # x, but delete first five entries of x
```

returns the output

```
[1] 6 7 8 9 10
```

because the first five elements have been removed.

### 4. Selecting elements using characters

```
x <- c(Water=1, Juice=2, Lemonade=3 )  
names(x)  
x["Juice"]
```

returns the output

```
[1] "Water"    "Juice"      "Lemonade"  
Juice  
2
```

## Standard Functions

Some standard functions and their roles in *R* are

<code>abs()</code>	Absolute value
<code>sqrt()</code>	Square root
<code>round(), floor(), ceiling()</code>	Rounding, up and down
<code>sum(), prod()</code>	Sum and product
<code>log(), log10(), log2()</code>	Logarithms
<code>exp()</code>	Exponential function
<code>sin(), cos(), tan(),</code>	Trigonometric functions
<code>asin(), acos(), atan()</code>	
<code>sinh(), cosh(), tanh(),</code>	Hyperbolic functions
<code>asinh(x), acosh(), atanh(x)</code>	

All functions can again be applied directly to numeric vectors.

## Statistical Functions

Some statistical functions and their roles in *R* are

<code>mean()</code> , <code>var()</code>	Mean and variance
<code>cov()</code> , <code>cor()</code>	Covariance and correlation
<code>min()</code> , <code>max()</code>	Minimum and maximum

*Note:* the arguments of the functions vary depending on the chosen method. For example, the `mean()` function can be applied to general *R* objects where averaging makes sense (numeric or logical vectors, but also, e.g. matrices). The functions `var()`, `cov()`, `cor()` expect one or two numeric vectors, matrices, or data frames. Minimum and maximum functions work also with a comma-separated list of values, i.e.

```
min(2, 6.7, 1.2, 8.0)
```

provides the same result (1.2) as

```
min(c(2, 6.7, 1.2, 8.0))
```

Examples:

```
mean( c(1,2,5,6) )
[1] 3.5
```

```
var( c(1,2,5,6) )
[1] 5.666667
```

Note that `var()`, `cov()` use the factor  $1/(n - 1)$  for the unbiased estimate of the variance instead of  $1/n$  for the empirical variance, i.e.  $1/3$  in the example above. Both functions can also be applied to several vectors or a matrix. Then the covariance matrix (and correlation matrix in case of `cor()`) is computed. For example, consider two variables

```
age.v <- c(25,30,35,40)
income.v <- c(2000, 2500, 2800, 3200)
```

Then both commands return the symmetric covariance matrix (with variances as the diagonal entries and covariances as the non-diagonal entries).

```
var(cbind(age.v, income.v))
      age.v income.v
age.v     41.66667   3250.0
income.v 3250.00000 255833.3
```

```
cov(cbind(age.v, income.v))
      age.v income.v
age.v     41.66667   3250.0
income.v 3250.00000 255833.3
```

The (Pearson) correlation between the two variables is calculated as 0.9954293.

```
cor(cbind(age.v, income.v))
      age.v income.v
age.v     1.0000000 0.9954293
income.v 0.9954293 1.0000000
```

The Spearman rank correlation is perfectly 1, since both vectors are in increasing order:

```
cor(cbind(age.v, income.v), method="spearman")
      age.v income.v
age.v       1         1
income.v    1         1
```

## More Useful Functions

Some more commonly used standard functions and their roles in *R* are as follows:

- Cumulative sum and product:

```
x <- c( 1,3, 2, 5)
cumsum(x)          # 1, 1+3, 1+3+2, 1+3+2+5
cumprod(x)         # 1, 1*3, 1*3*2, 1*3*2*5
```

give the output

```
[1]  1  4  6 11
[1]  1  3  6 30
```

- Factorial:

```
factorial(5)
```

returns  $5!$  as

```
[1] 120
```

- Binomial coefficient  $\binom{n}{k}$ :

```
choose(4,2)
```

returns  $\binom{4}{2}$  as

```
[1] 6
```

## Mathematical Constants

The number  $\pi$  is a mathematical constant, the ratio of a circle's circumference to its diameter, and is commonly approximated as 3.14159. It is directly available in R as `pi`.

```
pi
[1] 3.141593
```

Other “constants” are

<code>Inf</code> , <code>-Inf</code>	$\infty, -\infty$
<code>NaN</code>	Not a Number: e.g. $0/0$ [1] <code>NaN</code>
<code>NA</code>	Not Available: missing values
<code>NULL</code>	empty set

## Assignment Operator for Creating Functions

The assignment operator `<-` (“less than” sign followed by hyphen) has already been introduced above in the context of variables. Alternatively, `=` (equality sign) can be used. One can create one's own functions: the function is an object with a name which takes values specified in the round brackets and returns what has been specified in the curly braces. For example, the following function `myfunction` returns a polynomial of degree 3 for a given argument `x`. Note that by default all four coefficients are equal to 1.

---

```

my.function <- function(x,a=1,b=1,c=1,d=1){
  h <- a+b*x+c*x^2+d*x^3
  return(h)
}

my.function(2)
[1] 15

my.function(x=2, a=4, b=3)
[1] 22

```

## Loops and Conditions

The concept of loops is convenient when some operation has to be repeated. Loops can be utilized in various ways, for example, via `for` or `while`. Conditions are specified with the `if` statement. For example,

```

x <- 1:10
for(i in 1:10){ if(x[i]>5){x[i] <- x[i]+i}
}
x

```

returns

```
[1] 1 2 3 4 5 12 14 16 18 20
```

In this example, `x` is a vector with 10 elements: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. For each element `x[i]`, we replace `x[i]` with `x[i]+i` if the condition `x[i]>5` is true; otherwise we do not.

---

## Statistical Functions

Now we consider some basic statistical functions in *R*. For illustration, we use the `painters` data in the following example. This data is available after loading the library MASS (only a subset is shown below). The data lists the subjective assessment, on a 0 to 20 integer scale, of 54 classical painters. The painters were assessed on four characteristics: composition, drawing, colour, and expression. The data is due to the eighteenth-century art critic, de Piles. Use `?painters` for more information on the data set.

```

library(MASS)
painters

```

shows

	Composition	Drawing	Colour	Expression	School
Da Udine	10	8	16	3	A
Da Vinci	15	16	4	14	A
Del Piombo	8	13	16	7	A
Del Sarto	12	16	9	8	A
Fr. Penni	0	15	8	0	A

## The Summary Function

The summary function allows a quick overview of a data frame. For numeric variables, the five-point summary (which is also used in a simple box plot, see Sect. 3.3) is calculated together with the arithmetic mean. For factor variables, the absolute frequencies of the most frequent categories are printed. If the factor has more than six categories, the other categories are summarized in a separate category—Other.

```
summary(painters)
```

yields

Composition	Drawing	...	School
Min. : 0.00	Min. : 6.00	...	A : 10
1st Qu.: 8.25	1st Qu.: 10.00	...	D : 10
Median : 12.50	Median : 13.50	...	E : 7
Mean : 11.56	Mean : 12.46	...	G : 7
3rd Qu.: 15.00	3rd Qu.: 15.00	...	B : 6
Max. : 18.00	Max. : 18.00	...	C : 6
		...	(Other) : 8

The summary function can also be applied to a single variable:

```
summary(painters$School)
```

returns

A	B	C	D	E	F	G	H
10	6	6	10	7	4	7	4

## Accessing Subgroups in Data Frames

Subgroups, i.e. groups of observations which share the same feature(s) of one or several variables, can be accessed using the `subset` command.

```
subset(painters, School=="F")
```

accesses all painters for which `School=="F"` holds.

	Composition	Drawing	Colour	Expression	School
Durer	8	10	10	8	F
Holbein	9	10	16	13	F
Pourbus	4	15	6	6	F
Van Leyden	8	6	6	4	F

This is a more elegant method than selecting these observations by specifying a condition in squared brackets via the `[rows,columns]` argument.

```
painters[ painters[["School"]] == "F", ]
```

Note that the explained structure is an important one: we access the rows and columns of a matrix or data set by using the `[rows,columns]` argument. Here we access all rows for which the variable “school” is “F”. If, in addition, we also want to restrict the data set to the first two variables, we can write:

```
painters[ painters[["School"]] == "F", c(1,2)]
```

Similarly,

```
subset(painters, Composition <= 6)
```

gives the output

	Composition	Drawing	Colour	Expression	School
Fr. Penni	0	15	8	0	A
Perugino	4	12	10	4	A
Bassano	6	8	17	0	D
Bellini	4	6	14	0	D
Murillo	6	8	15	4	D
Palma Vecchio	5	6	16	0	D
Caravaggio	6	6	16	0	E
Pourbus	4	15	6	6	F

Uninteresting columns can be eliminated using negative indexing. For instance, in the following example,

```
subset(painters, School=="F", select=c(-3,-5) )
```

	Composition	Drawing	Expression
Durer	8	10	8
Holbein	9	10	13
Pourbus	4	15	6
Van Leyden	8	6	4

the third and the fifth columns (Colour and School) are not shown.

The operator `%in%` allows for more complex searches. For instance,

```
subset(painters, Drawing %in% c(6,7,8,9) & Composition==10)
```

returns the following output:

	Composition	Drawing	Colour	Expression	School
Da Udine	10	8	16	3	A
J. Jordaens	10	8	16	6	G
Bourdon	10	8	8	4	H

i.e. those painters with a drawing score between 6 and 9 (= any number which matches 6, or 7, or 8, or 9).

## Stratifying a Data Frame and Applying Commands to a List

Sometimes it is of interest to apply statistical commands (such as `summary`) to several subgroups. If this is the case, the data is partitioned into different groups using `split` and then `lapply` applies a function to each of these groups. The command `split` partitions the data set by values of a specific variable. For example, we first stratify the `painters` data with respect to the painter's school:

```
splitted <- split(painters, painters$School)
splitted
```

`$A`

	Composition	Drawing	Colour	Expression	School
Da Udine	10	8	16	3	A
Da Vinci	15	16	4	14	A
Del Piombo	8	13	16	7	A
Del Sarto	12	16	9	8	A
Fr. Penni	0	15	8	0	A
Guilio Romano	15	16	4	14	A

---

Michelangelo	8	17	4	8	A
Perino del Vaga	15	16	7	6	A
Perugino	4	12	10	4	A
Raphael	17	18	12	18	A

\$B

	Composition	Drawing	Colour	Expression	School
F. Zucarro	10	13	8	8	B
Fr. Salviata	13	15	8	8	B
Parmigiano	10	15	6	6	B
Primiticcio	15	14	7	10	B
T. Zucarro	13	14	10	9	B
Volterra	12	15	5	8	B

\$C

...

Note, that `splitted` is now a list,

```
is.list(splitted)
```

returns

```
[1] TRUE
```

while the objects `splitted$A` to `splitted$H` are data frames.

```
is.data.frame(splitted$A)
```

returns

```
[1] TRUE
```

Secondly, as indicated above, the command `lapply` allows us to apply a function to a list. For instance,

```
lapply(splitted, summary)
```

applies the `summary` function to all data frames in the list `splitted` (output not shown). See also `?apply`, `?sapply`, `?tapply`, and `?mapply` for similar operations.

## Sorting, Ranking, Finding Duplicates, and Unique Values

- Sorting a vector:

```
x <- c( 1,3, 2, 5)
sort(x)
sort(x, decreasing=TRUE)
```

returns the ordered values in decreasing order as

```
[1] 5 3 2 1
```

See also the command `order` for showing the order of vector elements.

- Calculating ranks;

```
x <- c( 10,30, 20, 50, 20)
rank(x)
```

returns the following output:

```
[1] 1.0 4.0 2.5 5.0 2.5
```

- Finding duplicate values:

```
x <- c( 1,3, 2, 5, 2)
duplicated(x)
```

indicates which values occur more than once:

```
[1] FALSE FALSE FALSE FALSE TRUE
```

- Removing duplicates:

```
x <- c( 1,3, 2, 5, 2)
unique(x)
```

shows the output as

```
[1] 1 3 2 5
```

This means `unique` finds out how many *different* values a vector has.

## Categorizing Numeric Variables

Continuous variables (vectors) can be categorized using the `cut` command.

```
x <- c(1.3, 1.5, 2.5, 3.8, 4.1, 5.9, 7.1, 8.4, 9.0)
xdiscrete <- cut(x, breaks=c(-Inf, 2, 5, 8, Inf) )
is.factor(xdiscrete)
xdiscrete
table(xdiscrete)
```

returns

```
[1] TRUE
[1] (-Inf,2] (-Inf,2] (2,5] (2,5] (2,5] (5,8] (5,8] (8,Inf]
[9] (8,Inf]
Levels: (-Inf,2] (2,5] (5,8] (8,Inf]
(-Inf,2] (2,5] (5,8] (8,Inf]
2           3           2           2
```

## Random Variables

- R has built-in functions for several probability density/mass functions (PMF/PDF), (probability) distribution function (i.e. the CDF), quantile functions and for generating random numbers.
- The function names use the following scheme:

First letter	Function	Further letters
d	density	distribution name
p	probability	distribution name
q	quantiles	distribution name
r	random number	distribution name

- Examples:

```
dnorm(x=0)
[1] 0.3989423
```

returns the value of the density function (i.e.  $P(X = x)$ ) of a  $N(0, 1)$ -distribution at  $x = 0$ , which is  $1/\sqrt{2\pi}$ .

```
pnorm(q=0)
pnorm(q=1.96)
[1] 0.5
[1] 0.9750021
```

---

returns the value of the CDF of a  $N(0, 1)$ -distribution at  $q$ , i.e.  $\Phi(q) = P(X \leq q)$ .

```
qnorm(p=0.95)
```

returns the value

```
[1] 1.644854
```

which is the 95 % quantile of a  $N(0, 1)$ -distribution.

```
X <- rnorm(n=4)
X
```

returns a vector of four normal random numbers of a  $N(0, 1)$ -distribution:

```
[1] -0.90826678 -0.09089654 -0.47679821 1.22137230
```

Note that a repeated application of this function leads to different random numbers. To get a reproducible sequence of random numbers, a `seed` value needs to be set:

```
set.seed(89234512)
X <- rnorm(n=4)
X
```

If all three commands are executed, then the sequence is (using the standard random generator)

```
[1] -1.07628865 0.37797715 0.04925738 -0.22137107
```

- The following functions for distributions can be used:

### *Model distributions*

Function	Distribution
<code>beta</code>	Beta
<code>binom</code>	Binomial
<code>cauchy</code>	Cauchy
<code>exp</code>	Exponential
<code>gamma</code>	Gamma
<code>geom</code>	Geometric
<code>hyper</code>	Hypergeometric
<code>lnorm</code>	Log-normal
<code>norm</code>	Normal
<code>pois</code>	Poisson
<code>unif</code>	Uniform
<code>mvtnorm</code>	Multivariate normal (in package <code>mvtnorm</code> )

### Test distributions

Function	Distribution
chisq	$\chi^2$
f	F
signrank	Wilcoxon signed rank (1 sample)
t	t
wilcox	Wilcoxon rank sum (2 samples)

- For convenience, we list a few important PDF and CDF values in Sect. C.

## Key Points and Further Issues

### Note:

- ✓ R uses the following data structures: vectors, matrices, arrays, lists, and data frames.
- ✓ Entries of matrices and data frames can be accessed using squared brackets. For example, `data[1:5, 2]` refers to the first five observations of the second column (variable) of the data. Variables of data frames can also be accessed via the \$ sign, e.g. via `data$variable`.
- ✓ If operations such as statistical analyses have to be repeated on several subgroups, using `split` together with `lapply` is a viable option. Alternatively, loops (e.g. `for`) together with conditions (such as `if`) can be used.
- ✓ R contains the most relevant statistical functions needed for descriptive and inductive analyses (as shown throughout the book). User-written packages can be installed using `install.packages(''package_name'')`.
- ✓ Readers who are interested in learning more about R are referred to Albert and Rizzo (2012), Crawley (2013), Dalgaard (2008), Ligges (2008), and Everitt and Hothorn (2011).

## Data Sets

From the data used in this book, we publish some relevant data sets, along with solutions of the *R*-exercises, on <https://chris.userweb.mwn.de/book/>. The important data sets are explained in the following paragraphs.

## Pizza Delivery Data

The pizza delivery data (`pizza_delivery.csv`, see also Table A.1) is a simulated data set. The data refers to an Italian restaurant which offers home delivery of pizza. It contains the orders received during a period of one month: May 2014. There are three branches of the restaurant. The pizza delivery is centrally managed: an operator receives a phone call and forwards the order to the branch which is nearest to the customer's address. One of the five drivers (two of whom only work part time at the weekend) delivers the order. The data set captures the number of pizzas ordered as well as the final bill (in €) which may also include drinks, salads, and pasta dishes. The owner of the business observed an increased number of complaints, mostly because pizzas arrive too late and too cold. To improve the service quality of his business, the owner wants to measure (i) the time from call to delivery and (ii) the pizza temperature at arrival (which can be done with a special device). Ideally, a pizza arrives within 30 min of the call; if it takes longer than 40 min, then the customers are promised a free bottle of wine (which is not always handed out though). The temperature of the pizza should be above 65 °C at the time of delivery. The analysis of the data aims to determine the factors which influence delivery time and temperature of the pizzas.

**Table A.1** First few rows of the pizza delivery data

	day	date	time	operator	branch	driver	temperature
1	Thursday	1 May 2014	35.1	Laura	East	Bruno	68.3
2	Thursday	1 May 2014	25.2	Melissa	East	Salvatore	71.0
3	Thursday	1 May 2014	45.6	Melissa	West	Salvatore	53.4
4	Thursday	1 May 2014	29.4	Melissa	East	Salvatore	70.3
5	Thursday	1 May 2014	30.0	Melissa	West	Salvatore	71.5
6	Thursday	1 May 2014	40.3	Melissa	Centre	Bruno	60.8

	bill	pizzas	free_wine	got_wine	discount_customer
1	58.4	4	0	0	1
2	26.4	2	0	0	0
3	58.1	3	1	0	0
4	35.2	3	0	0	0
5	38.4	2	0	0	0
6	61.8	4	1	1	0

**Table A.2** First few rows of the decathlon data from the 2004 Olympic Games in Athens data

	100m	Long.jump	Shot.put	High.jump	400m
Roman Sebrle	10.85	7.84	16.36	2.12	48.36
Bryan Clay	10.44	7.96	15.23	2.06	49.19
Dmitriy Karpov	10.50	7.81	15.93	2.09	46.81
Dean Macey	10.89	7.47	15.73	2.15	48.97
Chiel Warners	10.62	7.74	14.48	1.97	47.97
Attila Zsivoczky	10.91	7.14	15.31	2.12	49.40
...					
	110m.hurdle	Discus	Pole.vault	Javelin	1500m
Roman Sebrle	14.05	48.72	5.0	70.52	280.01
Bryan Clay	14.13	50.11	4.9	69.71	282.00
Dmitriy Karpov	13.97	51.65	4.6	55.54	278.11
Dean Macey	14.56	48.34	4.4	58.46	265.42
Chiel Warners	14.01	43.73	4.9	55.39	278.05
Attila Zsivoczky	14.95	45.62	4.7	63.45	269.54
...					

## Decathlon Data

This data (`decathlon.csv`, see also Table A.2) describes the results of the decathlon competition during the 2004 Olympic Games in Athens. The performance of all 30 athletes in the 100 m race (in seconds), long jump (in metres), shot-put (in metres), high jump (in metres), 400 m race (in seconds), 110 m hurdles race (in seconds), discus competition (in metres), pole vault (in metres), javelin competition (in metres), and 1500 m race (in seconds) are recorded in the data set.

## Theatre Data

This data (`theatre.csv`, see also Table A.3) summarizes a survey conducted on 699 participants in a big Swiss city. The survey participants are all frequent visitors to a local theatre and were asked about their age, sex (gender, female = 1), annual income (in 1000 SFR), general expenditure on cultural activities (“Culture”, in SFR per month), expenditure on theatre visits (in SFR per month), and their estimated expenditure on theatre visits in the year before the survey was done (in SFR per month).

**Table A.3** First few rows of the theatre data

	Age	Sex	Income	Culture	Theatre	Theatre_ly
1	31	1	90.5	181	104	150
2	54	0	73.0	234	116	140
3	56	1	74.3	289	276	125
4	36	1	73.6	185	75	130
5	24	1	109.0	191	172	140
6	25	0	93.1	273	168	130
...						

## Solutions to Chapter 1

### *Solution to Exercise 1.1*

- (a) The population consists of all employees of the airline. This may include administration staff, pilots, stewards, cleaning personnel, and others. Each single employee relates to an observation in the survey.
- (b) The population comprises all students who take part in the examination. Each student represents an observation.
- (c) All people suffering high blood pressure in the study area (city, province, country, . . .), are the population of interest. Each of these persons is an observation.

*Solution to Exercise 1.2* The *population* in this study refers to all leopards in the park. Only a few of the leopards are equipped with the GPS devices. This is the *sample* on which the study is conducted in. Each leopard refers to an *observation*. The measurements are taken for each leopard in the sample. The GPS coordinates allow to determine the position during the entire day. Important *variables* to capture would therefore be  $X_1$  = “latitude”,  $X_2$  = “longitude”, and  $X_3$  = “time”. Each variable would take on certain *values* for each observation; for example, the first leopard may have been observed at latitude  $32^\circ$  at a certain time point, and thus  $x_{11} = 32^\circ$ .

*Solution to Exercise 1.3*

Qualitative:	Preferred political party, eye colour, gender, blood type, subject line of an email.
Quantitative and discrete:	Shoe size, customer satisfaction on a scale from 1 to 10, number of goals in a hockey match.
Quantitative and continuous:	Time to travel to work, price of a canteen meal, wavelength of light, delivery time of a parcel, height of a child.

*Solution to Exercise 1.4*

- (a) The choice of a political party is measured on a nominal scale. The names of the parties do not have a natural order.
- (b) Typically the level of a computer game is measured on an ordinal scale: for example, level 10 may be more difficult than level 5, but this does not imply that level 10 is twice as difficult as level 5, or that the difference in difficulty between levels 2 and 3 is the same as the difference between levels 10 and 11.
- (c) The production time of a car is measured on a continuous scale (ratio scale). In practice, it may be measured in days from the start of the production.
- (d) This variable is measured on a continuous scale (ratio scale). Typically, the age is captured in years starting from the day of birth.
- (e) Calender year is a continuous variable which is measured on an interval scale. Note that the year which we define as “zero” is arbitrary, and it varies from culture to culture. Because the year zero is arbitrary, and we also have dates before this year, the calender year is measured on an interval scale.
- (f) The scale is continuous (ratio scale).
- (g) The scale of ID numbers is nominal. The ID number may indeed consist of numbers; however, “112233” does not refer to something half as much/good as “224466”. The number is descriptive.
- (h) The final rank is measured on an ordinal scale. The ranks can be clearly ordered, and the participants can be ranked by using their final results. However the first winner may not have “double” the beauty of the second winner, it is merely a ranking.
- (i) The intelligence quotient is a variable on a continuous scale. It is constructed in such a way that differences are interpretative—i.e. being 10 points above or 10 points below the average score of 100 points means the same deviation from the average. However, ratios cannot be interpreted, so the intelligence quotient is measured on an interval scale.

*Solution to Exercise 1.5*

- (a) The data is provided in .csv format. We thus read it in with the `read.csv()` command (after we have set a working directory with `setwd()`):

```
setwd('C:/directory')
pizza <- read.csv('pizza_delivery.csv')
```



- (b) The data can be viewed by means of the `fix()` or `View()` command or simply being printed:

```
fix(pizza)
pizza
```



- (c) We can access the data, as for any matrix, by using squared brackets `[, ]`, see also Appendix A.1. The first entry in the brackets refers to the row and the second entry to the columns. Each entry either is empty (referring to every row/column) or consists of a vector or sequence describing the columns/rows we want to select. This means that we obtain the first 5 rows and variables via `pizza[1:5,1:5]`. If we give the new data the name “`pizza2`” we simply need to type:

```
pizza2 <- pizza[1:5,1:5]
pizza2
```



We can save this new data either as a `.dat` file (with `write.table()`), or as a `.csv` file (with `write.csv()`), or directly as an *R* data file (with `save()`) which gives us access to our entire *R* session.

```
write.csv(pizza2,file='pizza2.csv')
write.table(pizza2,file='pizza2.dat')
save(pizza2,file='pizza2.Rdata')
```



- (d) We can access any variable by means of the `$` sign. If we type `pizza$new` we create a new variable in the `pizza` data set called “`new`”. Therefore, a simple way to add a variable to the data set is as follows:

```
pizza$NewTemperature <- 32+1.8*pizza$temperature
```



- (e)

```
attach(pizza)
NewTemperature
```



```
> str(pizza)
'data.frame': 1266 obs. of 13 variables:
 $ day           : Factor w/ 7 levels "Friday","Monday",...: 5 5 5 5 ...
 $ date          : Factor w/ 31 levels "01-May-14","02-May-14",...: 1
 $ time          : num 35.1 25.2 45.6 29.4 30 ...
 $ operator      : Factor w/ 2 levels "Laura","Melissa": 1 2 2 2 2 2 ...
 $ branch        : Factor w/ 3 levels "Centre","East",...: 2 2 3 2 3 ...
 $ driver         : Factor w/ 5 levels "Bruno","Domenico",...: 1 5 5 5 ...
 $ temperature   : num 68.3 71 53.4 70.3 71.5 ...
 $ bill           : num 58.4 26.4 58.1 35.2 38.4 61.8 57.9 35.8 36.6 ...
 $ pizzas         : int 4 2 3 3 2 4 3 2 2 5 ...
 $ free_wine     : int 0 0 1 0 0 1 1 0 0 0 ...
 $ got_wine      : int 0 0 0 0 0 1 1 0 0 0 ...
 $ discount_customer: int 1 0 0 0 0 0 0 0 0 0 ...
 $ NewTemperature: num 155 160 128 159 161 ...
```

**Fig.B.1** Applying `str()` to the pizza data

- (f) We can apply all these commands onto the object “pizza”. The command `str(pizza)` gives us an overview of the data set, see also Fig. B.1. The output shows that the data set contains 1266 observations (deliveries) and 13 variables. We can also see which of these variables are factors (categorical with defined categories) and which are numerical. We also see the first actual numbers for the each variable and the coding scheme used for the categories of the factor variables. The command `dim` summarizes the dimension of the data, i.e. the number of rows and columns of the data matrix. `Colnames` gives us the names of the variables from the data set, and so does `names`. The commands `nrow` and `ncol` give us the number of rows and columns, respectively. Applying `head` and `tail` to the data prints the first and last rows of the data, respectively.

### Solution to Exercise 1.6

- (a) The appropriate study design is a survey. The information would be obtained via a questionnaire given to a sample of parents. It is not a controlled experiment because we do not manipulate one particular variable, while controlling others; we rather collect data on all variables of interest.
- (b) There are different options to ask for parents’ attitudes: of course one could simply ask “what do you think of immunization?”; however, capturing long answers in a variable “attitude” may make it difficult to summarize and distil the information obtained. A common way to deal with such variables is to translate a concept into a score: for example, one could ask 5 “yes/no”-type questions (instead of one general question) which relate to attitudes towards immunization, such as “do you think immunization may be harmful for your child?” or “do you agree that it is a priority to immunize infants in their first year of life?” The number of answers that show a positive attitude towards immunization can be summed up. If there are 5 questions, there are up to 5 points “to earn”. Thus, each parent may be asked 5 questions and his/her attitude can be summarized on a scale ranging from 0 to 5, depending on the answers given.

(c) The following variables are needed:

- Attitude: possibly several variables are needed to capture parents' information in a score, see (b) for details. The scale is ordinal because a higher score relates to a more positive attitude towards immunization, but the differences between different score levels cannot be interpreted meaningfully.
- Immunized: a binary ("yes–no" type) variable capturing whether the parent agrees to immunization against chickenpox for their youngest child or not. This is a nominal variable.
- Gender: to compare "Immunized" for male and female parents. This is a nominal variable.
- Age: to compare the age distribution in the group of parents who would immunize their child with the age distribution in the group who would not immunize their child. Age is measured on a continuous scale.

(d) A data set might look as follows:

Parent	$A_1$	...	$A_5$	Attitude	Immunization	Gender	Age
1	yes	...	yes	3	yes	male	35
2	no	...	yes	2	no	female	26
:	:	:		:	:	:	:

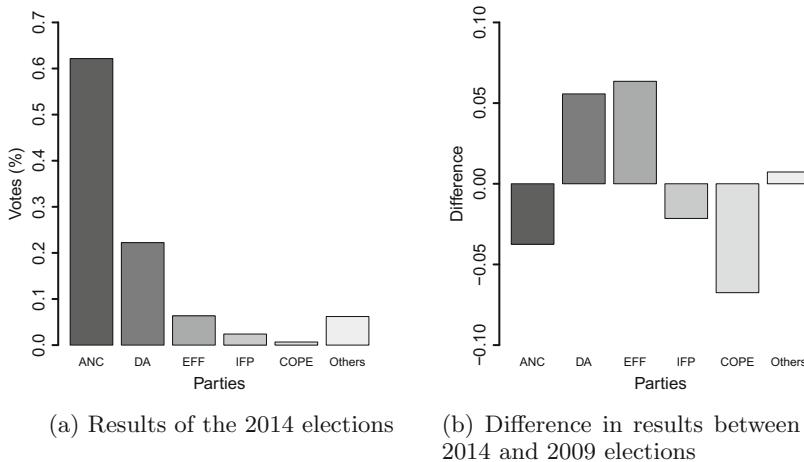
where  $A_1, \dots, A_5$  refer to variables capturing attitudes towards immunization and "Attitude" is the score variable summarizing these questions. The questions may be written down as follows:

- What is the average attitude score towards immunization among parents and how much does it vary?
- What percentage of parents answer "yes" when asked whether they would immunize their youngest child against chickenpox?
- What is the difference in the proportion calculated in (b) when stratified by gender?
- What is the average age of those parents who would immunize their child compared with the average age of those parents who would not immunize their child?

## Chapter 2

### *Solution to Exercise 2.1*

(a) The table shows the relative frequencies of each party and not the absolute frequencies. We can thus draw a bar chart where the relative frequencies of votes are plotted on the  $y$ -axis and different parties are plotted on the  $x$ -axis. In  $R$ , we can first type in the data by defining two vectors and then use the "barplot" command to draw the bar chart (Fig. B.2a). Typing "?barplot" and "?par" shows the graphical options to improve the presentation and look of the graph:

**Fig. B.2** Bar charts for national elections in South Africa

```
results2014 <- c(0.6215,0.2223,0.0635,0.0240,0.0067,0.0620)
barplot(results2014)
barplot(results2014,names.arg=c('ANC','DA','EFF','IFP','COPE',
'Others'), col=gray.colors(6),ylim=c(0,0.7),xlab='Parties',ylab =
'Votes(%)')
```



- (b) There are several options to compare the results. Of course, one can simply plot the two bar charts with each bar chart representing one election. It would be important for this solution to ensure that the  $y$ -axes are identical in both the plots. However, if we want to compare the election results in one graph then we can plot the difference between the two elections, i.e. the win/loss per party. The bar chart for the new variable “difference in proportion of votes between the two elections” is shown in Fig. B.2 and is obtained as follows:

```
results2009 <- c(0.6590,0.1666,0,0.0455,0.0742,0.0547)
difference <- results2014-results2009
barplot(difference)
```



*Remark* Another solution would be to create subcategories on the  $x$ -axis: for example, there would be categories such as “ANC 2009 results” and “ANC 2014 results”, followed by “DA 2009 results” and “DA 2014 results”, and so on.

### Solution to Exercise 2.2

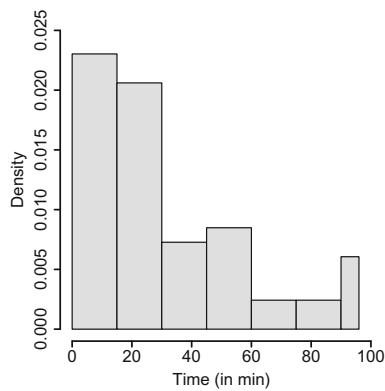
- (a) The scale of  $X$  is continuous. However, please note that the number of values  $X$  can practically take is limited (90 min plus extra time, recorded in 1 min intervals).

**Table B.1** Frequency table and other information for the variable “time until first goal”

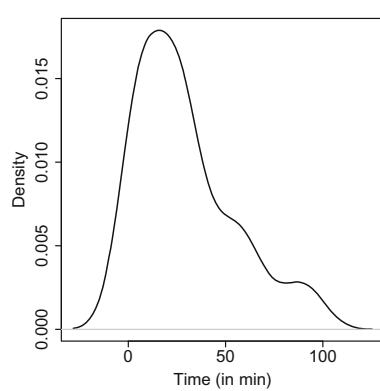
$j$	$[e_{j-1}, e_j)$	$n_j$	$f_j$	$d_j$	$h_j$	$F(x)$
1	[0, 15)	19	$\frac{19}{55}$	15	$\frac{19}{825}$	$\frac{19}{55}$
2	[15, 30)	17	$\frac{17}{55}$	15	$\frac{17}{825}$	$\frac{36}{55}$
3	[30, 45)	6	$\frac{6}{55}$	15	$\frac{6}{825}$	$\frac{42}{55}$
4	[45, 60)	5	$\frac{5}{55}$	15	$\frac{5}{825}$	$\frac{47}{55}$
5	[60, 75)	4	$\frac{4}{55}$	15	$\frac{4}{825}$	$\frac{51}{55}$
6	[75, 90)	2	$\frac{2}{55}$	15	$\frac{2}{825}$	$\frac{53}{55}$
7	[90, 96)	2	$\frac{2}{55}$	6	$\frac{2}{825}$	1
Total		56	1			

- (b) It is straightforward to obtain the frequency table, as well as the other information needed to obtain the histogram and the ECDF, see Table B.1.
- (c) We need to obtain the heights for each category to obtain the histogram using  $h_j = f_j/d_j$ , see Table B.1.
- (d) We obtain the histogram and kernel density plot in R (Fig. B.3a) using the commands

```
goals <- c(6,24,91,...,7)
hist(goals, breaks=c(0,15,30,45,60,75,90,96))
plot(density(goals, adjust=1,kernel='gaussian'))
```

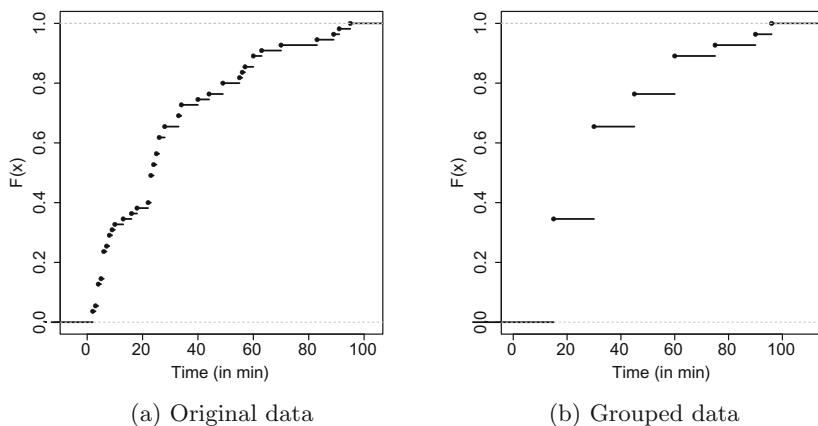


(a) Histogram



(b) Kernel density plot

**Fig. B.3** Distribution of time to goal



**Fig.B.4** Empirical cumulative distribution function for the variable “time until first goal”

- (e) The ECDF values for  $F(x)$  are calculated using the relative frequencies  $f(x)$ , see Table B.1.
  - (f) (i) We can easily plot the ECDF (Fig. B.4a) for the original data using the *R* command

```
plot.ecdf(goals)
```



- (ii) Generating the ECDF for the grouped data requires more effort and is not necessarily intuitive: first we categorize the continuous variable using the function `cut`. We use the `label` option to indicate that the name of each category corresponds to the upper limit of the respective interval. This new variable is a “factor” variable and the `plot.ecdf` function does not work with this type of variable. We need to first change the “factor” variable into a “character” variable with strings corresponding to the labels and coerce this into numeric values. Then we use `plot.ecdf`, see also Fig. B.4b. Alternatively, we can directly replace the raw values with numbers corresponding to the upper interval limits.

```
goals_cat <- cut(goals, breaks=c(0,15,30,45,60,75,90,96),
labels=c(15,30,45,60,75,90,96))
plot.ecdf(as.numeric(as.character(goals_cat)))
```



- (g) To solve the exercises, we simply use Formula (2.11) and Rules (2.3ff.)

$$(i) \ H(X \leq 45) = F(45) = \frac{42}{55} \approx 0.76.$$

$$(ii) \quad H(X > 80) = 1 - F(80) = 1 - \left( \frac{51}{55} + \frac{2/55}{15}(80 - 75) \right) \approx 0.085.$$

$$(iii) H(20 \leq X \leq 65) = F(65) - F(20) = \frac{47}{55} + \frac{4/55}{15} \cdot (65 - 60) - \left( \frac{19}{55} + \frac{17/55}{15} \cdot (20 - 15) \right) \approx 0.43.$$

(h) We know from (2.11) that

$$F(x_p) = p = F(e_{j-1}) + h_j(x_p - e_{j-1})$$

with  $h_j = f_j/d_j$  which relates to

$$x_p = e_{j-1} + \frac{p - F(e_{j-1})}{h_j}.$$

We are interested in  $x_{0.8}$  because 80 % of the matches have seen a goal at this time point:

$$x_{0.8} = 45 + \frac{0.8 - \frac{43}{56}}{\frac{1}{168}} = 50.4.$$

We conclude that 80 % of the “first goals” happened up to 50.4 min.

### *Solution to Exercise 2.3*

- (a) We obtain the relative frequencies for the first and fourth intervals as  $0.125$  ( $h_j \cdot d_j = 0.125 \cdot 1$ ). Accordingly, for the other two intervals, we obtain frequencies of  $f_j = 3 \cdot 0.125 = 0.375$ .
- (b) We obtain the absolute frequencies for the first and fourth intervals as  $250$  ( $2000 \cdot 0.125$ ). For the other intervals, we obtain  $750$  ( $2000 \cdot 0.375$ ).

### *Solution to Exercise 2.4*

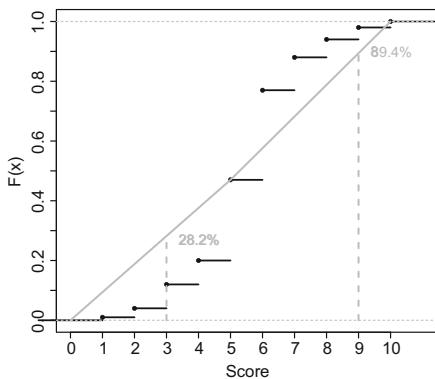
- (a) The absolute frequencies  $n_j$  are evident from the following table:

$j$	$e_{j-1}$	$e_j$	$F(e_j)$	$f_j$	$n_j (= f_j n)$	$d_j$	$a_j$
1	8	14	0.25	0.25	$0.25 \cdot 500 = 125$	6	11
2	14	22	0.40	0.15	75	8	18
3	22	34	0.75	0.35	175	12	28
4	34	50	0.97	0.22	110	16	42
5	50	82	1.00	0.03	15	32	66

- (b) We obtain  $F(X > 34) = 1 - F(34) = 1 - 0.75 = 0.25$ .

**Table B.2** Information needed to calculate the ECDF

Score	1	2	3	4	5	6	7	8	9	10
Results	1	3	8	8	27	30	11	6	4	2
$f_j$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{8}{100}$	$\frac{8}{100}$	$\frac{27}{100}$	$\frac{30}{100}$	$\frac{11}{100}$	$\frac{6}{100}$	$\frac{4}{100}$	$\frac{2}{100}$
$F_j$	$\frac{1}{100}$	$\frac{4}{100}$	$\frac{12}{100}$	$\frac{20}{100}$	$\frac{47}{100}$	$\frac{77}{100}$	$\frac{88}{100}$	$\frac{94}{100}$	$\frac{98}{100}$	1

**Fig. B.5** Empirical cumulative distribution function for the variable “score”

### Solution to Exercise 2.5

- (a) The data needed to calculate and draw the ECDF is summarized in Table B.2; the ECDF is plotted in Fig. B.5.
- (b) It follows from Table B.2 that  $F(3) = 12\%$  and  $F(9) = 98\%$ .
- (c) The grey solid line in Fig. B.5 summarizes the ECDF for the grouped data. It goes from  $(0, 0)$  to  $(1, 1)$  with a breakpoint at  $(5, 0.47)$  since  $F(5) = 0.47$  summarizes the information for the group “disagree”. Using (2.11) we calculate:

$$\begin{aligned}
 F(3) &= F(e_{j-1}) + \frac{f_j}{d_j}(x - e_{j-1}) \\
 &= F(0) + \frac{0.47}{5} \cdot (3 - 0) = 28.2\% \\
 F(9) &= F(5) + \frac{0.53}{5} \cdot (9 - 5) = 89.4\%.
 \end{aligned}$$

- (d) The results of (b) and (c) are very different. The formula applied in (c) assumes that the values in each category are uniformly distributed, i.e. that within each category, each value occurs as often as each other value. However, we know from (a) that this is not true: there are more values towards the central score numbers. The assumption used in (c) is therefore inappropriate as also highlighted in Fig. B.5.

*Solution to Exercise 2.6* We read in and attach the data as follows:

```
setwd('C:/directory')
pizza <- read.csv('pizza_delivery.csv')
attach(pizza)
```



- (a) We need the options `ylim`, `xlim`, `ylab`, `xlab`, `col` to adjust the histogram produced with `hist()`. We then add a dashed (`lty=2`) line (`type='l'`), which is thick (`lwd=3`), from (65, 0) to (65, 400) using the `lines()` command. See also Fig. B.6a.

```
hist(temperature,xlab='Temperature',xlim=c(40,90),
ylim=c(0,400),col='lightgrey',ylab='Number of deliveries')
lines(c(65,65),c(0,400),type='l',lty=2,lwd=3)
```



- (b) We can create the histogram as follows, see also Fig. B.6b:

```
library(ggplot2)
p1 <- ggplot(data=pizza,aes(x=temperature))
p2 <- p1 + geom_histogram(fill='darkgrey',alpha=0.5,binwidth=2.5)
+ scale_y_continuous('Number of deliveries')
plot(p2)
```



- (c) A possible solution is as follows, see also Fig. B.6c:

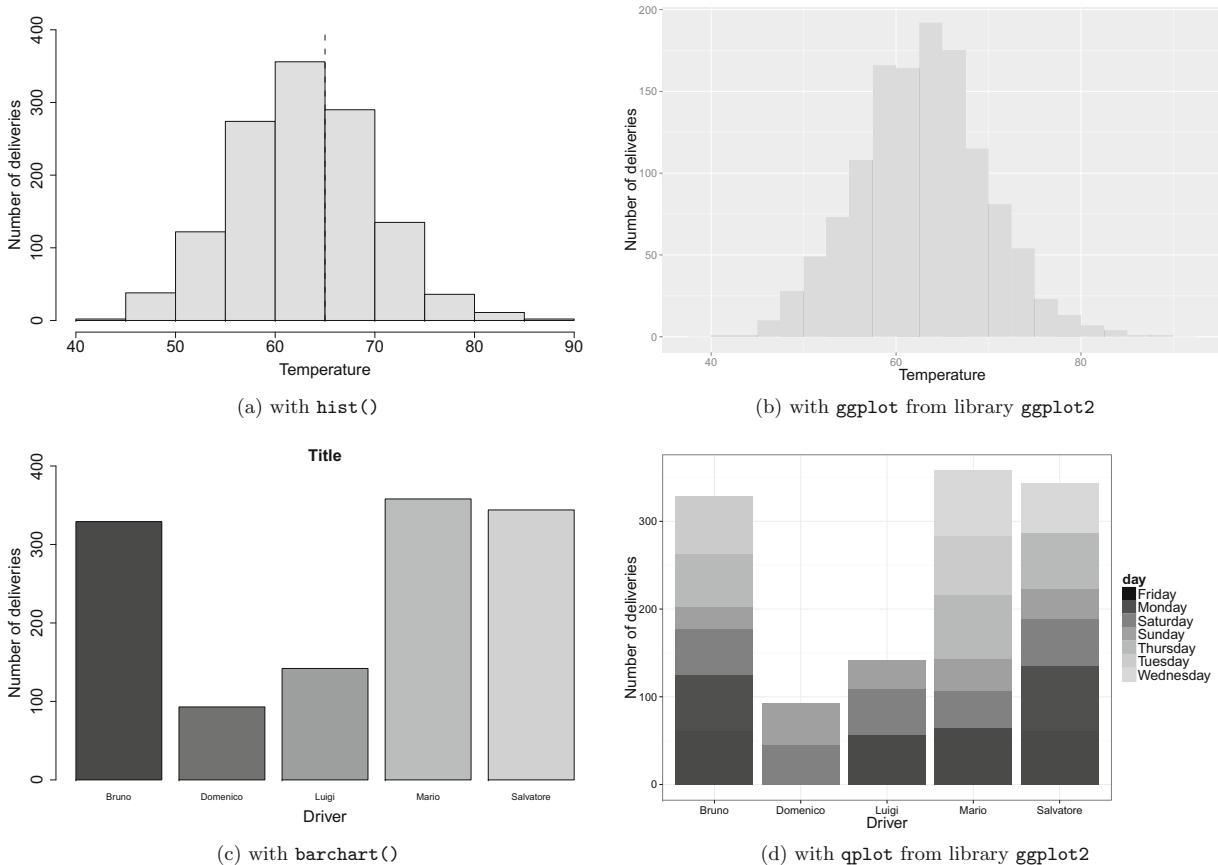
```
barplot(table(driver),ylim=c(0,200),col=gray.colors(7),
ylab='Number of deliveries', xlab='Driver',main='Title')
```



- (d) We can produce the graph (Fig. B.6d) as follows:

```
p3 <- qplot(driver,data=pizza,aes=('bar'),fill=day)
p4 <- p3 + scale_fill_grey() +theme_bw()
plot(p4)
```





**Fig.B.6** Creating figures with R

## Chapter 3

*Solution to Exercise 3.1*

- (a) The arithmetic means can be calculated as follows:

$$\bar{x}_D = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{1}{10}(12.5 + \dots + 17.5) = 17.32,$$

$$\bar{x}_A = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{1}{10}(342 + \dots + 466) = 612.4.$$

The ordered values of the two variables are:

i:	1	2	3	4	5	6	7	8	9	10
<hr/>										
D:	7.6	12.1	12.5	14.8	16.2	16.5	17.5	18.5	27.4	29.9
A:	238	342	398	466	502	555	670	796	912	1245

$$\tilde{x}_{0.5,D} = \frac{1}{2}(\tilde{x}_{(5)} + \tilde{x}_{(6)}) = \frac{1}{2}(16.2 + 16.5) = 16.35,$$

$$\tilde{x}_{0.5,A} = \frac{1}{2}(\tilde{x}_{(5)} + \tilde{x}_{(6)}) = \frac{1}{2}(502 + 555) = 528.5.$$

- (b) We have  $n\alpha = 10 \cdot 0.25 = 2.5$  which is not an integer. We can therefore calculate the 25 % quantiles, i.e. the first quartiles, as

$$\tilde{x}_{0.25,D} = \tilde{x}_{(3)} = 12.5; \quad \tilde{x}_{0.25,A} = \tilde{x}_{(3)} = 398.$$

Similarly,  $n\alpha = 10 \cdot 0.75 = 7.5$  is not an integer and thus

$$\tilde{x}_{0.75,D} = \tilde{x}_{(8)} = 18.5; \quad \tilde{x}_{0.75,A} = \tilde{x}_{(8)} = 796.$$

One can see that the distributions for both variables are not symmetric. For example, when looking at the distance hiked, the difference between the median and the first quartile ( $16.35 - 12.5$ ) is much larger than the difference between the median and the third quartile ( $18.5 - 16.35$ ); this indicates a distribution that is skewed towards the left.

- (c) We can calculate the interquartile ranges as follows:

$$d_{Q,A} = 796 - 398 = 398; \quad d_{Q,D} = 18.5 - 12.5 = 6.$$

The mean absolute deviations are:

$$D_D(\tilde{x}_{0.5}) = \frac{1}{10}(|7.6 - 16.35| + \dots + |29.9 - 16.35|) = 4.68,$$

$$D_A(\tilde{x}_{0.5}) = \frac{1}{10}(|238 - 528.5| + \dots + |1245 - 528.5|) = 223.2.$$

The variances of both the variables are

$$\tilde{s}_D^2 = \frac{1}{10}([7.6 - 16.35]^2 + \dots + [29.9 - 16.35]^2) \approx 41.5,$$

$$\tilde{s}_A^2 = \frac{1}{10}([238 - 528.5]^2 + \dots + [1245 - 528.5]^2) \approx 82,314.$$

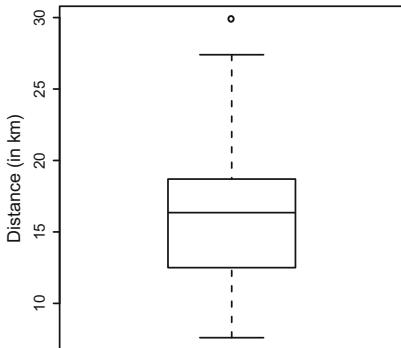
The standard deviations are therefore  $\tilde{s}_D = \sqrt{41.5}$  and  $\tilde{s}_A = \sqrt{82,314}$ .

(d) To answer this question, we can use the rules of linear transformations.

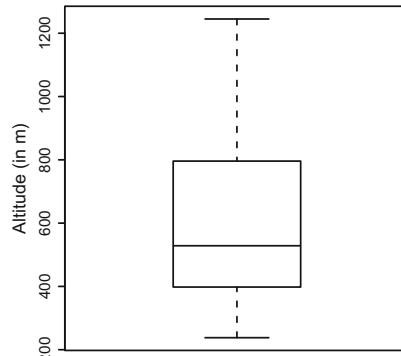
$$\bar{y} \stackrel{(3,4)}{=} 0 + 3.28\bar{x} = 3.28 \cdot 528.5 = 1722.48,$$

$$\tilde{s}_y^2 \stackrel{(3,29)}{=} b^2\tilde{s}_x^2 = 3.28^2 \cdot 272.2 \approx 2928.$$

(e) To draw the box plots, we can use the results from (a), (b), and (c) in this solution. The median, first quartile, third quartile, and the interquartile range have already been calculated. It is also easy to determine the minimum and maximum values from the table of ordered values in (a). What is missing is the knowledge of whether to treat some of the values as extreme values or not. For the distance hiked, an extreme value would be any value  $> 18.5 + 1.5 \times 6 = 27.5$  or  $< 12.5 - 1.5 \times 6 = 3.5$ . It follows that there is only one extreme value: 29.9 km. For the maximum altitude, there is no extreme value because there is no value  $> 796 + 1.5 \times 398 = 1292$  or  $< 398 - 1.5 \times 398 = -199$ . The box plots are shown in Fig. B.7a, b.



(a) Box plot for distance hiked



(b) Box plot for maximum altitude

**Fig. B.7** Box plots for Exercise 3.1

(f) The data can be summarized as follows:

Class intervals	(5; 15]	(15; 20]	(20; 30]
$n_j$	4	4	2
$f_j$	4/10	4/10	2/10
$\sum f_j$	4/10	8/10	1

We can calculate the weighted arithmetic mean by using the relative frequencies  $f_j$  and the middle of the intervals  $m_j$ :

$$\bar{x} = \sum_j f_j m_j = \frac{4}{10} \cdot 10 + \frac{4}{10} \cdot 17.5 + \frac{2}{10} \cdot 25 = 16.$$

To estimate the weighted median, we need to determine the class for which

$$\sum_{j=1}^{m-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^m f_j \geq 0.5$$

holds. This is clearly the case for the second class  $K_2 = (15; 20]$ . Thus

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m = 15 + \frac{0.5 - 0.4}{0.4} \cdot 5 = 16.25.$$

(g) If the raw data is known, then the variance for the grouped data will be identical to the variance calculated in (c). For educational purposes, we show the identity here. The variance for grouped data can be calculated as:

$$\tilde{s}^2 = \underbrace{\frac{1}{n} \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2}_{\text{between}} + \underbrace{\frac{1}{n} \sum_{j=1}^k n_j \tilde{s}_j^2}_{\text{within}}$$

Using the arithmetic mean  $\bar{x} = 17.32$  as well as the means in each class,  $\bar{x}_1 = 11.75$ ,  $\bar{x}_2 = 17.225$ , and  $\bar{x}_3 = 28.65$ , we can calculate the variance between the classes:

$$\begin{aligned} \tilde{s}_b^2 &= \frac{1}{10} ([4 \cdot (11.75 - 17.32)^2] + [4 \cdot (17.225 - 17.32)^2] \\ &\quad + [2 \cdot (28.65 - 17.32)^2]) = 38.08735. \end{aligned}$$

The variances within each class are:

$$\tilde{s}_1^2 = \frac{1}{4} [(7.6 - 11.75)^2 + \dots + (14.8 - 11.75)^2] = 6.8025,$$

$$\tilde{s}_2^2 = \frac{1}{4} [(16.2 - 17.225)^2 + \dots + (18.5 - 17.225)^2] = 0.956875,$$

$$\tilde{s}_3^2 = \frac{1}{2} [(27.4 - 28.65)^2 + (29.9 - 28.65)^2] = 1.5625.$$

We thus get

$$\tilde{s}_w^2 = \frac{1}{10}(4 \cdot 6.8025 + 4 \cdot 0.956875 + 2 \cdot 1.5625) = 3.41625.$$

The total variance is therefore  $\tilde{s}^2 = \tilde{s}_w^2 + \tilde{s}_b^2 = 3.41625 + 38.08735 \approx 41.5$ . The results will typically differ if we do not know the raw data: we have to replace the arithmetic means within each class,  $\bar{x}_j$ , with the middle of each class  $a_j$ , i.e.  $a_1 = 10, a_2 = 17.5, a_3 = 25$ :

$$\begin{aligned}\tilde{s}_b^2 &= \frac{1}{10}([4 \cdot (10 - 17.32)^2] + [4 \cdot (17.5 - 17.32)^2] \\ &\quad + [2 \cdot (25 - 17.32)^2]) = 33.2424.\end{aligned}$$

We further get

$$\begin{aligned}\tilde{s}_1^2 &= \frac{1}{4}[(7.6 - 10)^2 + \dots + (14.8 - 10)^2] = 9.865, \\ \tilde{s}_2^2 &= \frac{1}{4}[(16.2 - 17.5)^2 + \dots + (18.5 - 17.5)^2] = 0.9225, \\ \tilde{s}_3^2 &= \frac{1}{2}[(27.4 - 25)^2 + (29.9 - 25)^2] = 1.5625,\end{aligned}$$

and

$$\tilde{s}_w^2 = \frac{1}{10}(4 \cdot 9.865 + 4 \cdot 0.9225 + 2 \cdot 14.885) = 7.292.$$

The variance is  $\tilde{s}^2 = \tilde{s}_w^2 + \tilde{s}_b^2 = 7.292 + 33.2424 \approx 40.5$ . The approximation is therefore good. However, please note that the between-class variance was estimated too low, but the within-class variance was estimated too high; only the combination of the two variance components led to reasonable results in this example. It is evident that the approximation in the third class was not ideal. The middle of the interval, 25, was not a good proxy for the true mean in this class, 28.65.

- (h) It is easy to calculate the mean and the median:

```
distance <- c(12.5, 29.9, ..., 17.5)
altitude <- c(342, 1245, ..., 466)
mean(distance)
mean(altitude)
median(distance)
median(altitude)
```



We can use the `quantile` function, together with the `probs` option, to get the quantiles:

```
quantile(distance,probs=0.75)
quantile(distance,probs=0.25)
quantile(altitude,probs=0.75)
quantile(altitude,probs=0.25)
```



However, the reader will see that the results differ slightly from our results obtained in (b). As noted in Example 3.1.5, *R* offers nine different ways to obtain quantiles, each of which can be chosen by the `type` argument. The difference between these options cannot be understood easily without a background in probability theory. It may, however, be worth highlighting that we get the same results as in (b) if we choose the `type=2` option in this example. The interquartile ranges can be calculated by means of the difference of the quantiles obtained above. To determine the mean absolute deviation, we have to program our own function:

```
amd <- function(mv){1/length(mv)*sum(abs(mv-median(mv)))}
amd(distance)
amd(altitude)
```



We can calculate the variance using the `var` command. However, as noted in Example 3.2.4, on p. 52, *R* uses  $1/(n - 1)$  rather than  $1/n$  when calculating the variance. This important alternative formula for variance estimation is explained in Chap. 9, Theorem 9.2.1. To obtain the results from (c), we hence need to multiply the output from *R* by  $(n - 1)/n$ :

```
var(altitude)*9/10
var(distance)*9/10
```



The box plots can be drawn by using the `boxplot` command:

```
boxplot(altitude)
boxplot(distance)
```



The weighted mean is obtained as follows:

```
weighted.mean(c(10,17.5,25),c(4/10,4/10,2/10))
```



*Solution to Exercise 3.2*

(a) We need to solve the equation that defines the arithmetic mean:

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_i x_i \\ -90 &= \frac{1}{10}(200 + 600 - 200 - 200 - 200 - 100 - 100 - 400 + 0 + R) \\ -90 &= \frac{1}{10}(-400 + R) \\ \Rightarrow R &= -500.\end{aligned}$$

(b) The mode is  $\tilde{x}_M = -200$ . Using  $n\alpha = 2.5$  and  $n\alpha = 7.5$ , respectively, we can determine the quartiles as follows:

$$\begin{aligned}\tilde{x}_{0.25} &= x_{(3)} = -200, \\ \tilde{x}_{0.75} &= x_{(8)} = 0.\end{aligned}$$

The interquartile range is  $d_Q = 0 - (-200) = 200$ .

(c) It is not possible to use the coefficient of variation because some of the values are negative.

*Solution to Exercise 3.3* We calculate

$$n_m = n - n_w - n_c = 100 - 45 - 20 = 35.$$

Using the formula for the arithmetic mean for grouped data,

$$\bar{x} = \frac{1}{n}(n_w \bar{x}_w + n_m \bar{x}_m + n_c \bar{x}_c),$$

we further get

$$\begin{aligned}\bar{x}_m &= \frac{1}{n_m}(n \bar{x} - n_w \bar{x}_w - n_c \bar{x}_c) \\ &= \frac{1}{35}(100 \cdot 15 - 45 \cdot 16 - 20 \cdot 7.5) = 18.\end{aligned}$$

Similarly, we can use the Theorem of Variance Decomposition, Eq. (3.27), to calculate the variance for the grouped data:

$$\begin{aligned}s^2 &= s_w^2 + s_b^2 = \frac{1}{n}(n_w s_w^2 + n_m s_m^2 + n_c s_c^2) \\ &\quad + \frac{1}{n} (n_w (\bar{x}_w - \bar{x})^2 + n_m (\bar{x}_m - \bar{x})^2 + n_c (\bar{x}_c - \bar{x})^2).\end{aligned}$$

This yields

$$\begin{aligned}s_m^2 &= \frac{1}{n_m} [ns^2 - n_w s_w^2 - n_c s_c^2 - n_w (\bar{x}_w - \bar{x})^2 - n_m (\bar{x}_m - \bar{x})^2 - n_c (\bar{x}_c - \bar{x})^2] \\ &= \frac{1}{35} (100 \cdot 19.55 - 45 \cdot 6 - 20 \cdot 3 - 45 \cdot 1^2 - 35 \cdot 3^2 - 20 \cdot 7.5^2) = 4.\end{aligned}$$

*Solution to Exercise 3.4*

- (a) We evaluate a period of 6 years which means that  $T = 0, 1, 2, 3, 4, 5$ . To determine the average growth rate, we need to first calculate the geometric mean. There are two options to facilitate this:

(i) Solution:

$$B_t/B_{t-1} = (-, 1.04, 1.125, 0.925, 1.2, 0.933)$$

$$\bar{x}_G = (1.04 \cdot 1.125 \cdot 0.925 \cdot 1.2 \cdot 0.933)^{1/5} = 1.04.$$

(ii) Easier solution:

$$\bar{x}_G = (28/23)^{1/5} = 1.04.$$

Since  $\bar{x}_G = 1.04$ , the average growth rate is  $r = 1.04 - 1 = 4\%$ .

- (b) To predict the number of members in 2018, we could apply the average growth rate to the number of members in 2016 for two consecutive years:

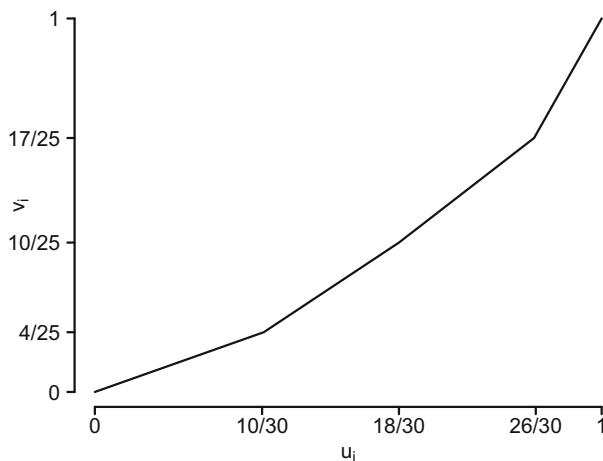
$$B_{2018} = \bar{x}_G B_{2017}, \quad B_{2017} = \bar{x}_G B_{2016}, \quad \Rightarrow B_{2018} = \bar{x}_G^2 B_{2016}$$

$$B_{2018} = 1.04^2 \cdot 28 = 30.28 \approx 31.$$

- (c) We could use the approach outlined in (b) to predict the number of members in 2025. However, this assumes that the average growth rate between 2011 and 2016 remains valid until 2025. This is rather unrealistic. The number of members of the club increases in some years, but decreases in other years. There is no guarantee that the pattern observed until 2016 can be generalized to the future. This highlights that statistical methodology should in general be used with care when making long-term future predictions.
- (d) The invested money is  $\sum_i x_i = € 250$  million. We deal with partially grouped data because the club's members invest in groups, but the invested sum is clearly defined and not part of a group or interval. We can thus summarize the data as follows:

$(i)$	1	2	3	4
Number of members	10	8	8	4
Rel. number of members $f_j$	10/30	8/30	8/30	4/30
$\tilde{u}_i = \sum_j f_j$	10/30	18/30	26/30	1
Money invested $x_i$	40	60	70	80
Rel. amount per group	40/250	60/250	70/250	80/250
$v_i$	40/250	100/250	170/250	1

The Lorenz curve is plotted in Fig. B.8.



**Fig. B.8** Lorenz curve

(e) The Gini coefficient can be calculated using formula (3.37) as follows:

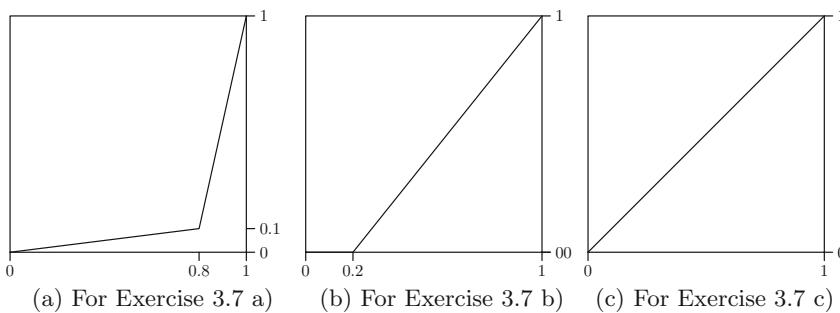
$$G = 1 - \frac{1}{30}(10(0 + 4/25) + 8(4/25 + 2/5) + 8(2/5 + 17/25) + 4(17/25 + 1)) = 1 - 268/375 = 107/375 = 0.2853.$$

$$G^+ \stackrel{(3.39)}{=} 30/29 \cdot 107/375 = 214/725 = 0.2952.$$

The concentration of investment is thus rather weak.

*Solution to Exercise 3.5* Let us look at Fig. 3.8b first. The quantiles coincide approximately with the bisection line. This means that the distribution of “length of service” is similar for men and women. They have worked for about the same amount of time for the company. For example, the median service time should be approximately the same for men and women, the first quartile should be very similar too, the third quartile should be similar too, and so on. However, Fig. 3.8a shows a somewhat different pattern: the quantiles for men are consistently higher than those for women. For example, the median salary will be higher for men. In conclusion, we see that men and women have a similar length of service in the company, but earn less.

*Solution to Exercise 3.6* There are many ways in which a “mode” function can be programmed. We present a simple and understandable solution, not the most efficient one. Recall that `table` constructs a frequency table. This table shows immediately which value(s) occur(s) most often in the data. How can we extract it? Applying the

**Fig. B.9** Lorenz curves

`names` function on the table returns the names of the values represented in the table. The only thing we need to do is to choose the name which corresponds to the value which has been counted most. This yields the following function:

```
mymode <- function(vec){
  mt <- table(vec)
  names(mt)[mt == max(mt)]
}
```



This function will work in general, though it returns a character vector. Using `as.numeric` is one option to make the character string numeric, if necessary.

### *Solution to Exercise 3.7*

- In this exercise, we do not have individual data; i.e. we do not know how much each inhabitant earns. The summarized data simply tells us about the wealth of two groups. For simplicity and for illustrative purposes, we assume that the wealth is equally distributed in each group. We determine  $(\tilde{u}_i, \tilde{v}_i)$  as  $(0.8, 0.1)$  and  $(1, 1)$  because 80 % of the population earn 10 % of the wealth and 100 % of the population earn everything. The respective Lorenz curve is illustrated in Fig. B.9a.
- The upper class lost its wealth. This means that 20 % of the population do not own anything at all. However, the remaining 80 % owns the rest. This yields  $(\tilde{u}_i, \tilde{v}_i)$  of  $(0.2, 0)$  and  $(1, 1)$ , see also Fig. B.9b.
- In this scenario, 20 % of the population leave the country. However, the remaining 80 %—which are now 100 % of the population—earn the rest. The money is equally distributed between the population. Figure B.9c shows this situation.

*Solution to Exercise 3.8*

- (a) It is necessary to use the harmonic mean to calculate the average speed. Using  $w_1 = n_1/n = 180/418 \approx 0.43$ ,  $w_2 = 117/418 \approx 0.28$ , and  $w_3 = 121/418 \approx 0.29$  we get

$$\begin{aligned}\bar{x}_H &= \frac{1}{\sum_{i=1}^k \frac{w_i}{x_i}} \\ &= \frac{1}{0.43/48 + 0.28/37 + 0.29/52} \approx 45.2 \text{ km/h.}\end{aligned}$$

- (b) Travelling at 45.2 km/h means travelling about 361 km in 8 h. The bus will not be in time.

*Solution to Exercise 3.9*

- (a) The sum of investment is  $\sum_i x_i = €18,020$ . To calculate and draw the Lorenz curve, we need the following table:

$(i)$	1	2	3	4
investment $x_i$	800	2220	4700	10300
$f_j$	1/4	1/4	1/4	1/4
$u_i$	1/4	2/4	3/4	1
relative investment	0.044	0.123	0.261	0.572
$v_i$	0.044	0.168	0.428	1

The curve is presented in Fig. B.10.

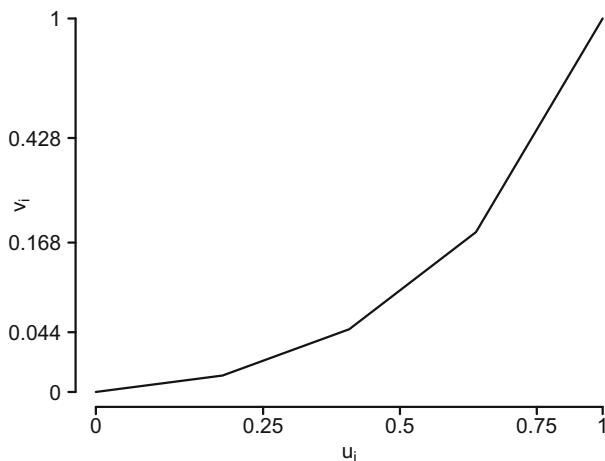
- (b) We can calculate the Gini coefficient as follows:

$$\begin{aligned}G &\stackrel{(3.37)}{=} 1 - \frac{1}{4}[(0 + 0.044) + (0.044 + 0.168) + (0.168 + 0.428) \\ &\quad + (0.428 + 1)] = 1 - \frac{1}{4} \cdot 2.28 = 0.43. \\ G^+ &\stackrel{(3.39)}{=} \frac{n}{n-1} G = \frac{4}{3} \cdot 0.43 = 0.57.\end{aligned}$$

- (c) The Gini coefficient remains the same as the relative investment stays the same.  
 (d) Using the library `ineq` we can easily reproduce the results in *R*:

```
library(ineq)
investment <- c(800, 10300, 4700, 2200)
plot(Lc(investment))
ineq(investment)
```





**Fig. B.10** Lorenz curve for investment sum

However, please note that the `ineq` command calculates the unstandardized Gini coefficient.

### Solution to Exercise 3.10

- (a) The easiest way to get all these measures is to use the `summary` function and apply it to the data columns which relate to quantitative variables:

```
setwd('C:/yourpath')
pizza <- read.csv('pizza_delivery.csv')
attach(pizza)
summary(pizza[,c('time','temperature','bill','pizzas')])
```



We then get the following output:

time	temperature	bill	pizzas
Min. :12.27	Min. :41.76	Min. : 9.10	Min. : 1.000
1st Qu.:30.06	1st Qu.:58.24	1st Qu.:35.50	1st Qu.: 2.000
Median :34.38	Median :62.93	Median :42.90	Median : 3.000
Mean :34.23	Mean :62.86	Mean :42.76	Mean : 3.013
3rd Qu.:38.58	3rd Qu.:67.23	3rd Qu.:50.50	3rd Qu.: 4.000
Max. :53.10	Max. :87.58	Max. :75.00	Max. :11.000

- (b) We can use the quantile function:

```
quantile(time,probs=0.99)
quantile(temperature,probs=0.99)
```



The results are 48.62 min for delivery time and 79.87 °C for temperature. This means 99 % of the delivery times are less than or equal to 48.62 min and 1 % of deliveries are greater than or equal to 48.62 min. Similarly, only 1 % of pizzas were delivered with a temperature greater than 79.87 °C.

- (c) The following simple function calculates the absolute mean deviation:

```
amdev <- function(mv){1/length(mv)*sum(abs(mv-mean(mv)))}
amdev(temperature)
```



- (d) We can use the `scale`, `mean`, and `var` commands, respectively.

```
sc.time <- scale(time)
mean(sc.time)
var(sc.time)
```



As one would expect, the mean is zero and the variance is 1 for the scaled variable.

- (e) The `boxplot` command draws a box plot; the `range` option specifies the range for which extreme values are defined. As specified in the help files, `range=0` produces a box plot with no extreme values.

```
boxplot(temperature,range=0)
boxplot(time,range=0)
```



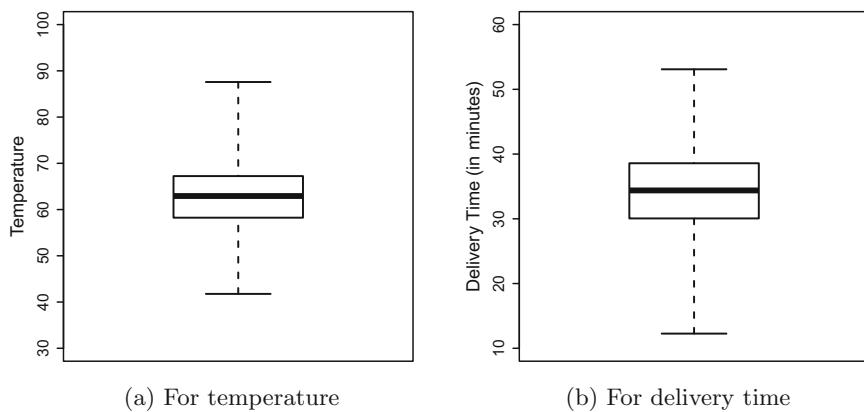
The box plots are displayed in Fig. B.11.

- (f) We use the `cut` command to create a variable which has the categories (10, 20], (20, 30], (30, 40], (40, 50], (50, 60], respectively. Using the interval mid-points, as well as the relative frequencies in each class (obtained via the `table` command), we get:

```
tc <- cut(time,breaks=seq(10,60,10))
weighted.mean(c(15,25,35,45,55),table(tc)/sum(table(tc)))
[1] 34.18641
mean(time)
[1] 34.22955
```



The weighted mean is very similar to the mean from the raw data, see output above.

**Fig. B.11** Box plots

(g) The plots can be reproduced by using the `qqplot` command:

```
qqplot(time[driver=='Luigi'],time[driver=='Domenico'])
qqplot(time[driver=='Mario'],time[driver=='Salvatore'])
```



## Chapter 4

### *Solution to Exercise 4.1*

(a) We need the following table to calculate the correlation coefficient  $R$ :

Café ( $i$ )	$x_i$	$R(x_i)$	$y_i$	$R(y_i)$	$d_i$	$d_i^2$
1	3	1	6	2	-1	1
2	8	4	7	3	1	1
3	7	3	10	5	-2	4
4	9	5	8	4	1	1
5	5	2	4	1	1	1

$$R = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6(1 + 1 + 4 + 1 + 1)}{5(25 - 1)} = 1 - 0.4 = 0.6.$$

There is a moderate-to-strong positive correlation between the ratings of the two coffee enthusiasts. In general, a high rating from one staff member implies a rather high rating from the other staff member.

(b) Above we have assigned ranks in an increasing order; i.e. the lowest  $x_i/y_i$  gets the lowest rank (1) and the highest  $x_i/y_i$  gets the highest rank (5). If we use

decreasing order and assign the lowest rank to the highest values, we get the following results:

Café ( $i$ )	$x_i$	$R(x_i)$	$y_i$	$R(y_i)$	$d_i$	$d_i^2$
1	3	5	6	4	1	1
2	8	2	7	3	-1	1
3	7	3	10	1	2	4
4	9	1	8	2	-1	1
5	5	4	4	5	-1	1

As in (a), we have  $\sum_i d_i^2 = 8$  and therefore, the results are identical:  $R = 0.6$ . Depending on whether ranks are assigned in an increasing order or a decreasing order, the sign of  $d_i$  differs, but the calculation of  $R$  is not affected since the squared values of  $d_i$  are used for its calculation and the sign of  $d_i$  is thus not important.

- (c) We can summarize the ratings in a  $2 \times 2$  table:

		X	Y
Coffee	Bad	2	1
	Good	3	4

The odds ratio is  $OR = (2 \times 4)/(3 \times 1) = 2$ . The chance of rating a coffee as good is twice as likely for person X compared to person Y.

### Solution to Exercise 4.2

- (a) The expected frequencies (under independence) are:

	Satisfied	Unsatisfied
Car	$\frac{74 \cdot 58}{150} = 28.61$	$\frac{76 \cdot 58}{150} = 29.39$
Car (diesel engine)	$\frac{74 \cdot 60}{150} = 29.6$	$\frac{76 \cdot 60}{150} = 30.4$
Motorbike	$\frac{74 \cdot 32}{150} = 15.79$	$\frac{76 \cdot 32}{150} = 16.21$

We therefore have

$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i+j}}{n}\right)^2}{\frac{n_{i+j}}{n}} \\ &= \frac{(33 - 28.61)^2}{28.61} + \frac{(25 - 29.39)^2}{29.39} + \frac{(29 - 29.6)^2}{29.6} \\ &\quad + \frac{(31 - 30.4)^2}{30.4} + \frac{(12 - 15.79)^2}{15.79} + \frac{(20 - 16.21)^2}{16.21} \\ &= 0.6736 + 0.6557 + 0.0122 + 0.0112 + 0.9097 + 0.8861 = 3.1485.\end{aligned}$$

The maximum value  $\chi^2$  can take is  $150(2 - 1) = 150$  which indicates that there is almost no association between type of vehicle and satisfaction with the insurance. The other measures can be calculated as follows:

$$V = \sqrt{\frac{\chi^2}{n(\min(k, l) - 1)}} = \sqrt{\frac{3.1485}{150(2 - 1)}} = 0.14.$$

$C_{\text{corr}}$ :

$$\begin{aligned}C_{\text{corr}} &= \sqrt{\frac{\min(k, l)}{\min(k, l) - 1}} \sqrt{\frac{\chi^2}{\chi^2 + n}} = \\ &= \sqrt{\frac{2}{1}} \sqrt{\frac{3.1485}{3.1485 + 150}} = \sqrt{2} \sqrt{0.02056} \approx 0.20.\end{aligned}$$

The small values of  $V$  and  $C_{\text{corr}}$  confirm that the association is rather weak.

- (b) The summarized table looks as follows:

	Satisfied	Unsatisfied
Car	62	56
Motorbike	12	20

Using (4.7), we obtain

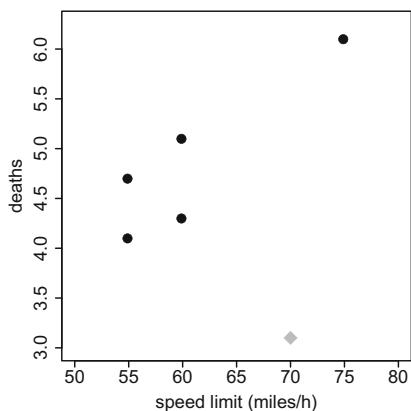
$$\begin{aligned}\chi^2 &= \frac{n(ad - bc)^2}{(a + d)(c + d)(a + c)(b + d)} \\ &= \frac{150(1240 - 672)^2}{118 \cdot 32 \cdot 74 \cdot 76} = \frac{48,393,600}{21,236,224} \approx 2.2788.\end{aligned}$$

The maximum value  $\chi^2$  can take is  $150(2 - 1)$ . The association is therefore weak. The odds ratio is

$$OR = \frac{ad}{bc} = \frac{62 \cdot 20}{12 \cdot 56} = \frac{1240}{672} \approx 1.845.$$

The chances of being satisfied with the insurance are 1.845 times higher among those who drive a car.

**Fig. B.12** Scatter diagram for speed limit and number of deaths



- (c) All  $\chi^2$ -based statistics suggest that there is only a small association between the two variables, for both the  $2 \times 3$  and the  $2 \times 2$  tables. However, the odds ratio gives us a more nuanced interpretation, showing that customers driving a car are somewhat more satisfied with their insurance. The question is whether the additional information from the odds ratio is stable and trustworthy. Confidence intervals for the odds ratio can provide guidance under such circumstances, see Sect. 9.4.4 for more details.

### Solution to Exercise 4.3

- (a) The scatter plot is given in Fig. B.12. The black circles show the five observations. A positive relationship can be discovered: the higher the speed limit, the higher the number of deaths. However, “Italy” (the observation on the top right) is the observation which gives the graph a visible pattern and drives our impression about the potential relationship.
- (b) Using  $\bar{x} = 61$  and  $\bar{y} = 4.86$  we obtain

$$S_{xx} = (55 - 61)^2 + (55 - 61)^2 + \dots + (75 - 61)^2 = 270$$

$$S_{yy} = (4.1 - 4.86)^2 + (4.7 - 4.86)^2 + \dots + (6.1 - 4.86)^2 = 2.512$$

$$S_{xy} = (55 - 61)(4.1 - 4.86) + \dots + (75 - 61)(6.1 - 4.86) = 23.2$$

and therefore

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \frac{23.2}{\sqrt{270 \cdot 2.512}} = 0.891.$$

The correlation coefficient of Spearman is calculated using the following table:

Country ( $i$ )	$x_i$	$R(x_i)$	$y_i$	$R(y_i)$	$d_i$	$d_i^2$
Denmark	55	4.5	4.1	5	-0.5	0.25
Japan	55	4.5	4.7	3	1.5	2.25
Canada	60	2.5	4.3	4	-1.5	2.25
Netherlands	60	2.5	5.1	2	0.5	0.25
Italy	75	1	6.1	1	0	0

This yields

$$R = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 5}{5(25 - 1)} = 0.75.$$

Please note that above we averaged the ranks for ties. The  $R$  function `cor` uses a more complicated approach; this is why the results differ slightly when using  $R$ .

- (c) The results stay the same. Pearson's correlation coefficient is invariant with respect to linear transformations which means that it does not matter whether we use miles/h or km/h.
- (d) (i) The grey square in Fig. B.12 represents the additional observation. The overall pattern of the scatter plot changes with this new observation pair: the positive relationship between speed limit and number of traffic deaths is not so clear anymore. This emphasizes how individual observations may affect our impression of a scatter plot.
- (ii) Based on the same approach as in (b), we can calculate  $\bar{x} = 62.5$ ,  $\bar{y} = 4.6333$ ,  $S_{xx} = 337.5$ ,  $S_{yy} = 4.0533$ ,  $S_{xy} = 13$ , and  $r = 0.3515$ . A single observation changes our conclusions from a strong positive relationship to a moderate-to-weak relationship. It is evident that Pearson's correlation coefficient is volatile and may be affected heavily by outliers or extreme observations.

#### *Solution to Exercise 4.4*

- (a) The contingency table for the absolute frequencies is as follows:

	1. Class	2. Class	3. Class	Staff	Total
Rescued	202	125	180	211	718
Not rescued	135	160	541	674	1510
Total	337	285	721	885	2228

- (b) To obtain the conditional relative frequency distributions, we calculate the proportion of passengers rescued ( $X$ ) for each travel class ( $Y$ ). In the notation of Definition 4.1.1, we determine  $f_{i|j}^{X|Y} = f_{ij}/f_{+j} = n_{ij}/n_{+j}$  for all  $i$  and  $j$ . For example,  $f_{\text{rescued}|1. \text{ class}} = 202/337 = 0.5994$ . This yields

	1. Class	2. Class	3. Class	Staff
Rescued	0.5994	0.4386	0.2497	0.2384
Not rescued	0.4006	0.5614	0.7503	0.7616

It is evident that the proportion of passengers being rescued differs by travel class. It can be speculated that there is an association between the two variables pointing towards better chances of being rescued among passengers from higher travel classes.

- (c) Using (4.3), we get

	1. Class	2. Class	3. Class	Staff	Total
Rescued	108.6	91.8	232.4	285.2	718
Not rescued	228.4	193.2	488.6	599.8	1510
Total	337	285	721	885	2228

which can be used to calculate  $\chi^2$  and  $V$  as follows:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}} = \frac{(202 - 108.6)^2}{108.6} + \frac{(125 - 91.8)^2}{91.8} \\ &\quad + \frac{(180 - 232.4)^2}{232.4} + \frac{(211 - 285.2)^2}{285.2} + \frac{(135 - 228.4)^2}{228.4} \\ &\quad + \frac{(160 - 193.2)^2}{193.2} + \frac{(541 - 488.6)^2}{488.6} + \frac{(674 - 599.8)^2}{599.8} \\ &= 80.33 + 12.01 + 11.82 + 19.30 + 38.19 + 5.71 + 5.62 + 9.18 \\ &= 182.16.\end{aligned}$$

$$V = \sqrt{\frac{\chi^2}{n(\min(k, l) - 1)}} = \sqrt{\frac{182.16}{2228(2 - 1)}} = 0.286.$$

The value of  $V$  indicates a moderate or weak relationship between the two variables. This is in contradiction to the hypothesis derived from the conditional distributions in (b).

- (d) The table is as follows:

	1. Class/2. Class	3. Class/Staff	Total
Rescued	327	391	718
Not rescued	295	1215	1510
Total	622	1606	2228

Using (4.7) we get

$$\chi^2 = \frac{2228(327 \cdot 1215 - 295 \cdot 391)^2}{718 \cdot 1510 \cdot 622 \cdot 1606} = 163.55.$$

and therefore  $V = \sqrt{\frac{163.55}{2228}} = 0.271$ .

There are several relative risks that can be calculated, for example:

$$\begin{aligned}\frac{f_{1|1}}{f_{1|2}} &= \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{327/622}{391/1606} \approx 2.16, \\ \frac{f_{2|1}}{f_{2|2}} &= \frac{n_{21}/n_{+1}}{n_{22}/n_{+2}} = \frac{295/622}{1215/1606} \approx 0.63.\end{aligned}$$

The proportion of passengers who were rescued was 2.16 times higher in the 1./2. class compared to the 3. class and staff. Similarly, the proportion of passengers who were not rescued was 0.62 times lower in the 1./2. class compared to the 3. class and staff. The odds ratio is  $OR = \frac{a \cdot d}{b \cdot c} = \frac{397,305}{115,345} = 3.444$ . This is nothing but the ratio of the relative risks, i.e.  $2.16/0.63$ . The chance of being rescued (i.e. the ratio rescued/not rescued) was almost 3.5 times higher for the 1./2. class compared to the 3. class and staff.

- (e) While Cramer's  $V$  led to rather conservative conclusions regarding a possible relationship of travel class and rescue status, the conditional relative frequency distribution, the relative risks, and the odds ratio support the hypothesis that, at least to some degree, the chances of being rescued were higher in better travel classes. This makes sense because better travel classes were located towards the top of the ship with best access to the lifeboats while both third-class passengers and large numbers of the staff were located and working in the bottom of the ship, where the water came in first.

### *Solution to Exercise 4.5*

- (a) Using (4.17) and (4.18), we get

$$\begin{aligned}r &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum_{i=1}^{36} x_i y_i - 36 \bar{x} \bar{y}}{\sqrt{n \tilde{s}_x^2 n \tilde{s}_y^2}} = \frac{22776 - 36 \cdot 12.22 \cdot 51.28}{n \sqrt{\tilde{s}_x^2 \tilde{s}_y^2}} \\ &= \frac{216.9}{36 \sqrt{76.95 \cdot 706.98}} \approx 0.026.\end{aligned}$$

This indicates that there is no linear relationship between temperature and hotel occupancy.

- (b) The scatter plot shows no clear pattern. This explains why the correlation coefficient is close to 0. However, if we look only at the points for each city separately, we see different structures for different cities: a possible negative relationship for Davos (D), a rather positive relationship for Polenca (P) and no visible relationship for Basel (B). This makes sense because for winter holiday destinations hotel occupancy should be higher when the temperature is low and for summer holiday destinations occupancy should be high in the summer months.
- (c) We type in the data by specifying three variables: temperature ( $X$ ), occupancy ( $Y$ ) and city ( $Z$ ). We then simply use the `cor` command to calculate the correlation—and condition on the values of  $Z$  which we are interested in:

```
X <- c(-6,-5,2,...,9,4)
Y <- c(91,89,76,...,9,12)
Z <- c(rep('Davos',12),rep('Polenca',12),rep('Basel',12))
cor(X[Z=='Davos'],Y[Z=='Davos'])
cor(X[Z=='Basel'],Y[Z=='Basel'])
cor(X[Z=='Polenca'],Y[Z=='Polenca'])
```



This yields correlation coefficients of  $-0.87$  for Davos,  $0.42$  for Basel and  $0.82$  for Polenca. It is obvious that looking at  $X$  and  $Y$  only indicates no correlation, but the information from  $Z$  shows strong linear relationships in subgroups. This example shows the limitations of using correlation coefficients and the necessity to think in a multivariate way. Regression models offer solutions. We refer the reader to Chap. 11, in particular Sect. 11.7.3 for more details.

### *Solution to Exercise 4.6*

- (a) We use the visual rule of thumb and work from the top left to the bottom right for the concordant pairs and from the top right to the bottom left for the discordant pairs:

$$K = 82 \cdot 43 + 82 \cdot 9 + 82 \cdot 2 + 82 \cdot 10 + 8 \cdot 2 + 8 \cdot 10 + 4 \cdot 10 + 4 \cdot 9$$

$$+ 43 \cdot 10 = 5850$$

$$D = 4 \cdot 8 + 9 \cdot 2 = 50$$

$$\gamma = \frac{K - D}{K + D} = \frac{5800}{5900} = 0.98.$$

- (b)

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i+j}}{n}\right)^2}{\frac{n_{i+j}}{n}} = \frac{(82 - \frac{86.90}{158})^2}{\frac{86.90}{158}} + \frac{(4 - \frac{49.86}{158})^2}{\frac{49.86}{158}}$$

$$\begin{aligned}
& + \frac{(0 - \frac{19.86}{158})^2}{\frac{19.86}{158}} + \frac{(8 - \frac{90.60}{158})^2}{\frac{90.60}{158}} + \frac{(43 - \frac{49.60}{158})^2}{\frac{49.60}{158}} + \frac{(9 - \frac{60.19}{158})^2}{\frac{60.19}{158}} \\
& + \frac{(0 - \frac{12.90}{158})^2}{\frac{12.90}{158}} + \frac{(2 - \frac{12.49}{158})^2}{\frac{12.49}{158}} + \frac{(10 - \frac{12.19}{158})^2}{\frac{12.19}{158}} \\
= & 22.25 \cdot 19.27 + 10.34 + 20.05 + 31.98 + 0.47 + 6.84 + 0.80 + 50.74 \\
= & 162.74.
\end{aligned}$$

$$V = \sqrt{\frac{\chi^2}{n(\min(k, l) - 1)}} = \sqrt{\frac{162.74}{158 \cdot 2}} \approx 0.72.$$

(c) The table is as follows:

		Use a leash		Total
		Agree or no.	Disagree	
Use for concerts	Agree or no.	137	9	146
	Disagree	2	10	12
	Total	139	19	158

(d) The relative risk can either be summarized as:

$$\frac{2/139}{10/19} \approx 0.03 \quad \text{or} \quad \frac{10/19}{2/139} \approx 36.6.$$

The proportion of those who disagree with using the park for summer concerts is 0.03 times lower in the group who agree or have no opinion about using leashes for dogs compared to those who disagree. Similarly, the proportion of those who disagree with using the park for summer concerts is 36.6 times higher in the group who also disagree with using leashes for dogs compared to those who do not disagree.

(e) The odds ratio is  $\text{OR} = (137 \cdot 10) / (2 \cdot 9) \approx 36.1$ .

- The chance of not disagreeing with the concert proposal is 36.1 times higher for those who also do not disagree with the leash proposal.
- The chance of not disagreeing with the leash proposal is 36.1 times higher for those who also do not disagree with the concert proposal.
- In simpler words: The chance of agreeing or having no opinion for one of the questions is 36.1 times higher if the person also has no opinion or agrees with the other question.

(f)

$$\gamma = \frac{137 \cdot 10 - 9 \cdot 2}{137 \cdot 10 + 9 \cdot 2} = \frac{1352}{1388} = 0.974.$$

- (g) In general, it makes sense to use all the information available, i.e. to use the ordinal structure of the data and all three categories. While it is clear that  $\gamma$  is superior to  $V$  in our example, one may argue that the relative risks or the odds ratio could be more useful because they provide an intuitive quantification on how the two variables relate to each other rather than just giving a summary of strength and direction of association. However, as we have seen earlier, the interpretations of the relative risks and the odds ratio are quite clumsy in this example. It can be difficult to follow the somewhat complicated interpretation. A simple summary would be to say that agreement with both questions was strongly associated ( $\gamma = 0.98$ ).

*Solution to Exercise 4.7* We use the definition of the correlation coefficient, replace  $y_i$  with  $a + bx_i$  and replace  $\bar{y}$  with  $a + b\bar{x}$  to obtain

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(a + bx_i - (a + b\bar{x}))}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (a + bx_i - (a + b\bar{x}))^2}}.$$

This equates to:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(b(x_i - \bar{x}))}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (b(x_i - \bar{x}))^2}} = \frac{b \sum_{i=1}^n (x_i - \bar{x})^2}{\sqrt{b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (x_i - \bar{x})^2}} = 1.$$

*Solution to Exercise 4.8*

- (a) We read in the data, make sure the first column is recognized as containing the row names, attach the data, and obtain the correlation using the `cor` command:

```
decathlon <- read.csv('decathlon.csv', row.names=1)
attach(decathlon)
cor(X.Discus,X.High.jump)
```



The correlation is 0.52984. This means there is a moderate-to-strong positive correlation; i.e. the longer the distance the discus is thrown, the higher the height in the high jump competition.

- (b) There are 10 variables. For the first variable, we can calculate the correlation with 9 other variables. For the second variable, we can also calculate the correlation with 9 other variables. However, we have already calculated one out of the 9 correlations, i.e. when analysing variable number one. So it is sufficient to calculate 8 correlations for the second variable. Similarly, we need another 7 correlations for the third variable, 6 correlations for the fourth variable, and so on. In total, we therefore need to have  $9 + 8 + 7 + \dots + 1 = 45$  correlations. Since the correlation coefficient describes the relationship between two variables, it

**Fig. B.13** Correlation matrix  
for the decathlon data

	X.100m	X.Long.jump	X.Shot.put	X.High.jump	X.400m	X.110m.hurdle	X.Discus	X.Pole.vault	X.Javeline	X.1500m
X.100m	1	-0.7	-0.37	-0.31	0.63	0.54	-0.23	-0.26	-0.01	0.06
X.Long.jump	-0.7	1	0.2	0.35	-0.67	-0.54	0.25	0.29	0.09	-0.15
X.Shot.put	-0.37	0.2	1	0.61	-0.2	-0.25	0.67	0.02	0.38	0.13
X.High.jump	-0.31	0.35	0.61	1	-0.17	-0.33	0.52	-0.04	0.2	0
X.400m	0.63	-0.67	-0.2	-0.17	1	0.52	-0.14	-0.12	-0.05	0.55
X.110m.hurdle	0.54	-0.54	-0.25	-0.33	0.52	1	-0.22	-0.15	-0.08	0.18
X.Discus	-0.23	0.25	0.67	0.52	-0.14	-0.22	1	-0.18	0.25	0.22
X.Pole.vault	-0.26	0.29	0.02	-0.04	-0.12	-0.15	-0.18	1	-0.07	0.18
X.Javeline	-0.01	0.09	0.38	0.2	-0.05	-0.08	0.25	-0.07	1	-0.25
X.1500m	0.06	-0.15	0.13	0	0.55	0.18	0.22	0.18	-0.25	1

makes sense to summarize the results in a contingency table, similar to a matrix, see Fig. B.13.

- (c) Using `cor(decathlon)` yields the correlation coefficients between all variable pairs. This is much simpler than calculating the correlation coefficient for each of the 45 comparisons. Note that the correlation matrix provides the 45 comparisons both in the upper triangle and in the lower triangle of the table. We know that  $r(X, Y) = r(Y, X)$ , but  $R$  still provides us with both, although they are identical. Note that the diagonal elements are 1 because  $r(X, X) = 1$ .
- (d) One way to omit rows with missing data automatically is to use the `na.omit` command:

```
cor(na.omit(decathlon))
```



The results are displayed in Fig. B.13. We see moderate-to-strong correlations between the 100 m race, 400 m race, 110 m hurdle race and long jump. This may reflect the speed-and-athletic component of the decathlon contest. We also see moderate-to-strong correlations between the shot-put, high jump, and discuss events. This may reflect the strength-and-technique component of the contest. The other disciplines also show correlations which make sense, but they are rather weak and may reflect the uniqueness of these disciplines.

### *Solution to Exercise 4.9*

- (a) A possible code is listed below:

```
 pizza <- read.csv('pizza_delivery.csv')
 pizza$tempcat <- cut(pizza$temperature, breaks=c(0,65,100))
 pizza$timecat <- cut(pizza$time, breaks=c(0,30,100))
 attach(pizza)
 addmargins(table(tempcat,timecat))
```



		timecat	
tempcat		(0,30]	(30,100]
		Sum	
(0,65]	101	691	792
(65,100]	213	261	474
Sum	314	952	1266

We can see that there is a higher proportion of high temperature ((65, 100]) in the category of short delivery times ((0, 30]) compared to long delivery times ((30, 100]).

- (b) Using the data from (a), we can calculate the odds ratio:

```
(101*261)/(213*691)
```

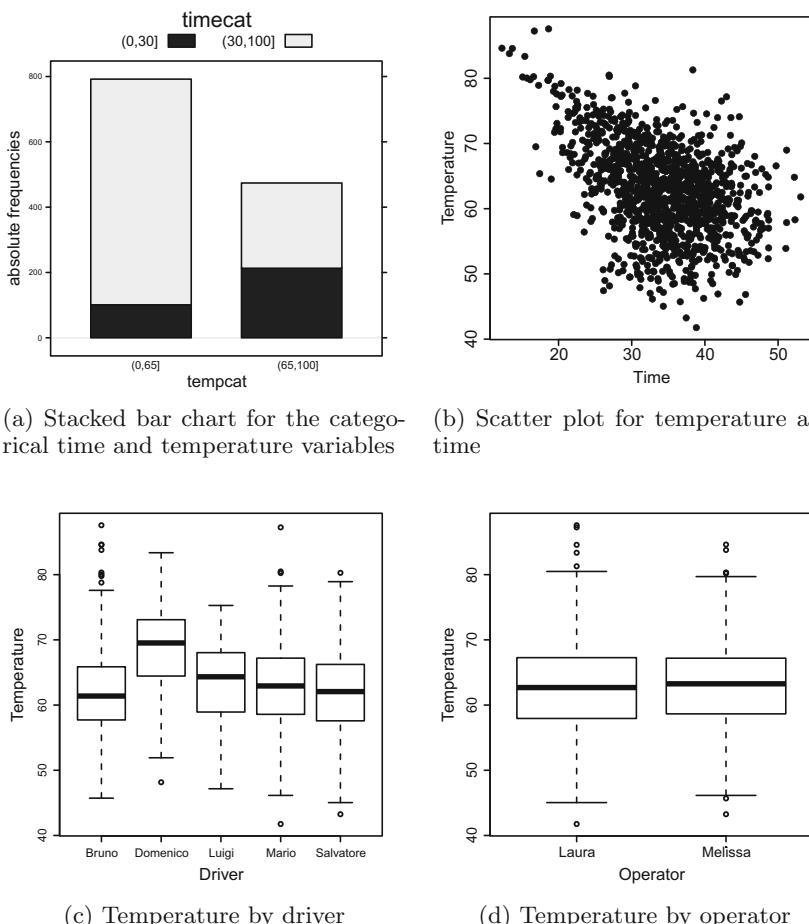


Thus, the chances of receiving a cold pizza are 0.18 lower if the delivery time is short.

- (c) We use the vcd and ryouready packages to determine the desired measures of association:

```
library(vcd)
library(ryouready)
library(lattice)
assocstats(xtabs(~tempcat+timecat))
ord.gamma(table(tempcat,timecat))
ord.tau(table(tempcat,timecat))
barchart(table(tempcat,timecat),horizontal=F,stack=T)
```



**Fig. B.14** Plots for Exercise 4.9

Cramer's  $V$  is 0.361, Stuart's  $\tau_c$  is  $-0.302$ , and  $\gamma$  is  $-0.696$ . The first two measures indicate a moderate-to-weak relationship between temperature and delivery time. It is clear from the last two measures that this relationship is negative, i.e. that shorter delivery times imply higher temperatures. However, interestingly,  $\tau_c$  and  $\gamma$  provide us with a different strengths of association. In any case, it is clear that for shorter delivery times the customers receive warmer pizzas, as evident from the stacked bar chart (Fig. B.14a).

- (d) The scatter plot (Fig. B.14b) shows a decreasing temperature for an increasing delivery time. This is also highlighted in the correlation coefficients which are  $-0.43$  and  $-0.39$  for Bravais–Pearson and Spearman, respectively.

```
plot(time,temperature)
cor(time,temperature)
cor(time,temperature,method='spearman')
```



- (e) It makes sense to compare continuous variables (temperature, number of pizzas, bill) using the correlation coefficient from Bravais–Pearson. The relationships between temperature and driver and operator could be visualized using stratified box plots.

```
boxplot(temperature~driver)
boxplot(temperature~operator)
cor(temperature,pizzas)
cor(temperature,bill)
```



The correlation coefficients are  $-0.37$  and  $-0.42$  for number of pizzas and bill, respectively. More pizzas and a higher bill are associated with a lower temperature. The box plots (Fig. B.14c, d) show variation in the delivery times of the different drivers, but almost identical delivery times for the two operators. These results give us a first idea about the relationship in the data. However, they do not tell us the full story. For example: is the pizza temperature for higher bills low because a higher bill means more pizzas, and therefore, a less efficient preparation of the food? Is one driver faster because he mainly delivers for a particular branch? Could it be that the operators have a different performance but because they deal with different branches and drivers, these differences are not visible? To address these questions, a multivariate perspective is needed. Regression models, as introduced in Chap. 11, provide some answers.

---

## Chapter 5

*Solution to Exercise 5.1* There are  $n = 10$  guests and  $m = 2$  guests shake hands with each other. The order is not important: two guests shake hands no matter who is “drawn first”. It is also not possible to shake hands with oneself which means that in terms of the urn model, we have the “no replacement” case. Therefore, we obtain the solution as

$$\binom{n}{m} = \binom{10}{2} = \frac{10 \cdot 9}{2 \cdot 1} = 45$$

handshakes in total.

*Solution to Exercise 5.2* We assume that it is not relevant to know which student gets tested at which time point. We are thus dealing with combinations without considering the order. We have  $n = 25$  and  $m = 5$  and therefore obtain:

- (a) a total number of  $\binom{25}{5} = 53,130$  possibilities.
- (b) a total number of  $\binom{25+5-1}{5} = \binom{29}{5} = 118,755$  possibilities.

In  $R$ , the commands `choose(25,5)` and `choose(29,5)`, respectively provide the required results.

*Solution to Exercise 5.3* The board consists of  $n = 381$  knots. Each knot is either occupied or not. We may thus assume a “drawing” without replacement in the sense that each knot can be drawn (occupied) only once. If we place  $m = 64$  counters on the board, then we can simultaneously think of “drawing” 64 occupied knots out of a total of 381 knots. The order in which we draw is not relevant—either a knot is occupied or not. The total number of combinations is  $\binom{n}{m} = \binom{381}{64} \approx 4.35 \cdot 10^{73}$ . We obtain the final number in  $R$  using the command `choose(381,64)`.

*Solution to Exercise 5.4* We obtain the results (again using the command `choose(n,m)` in  $R$ ) as follows:

- (a) The customer takes the beers “with replacement” because the customer can choose among any type of beer for each position in the tray. One can also think of an urn model with 6 balls relating to the six different beers, where beers are drawn with replacement and placed on the tray. The order in which the beers are placed on the tray is not relevant. We thus have

$$\binom{n+m-1}{m} = \binom{6+20-1}{20} = \binom{25}{20} = 53,130$$

combinations.

- (b) If the customer insists on having at least one beer per brewery on his tray, then 6 out of the 20 positions of the tray are already occupied. Based on the same thoughts as in (a), we calculate the total number of combinations as

$$\binom{n+m-1}{m} = \binom{6+14-1}{14} = \binom{19}{14} = 11,628.$$

*Solution to Exercise 5.5* Since each team has exactly one final place, we have a “without replacement” situation. Using  $n = 32$  and  $m = 3$  (and `choose(n,m)` in  $R$ ) yields

- (a)  $\frac{32!}{(32-3)!} = \binom{32}{3} 3! = 29,760$  and  
 (b)  $\binom{32}{3} = 4960$ .

*Solution to Exercise 5.6* There are  $n = 12$  different letters for  $m = 4$  positions of the membership code. Each letter can be used more than once if desired and we thus obtain  $n^m = 12^4 = 20,736$  possible combinations. We therefore conclude that sufficient membership codes are left. However, this may not last long and the book store may still wish to create another solution for its membership codes.

*Solution to Exercise 5.7* For each member of the jury, there are 61 scoring options:

0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
.	.	.	.	.	.	.	.	.	.
5	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9
6									

Different jury members are allowed to assign the same scores. We thus deal with combinations “with replacement”. To verify this, just think of an urn with 61 balls where each ball refers to one possible score. Now one ball is drawn, assigned to a specific jury member and then put back into the urn. Since each score is “attached” to a particular jury member, we have combinations with consideration of the order and therefore obtain a total of  $n^m = 61^9 \approx 1.17 \cdot 10^{16}$  possibilities. If you have difficulties in understanding the role of “replacement” and “order” in this example, recall that each member has 61 scoring options: thus,  $61 \times 61 \times \dots \times 61$  (9 times) combinations are possible.

*Solution to Exercise 5.8*

- (a) We obtain:

$$\begin{aligned} \binom{2}{0} = 1 &\leftrightarrow \binom{n}{2} = \binom{2}{2} = 1; & \binom{3}{1} = 3 &\leftrightarrow \binom{n}{2} = \binom{3}{2} = 3; \\ \binom{4}{2} = 6 &\leftrightarrow \binom{n}{2} = \binom{4}{2} = 6; & \binom{5}{3} = 10 &\leftrightarrow \binom{n}{2} = \binom{5}{2} = 10. \end{aligned}$$

- (b) Based on the observations from (a) we conclude that each entry on the diagonal line can be represented by  $\binom{n}{2}$ . The sum of two consecutive entries is thus  $\binom{n}{2} +$

$\binom{n+1}{2}$ . Using the fourth relation in (5.5), it follows that:

$$\begin{aligned}\binom{n}{2} + \binom{n+1}{2} &\stackrel{(5.5)}{=} \frac{n(n-1)}{2} + \frac{(n+1)n}{2} \\ &= \frac{n(n-1+n+1)}{2} = \frac{n \cdot 2n}{2} = n^2.\end{aligned}$$

## Chapter 6

*Solution to Exercise 6.1*

(a) We obtain

- $A \cap B = \{8\}$ .
- $B \cap C = \{\emptyset\}$ .
- $A \cap C = \{0, 4, 8, 9, 15\}$ .
- $C \setminus A = \{4, 9, 15\}$ .
- $\Omega \setminus (B \cup A \cup C) = \{6, 7, 11, 13, 14\}$ .

(b) The solutions are:

- $P(\bar{F}) = 1 - P(F) = 0.5$ .
- $P(G) = 1 - P(E) - P(F) = 0.3$ .
- $P(E \cap G) = 0$  because the events are pairwise disjoint.
- $P(E \setminus E) = 0$ .
- $P(E \cup F) = P(E) + P(F) = 0.7$ .

*Solution to Exercise 6.2* We know that the probability of failing the practical examination is  $P(PE) = 0.25$ , of failing the theoretical examination is  $P(TE) = 0.15$ , and of failing both is  $P(PE \cap TE) = 0.1$ .

- (a) If we ask for the probability of failing in *at least* one examination, we imply that either the theoretical examination, or the practical examination, or both are not passed. We can therefore calculate the union of events  $P(PE \cup TE) = P(PE) + P(TE) - P(TE \cap PE) = 0.25 + 0.15 - 0.1 = 0.3$ .
- (b)  $P(PE \setminus TE) = P(PE) - P(PE \cap TE) = 0.25 - 0.1 = 0.15$ .
- (c)  $P(\overline{PE \cup TE}) = 1 - P(PE \cup TE) = 1 - 0.3 = 0.7$ .

- (d) We are interested in the probability of the person failing exactly in one exam. This corresponds to  $P(M \setminus C \cup C \setminus M) = P(M \cup C) - P(C \cap M) = 0.3 - 0.1 = 0.2$ .

*Solution to Exercise 6.3* The total number of possible simple events is  $|\Omega| = 12$ . The number of favourable simple events is

- (a)  $|A| = 6$  (i.e. the numbers 2, 4, 6, 8, 10, 12). Hence,  $P(A) = \frac{6}{12} = \frac{1}{2}$ .
- (b)  $|B| = 3$  (i.e. the numbers 10, 11, 12). Hence,  $P(B) = \frac{3}{12} = \frac{1}{4}$ .
- (c)  $|C| = 2$  (i.e. the numbers 10, 12). Hence,  $P(A \cap B) = \frac{2}{12} = \frac{1}{6}$ .
- (d)  $|D| = 7$  (i.e. the numbers 2, 4, 6, 8, 10, 11, 12). Hence,  $P(A \cup B) = \frac{7}{12}$ .

*Solution to Exercise 6.4* The total number of simple events is  $\binom{12}{2}$ .

- (a) The number of favourable simple events is one and therefore  $P(\text{right two presents}) = \frac{|A|}{|\Omega|} = \frac{1}{\binom{12}{2}} \approx 0.015$ .
- (b) The number of favourable simple events is  $\binom{10}{2}$  because the person draws two presents out of the ten “wrong” presents:  
 $P(\text{wrong two presents}) = \frac{|A|}{|\Omega|} = \binom{10}{2}/\binom{12}{2} \approx 0.682$ . In Sect. 8.1.8, we explain the hypergeometric distribution which will shed further light on this exercise.

*Solution to Exercise 6.5*

- (a) Let  $V$  denote the event that there is too much salt in the soup and let  $L$  denote the event that the chef is in love. We know that

$$P(V) = 0.2 \Rightarrow P(\bar{V}) = 0.8.$$

Similarly, we have

$$P(L) = 0.3 \Rightarrow P(\bar{L}) = 0.7.$$

We therefore get:

$$P(V \cap L) = P(V|L) \cdot P(L) = 0.6 \cdot 0.3 = 0.18.$$

$$P(\bar{V} \cap L) = P(L) - P(V \cap L) = 0.3 - 0.18 = 0.12.$$

$$P(V \cap \bar{L}) = P(V) - P(V \cap L) = 0.2 - 0.18 = 0.02.$$

$$P(\bar{V} \cap \bar{L}) = P(\bar{V}) - P(\bar{V} \cap L) = 0.8 - 0.12 = 0.68.$$

This yields the following contingency table:

	$V$	$\bar{V}$	Total
$L$	0.18	0.12	0.3
$\bar{L}$	0.02	0.68	0.7
Total	0.2	0.8	1

- (b) The variables are not stochastically independent since, for example,  $P(V) \cdot P(L) = 0.3 \cdot 0.2 = 0.06 \neq 0.18 = P(V \cap L)$ .

*Solution to Exercise 6.6*

- (a) We define the following events:  $G = \text{Basil is treated well}$ ,  $\bar{G} = \text{Basil is not treated well}$ ;  $E = \text{Basil survives}$ ,  $\bar{E} = \text{Basil dies}$ . We know that

$$P(\bar{G}) = \frac{1}{3} \Rightarrow P(G) = \frac{2}{3}; \quad P(E|G) = \frac{1}{2}; \quad P(E|\bar{G}) = \frac{3}{4}.$$

Using the Law of Total Probability, we get

$$\begin{aligned} P(E) &= P(E|G) \cdot P(G) + P(E|\bar{G}) \cdot P(\bar{G}) \\ &= \frac{1}{2} \cdot \frac{2}{3} + \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{3} + \frac{1}{4} = \frac{7}{12} \approx 0.58. \end{aligned}$$

- (b) We can use Bayes' Theorem to answer the question:

$$P(\bar{G}|E) = \frac{P(E|\bar{G}) \cdot P(\bar{G})}{P(E|\bar{G}) \cdot P(\bar{G}) + P(E|G) \cdot P(G)} = \frac{\frac{3}{4} \cdot \frac{1}{3}}{\frac{7}{12}} = \frac{3}{7} \approx 0.43.$$

*Solution to Exercise 6.7* We define the following events and probabilities

- A: Bill never paid,  $P(A) = 0.05 \Rightarrow P(\bar{A}) = 0.95$ .
- M: Bill paid too late,  $P(M) = ?$
- $P(M|\bar{A}) = 0.2$ .
- $P(M|A) = 1$  because someone who never pays will always pay too late.

- (a) We are interested in  $P(M)$ , the probability that someone does not pay his bill in a particular month, either because he is not able to or he pays too late. We can use the Law of Total Probability to obtain the results:

$$\begin{aligned} P(M) &= P(M|A)P(A) + P(M|\bar{A})P(\bar{A}) \\ &= 0.05 \cdot 1 + 0.2 \cdot 0.95 = 0.24. \end{aligned}$$

(b) We can use Bayes' Theorem to obtain the results:

$$P(A | M) = \frac{P(A)P(M | A)}{P(M)} = \frac{0.05}{0.24} = 0.208.$$

(c) If the bill was not paid in a particular month, the probability is 20.8 % that it will never be paid, and 78.2 % that it will be paid. One could argue that a preventive measure that affects almost 79 % of trustworthy customers are not ideal and the bank should therefore not block a credit card if a bill is not paid on time.

### *Solution to Exercise 6.8*

(a) The “and” operator refers to the joint distribution of two variables. In our example, we want to know the probability of being infected *and* having been transported by the truck. This probability can be directly obtained from the respective entry in the contingency table: 40 out of 200 cows fulfil both criteria and thus

$$P(B \cap A) = \frac{40}{200}.$$

(b) We can use  $P(A) = \frac{100}{200} = P(\bar{A})$  to obtain:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{40/200}{100/200} = \frac{40}{100}.$$

(c) Using these results and  $P(B) = \frac{60}{200}$ , and  $P(\bar{B}) = \frac{140}{200} = 1 - P(B)$ , we obtain

$$P(B|\bar{A}) = \frac{P(B \cap \bar{A})}{P(\bar{A})} = \frac{20/200}{100/200} = \frac{20}{100}$$

by using the Law of Total Probability. We can thus calculate

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \\ &= 0.40 \cdot 0.50 + 0.20 \cdot 0.50 = 0.30. \end{aligned}$$

This means that the probability of a cow being infected is 30 %. Alternatively, we could have simply looked at the marginal distribution of the contingency table to get  $P(B) = 60/200 = 0.3$ .

### *Solution to Exercise 6.9*

(a) The two shots are independent of each other and thus

$$P(A \cap B) = 0.4 \cdot 0.5 = 0.2.$$

$$P(A \cup B) = 0.4 + 0.5 - 0.2 = 0.7.$$

(b) We need to calculate

$$P(A \setminus B \cup B \setminus A) = 0.4 - 0.2 + 0.5 - 0.2 = 0.5.$$

(c)

$$P(B \setminus A) = 0.5 - 0.2 = 0.3.$$

## Chapter 7

*Solution to Exercise 7.1*

- (a) The first derivative of the CDF yields the PDF,  $F'(x) = f(x)$ :

$$f(x) = \begin{cases} 0 & \text{if } x < 2 \\ -\frac{1}{2}x + 2 & \text{if } 2 \leq x \leq 4 \\ 0 & \text{if } x > 4. \end{cases}$$

- (b) We know from Theorem 7.2.3 that for any continuous variable  $P(X = x_0) = 0$  and therefore  $P(X = 4) = 0$ . We calculate  $P(X < 3) = P(X \leq 3) - P(X = 3) = F(3) - 0 = -\frac{9}{4} + 6 - 3 = 0.75$ .

- (c) Using (7.15), we obtain the expectation as

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^2 x0 dx + \int_2^4 x \left(-\frac{1}{2}x + 2\right) dx + \int_4^{\infty} x0 dx \\ &= 0 + \int_2^4 \left(-\frac{1}{2}x^2 + 2x\right) dx + 0 \\ &= \left[-\frac{1}{6}x^3 + x^2\right]_2^4 = \left(-\frac{64}{6} + 16\right) - \left(-\frac{8}{6} + 4\right) = \frac{8}{3}. \end{aligned}$$

Given that we have already calculated  $E(X)$ , we can use Theorem 7.3.1 to calculate the variance as  $\text{Var}(X) = E(X^2) - [E(X)]^2$ . The expectation of  $X^2$  is

$$\begin{aligned} E(X^2) &= \int_2^4 x^2 \left(-\frac{1}{2}x + 2\right) dx = \int_2^4 \left(-\frac{1}{2}x^3 + 2x^2\right) dx \\ &= \left[-\frac{1}{8}x^4 + \frac{2}{3}x^3\right]_2^4 = \left(-32 + \frac{128}{3}\right) - \left(-2 + \frac{16}{3}\right) = \frac{22}{3}. \end{aligned}$$

We thus obtain  $\text{Var}(X) = \frac{22}{3} - \left(\frac{8}{3}\right)^2 = \frac{66-64}{9} = \frac{2}{9}$ .

*Solution to Exercise 7.2*

- (a) The probability mass function of  $X$  is

$x_i$	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{3}{9}$

Using (7.16), we calculate the expectation as

$$\begin{aligned} E(X) &= 1 \cdot \frac{1}{9} + 2 \cdot \frac{1}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{2}{9} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{3}{9} \\ &= \frac{1+2+3+8+5+18}{9} = \frac{37}{9} \approx 4.1. \end{aligned}$$

To obtain the variance, we need

$$\begin{aligned} E(X^2) &= 1 \cdot \frac{1}{9} + 4 \cdot \frac{1}{9} + 9 \cdot \frac{1}{9} + 16 \cdot \frac{2}{9} + 25 \cdot \frac{1}{9} + 36 \cdot \frac{3}{9} \\ &= \frac{1+4+9+32+25+108}{9} = \frac{179}{9}. \end{aligned}$$

Therefore, using  $\text{Var}(X) = E(X^2) - [E(X)]^2$ , we get

$$\text{Var}(X) = \frac{179}{9} - \left(\frac{37}{9}\right)^2 = \frac{1611 - 1369}{81} = \frac{242}{81} \approx 2.98.$$

The manipulated die yields on average higher values than a fair die because its expectation is  $4.1 > 3.5$ . The variability of is, however, similar because  $2.98 \approx 2.92$ .

- (b) The probability mass function of  $Y = \frac{1}{X}$  is:

$y_i = \frac{1}{x_i}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$
$P\left(\frac{1}{X} = y\right)$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{3}{9}$

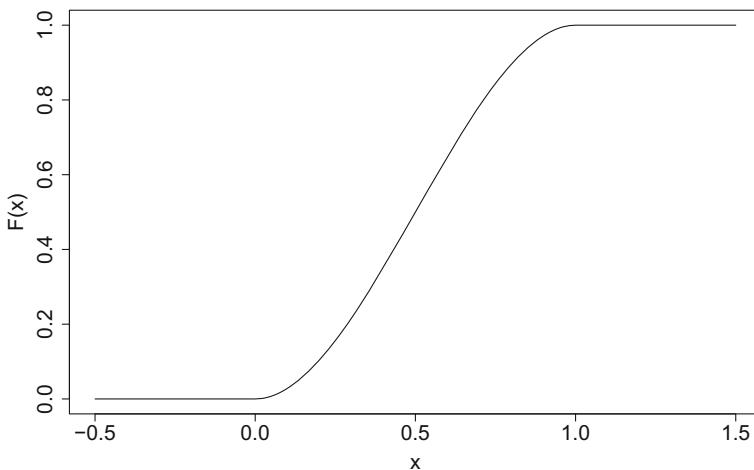
The expectation can hence be calculated as

$$\begin{aligned} E(Y) &= E\left(\frac{1}{X}\right) = 1 \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9} + \frac{1}{3} \cdot \frac{1}{9} + \frac{1}{4} \cdot \frac{2}{9} + \frac{1}{5} \cdot \frac{1}{9} + \frac{1}{6} \cdot \frac{3}{9} \\ &= \frac{1}{9} + \frac{1}{18} + \frac{1}{27} + \frac{1}{18} + \frac{1}{45} + \frac{1}{18} = \frac{91}{270}. \end{aligned}$$

Comparing the results from (a) and (b) shows clearly that  $E\left(\frac{1}{X}\right) \neq \frac{1}{E(X)}$ . Recall that  $E(bX) = bE(X)$ . It is interesting to see that for some transformations  $T(X)$  it holds that  $E(T(X)) = T(E(X))$ , but for some it does not. This reminds us to be careful when thinking of the implications of transformations.

### Solution to Exercise 7.3

- (a) There are several ways to plot the CDF. One possibility is to define the function and plot it with the `curve` command. Since the function has different definitions for the intervals  $[\infty, 0]$ ,  $[0, 1]$ ,  $(1, \infty]$ , we need to take this into account. Remember that a logical statement in R corresponds to a number, i.e. `TRUE = 1` and `FALSE = 0`; we can thus simply add the different pieces of the function and multiply them with a condition which specifies if  $X$  is contained in the interval or not (Fig. B.15):



**Fig. B.15** Cumulative distribution function for the proportion of wine sold

```
cdf <- function(x){
  (3 * x^2 - 2 * x^3) * (x <= 1 & x >= 0) + 1 * (x > 1) + 0 * (x < 0)
}
curve(cdf,from=-0.5,to=1.5)
```



(b) The PDF is

$$\frac{d}{dx}F(x) = F'(x) = f(x) = \begin{cases} 6(x - x^2) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

(c)

$$\begin{aligned} P\left(\frac{1}{3} \leq X \leq \frac{2}{3}\right) &= \int_{\frac{1}{3}}^{\frac{2}{3}} f(x)dx = F\left(\frac{2}{3}\right) - F\left(\frac{1}{3}\right) \\ &= \left[3\left(\frac{2}{3}\right)^2 - 2\left(\frac{2}{3}\right)^3\right] - \left[3\left(\frac{1}{3}\right)^2 - 2\left(\frac{1}{3}\right)^3\right] \\ &= 0.48149. \end{aligned}$$

(d) We have already defined the CDF in (a). We can now simply plug in the  $x$ -values of interest:

```
cdf(2/3)-cdf(1/3)
```



(e) The variance can be calculated as follows:

$$\mathbb{E}(X) = \int_0^1 x \cdot 6(x - x^2) dx = 6 \cdot \int_0^1 (x^2 - x^3) dx$$

$$= 6 \left[ \frac{1}{3}x^3 - \frac{1}{4}x^4 \right]_0^1 = 6 \cdot \left( \frac{1}{3} - \frac{1}{4} \right) = 0.5$$

$$\mathbb{E}(X^2) = \int_0^1 x^2 \cdot 6(x - x^2) dx = 6 \cdot \int_0^1 (x^3 - x^4) dx$$

$$= 6 \left[ \frac{1}{4}x^4 - \frac{1}{5}x^5 \right]_0^1 = 6 \cdot \left( \frac{1}{4} - \frac{1}{5} \right) = 0.3$$

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = 0.3 - 0.5^2 = 0.05.$$

*Solution to Exercise 7.4*

(a) Two conditions need to be satisfied for  $f(x)$  to be a proper PDF:

$$(i) \quad \int_0^2 f(x) dx = 1:$$

$$\begin{aligned} \int_0^2 f(x) dx &= \int_0^2 c \cdot x(2-x) dx = c \int_0^2 x(2-x) dx \\ &= c \int_0^2 (2x - x^2) dx = c \left[ x^2 - \frac{1}{3}x^3 \right]_0^2 \\ &= c \left[ 4 - \frac{8}{3} - (0 - 0) \right] = c \cdot \frac{4}{3} \stackrel{!}{=} 1 \\ \implies c &= \frac{3}{4}. \end{aligned}$$

$$(ii) \quad f(x) \geq 0:$$

$$f(x) = \frac{3}{4}x(2-x) \geq 0 \quad \forall x \in [0, 2].$$

(b) We calculate

$$F(x) = P(X \leq x) = \int_0^x f(t) dt = \int_0^x \frac{3}{4}t(2-t) dt$$

$$= \frac{3}{4} \int_0^x (2t - t^2) dt = \frac{3}{4} \left[ t^2 - \frac{1}{3}t^3 \right]_0^x$$

$$= \frac{3}{4} \left[ x^2 - \frac{1}{3}x^3 - 0 \right] = \frac{3}{4}x^2 \left( 1 - \frac{1}{3}x \right)$$

and therefore

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{3}{4}x^2(1 - \frac{1}{3}x) & \text{if } 0 \leq x \leq 2 \\ 1 & \text{if } x > 2. \end{cases}$$

(c) The expectation is

$$\begin{aligned} E(X) &= \int_0^2 xf(x)dx = \frac{3}{4} \int_0^2 (2x^2 - x^3)dx \\ &= \frac{3}{4} \left[ \frac{2}{3}x^3 - \frac{1}{4}x^4 \right]_0^2 = \frac{3}{4} \left[ \frac{2}{3} \cdot 8 - \frac{1}{4} \cdot 16 - 0 \right] \\ &= \frac{3}{4} \left[ \frac{16}{3} - \frac{12}{3} \right] = \frac{3}{4} \cdot \frac{4}{3} = 1. \end{aligned}$$

Using  $\text{Var}(X) = E(X^2) - (E(X))^2$ , we calculate the variance as

$$\begin{aligned} E(X^2) &= \int_0^2 x^2 f(x)dx = \frac{3}{4} \int_0^2 (2x^3 - x^4)dx \\ &= \frac{3}{4} \left[ \frac{2}{4}x^4 - \frac{1}{5}x^5 \right]_0^2 = \frac{3}{4} \left[ \frac{2}{4} \cdot 16 - \frac{1}{5} \cdot 32 - 0 \right] \\ &= 6 - \frac{3 \cdot 32}{4 \cdot 5} = 6 - \frac{3 \cdot 8}{5} = \frac{6}{5} \\ \text{Var}(X) &= \frac{6}{5} - 1^2 = \frac{1}{5}. \end{aligned}$$

(d)

$$P(|X - \mu| \leq 0.5) \geq 1 - \frac{\sigma^2}{c^2} = 1 - \frac{(\frac{1}{5})}{(0.5)^2} = 1 - 0.8 = 0.2.$$

*Solution to Exercise 7.5*

(a) The marginal distributions are obtained by the row and column sums of the joint PDF, respectively. For example,  $P(X = 1) = \sum_{j=1}^J p_{1j} = p_{1+} = 1/4$ .

X	$P(X = x_i)$	Y	$P(Y = y_i)$
0	3/4	1	1/6
1	1/4	2	7/12
		3	1/4

The marginal distribution of  $X$  tells us how many customers sought help via the telephone hotline (75 %) and via email (25 %). The marginal distribution of  $Y$  represents the distribution of the satisfaction level, highlighting that more than half of the customers (7/12) were “satisfied”.

- (b) To determine the 75 % quantile with respect to  $Y$ , we need to find the value  $y_{0.75}$  for which  $F(y_{0.75}) \geq 0.75$  and  $F(y) < 0.75$  for  $y < y_{0.75}$ .  $Y$  takes the values 1, 2, 3. The quantile cannot be  $y_{0.75} = 1$  because  $F(1) = 1/6 < 0.75$ . The 75 % quantile is  $y_{0.75} = 2$  because  $F(2) = 1/6 + 7/12 = 3/4 \geq 0.75$  and for all values which are smaller than 2 we get  $F(x) < 0.75$ .
- (c) We can calculate the conditional distribution using  $P(Y = y_j | X = 1) = p_{1j} / p_{1+} = p_{1j} / (1/6 + 1/12 + 0) = p_{1j} / (0.25)$ . Therefore,

$$\begin{aligned} P(Y = 1 | X = 1) &= \frac{1/6}{1/4} = \frac{2}{3}, \\ P(Y = 2 | X = 1) &= \frac{1/12}{1/4} = \frac{1}{3}, \\ P(Y = 3 | X = 1) &= \frac{0}{1/4} = 0. \end{aligned}$$

Among those who used the email customer service two-thirds were unsatisfied, one-third were satisfied, and no one was very satisfied.

- (d) As we know from (7.27), two discrete random variables are said to be independent if  $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$ . However, in our example,  $P(X = 0, Y = 1) = P(X = 0)P(Y = 1) = \frac{3}{4} \cdot \frac{1}{6} \neq 0$ . This means that  $X$  and  $Y$  are not independent.
- (e) The covariance of  $X$  and  $Y$  is defined as  $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$ . We calculate

$$\begin{aligned} E(X) &= 0 \cdot \frac{3}{4} + 1 \cdot \frac{1}{4} = \frac{1}{4} \\ E(Y) &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{7}{12} + 3 \cdot \frac{1}{4} = \frac{25}{12} \\ E(XY) &= 0 \cdot 1 \cdot 0 + 1 \cdot 1 \cdot \frac{1}{6} + 0 \cdot 2 \cdot \frac{1}{2} + 1 \cdot 2 \cdot \frac{1}{12} + 0 \cdot 3 \cdot \frac{1}{4} + 1 \cdot 3 \cdot 0 \\ &= \frac{2}{6} \\ \text{Cov}(X, Y) &= \frac{2}{6} - \frac{1}{4} \cdot \frac{25}{12} = -\frac{3}{16}. \end{aligned}$$

Since  $\text{Cov}(X, Y) < 0$ , we conclude that there is a negative relationship between  $X$  and  $Y$ : the higher the values of  $X$ , the lower the values of  $Y$ —and vice versa. For example, those who use the email-based customer service ( $X = 1$ ) are less satisfied than those who use the telephone customer service ( $X = 0$ ). It is, however, evident that in this example the values of  $X$  have no order and therefore, care must be exercised in interpreting the covariance.

*Solution to Exercise 7.6* Using Tschebyschev's inequality (7.24)

$$P(|X - \mu| < c) \geq 0.9 = 1 - \frac{\text{Var}(X)}{c^2},$$

we can determine  $c$  as follows:

$$\begin{aligned} 1 - \frac{\text{Var}(X)}{c^2} &= 0.9 \\ c^2 &= \frac{\text{Var}(X)}{0.1} = \frac{4}{0.1} = 40 \\ c &= \pm\sqrt{40} = \pm 6.325. \end{aligned}$$

Thus, the interval is  $[15 - 6.325; 15 + 6.325] = [8.675; 21.325]$ .

*Solution to Exercise 7.7*

(a) The joint PDF is:

		$Y$		
		0	1	2
$X$	-1	0.3	0.2	0.2
	2	0.1	0.1	0.1

(b) The marginal distributions are obtained from the row and column sums of the joint PDF, respectively:

$X$	-1	2	$Y$	0	1	2
$P(X = x)$	0.7	0.3	$P(Y = y)$	0.4	0.3	0.3

(c) The random variables  $X$  and  $Y$  are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall x, y.$$

However, in our example we have, for example,

$$P(X = -1, Y = 0) = 0.3 \neq P(X = -1) \cdot P(Y = 0) = 0.7 \cdot 0.4 = 0.28.$$

Hence, the two variables are not independent.

(d) The joint distribution of  $X$  and  $Y$  can be used to obtain the desired distribution of  $U$ . For example, If  $X = -1$  and  $Y = 0$ , then  $U = X + Y = -1$ . The respective probability is  $P(U = -1) = 0.3$  because  $P(U = -1) = P(X = -1, Y = 0) = 0.3$  and there is no other combination of  $X$ - and  $Y$ -values which yields  $X + Y = -1$ . The distribution of  $U$  is therefore as follows:

$k$	-1	0	1	2	3	4
$P(U = k)$	0.3	0.2	0.2	0.1	0.1	0.1

(e) We calculate

$$\mathbb{E}(U) = \sum_{k=-1}^4 k \cdot P(U = k) = 0.8$$

$$\mathbb{E}(X) = (-1)0.7 + 2 \cdot 0.3 = -0.1$$

$$\mathbb{E}(Y) = 0 \cdot 0.4 + 1 \cdot 0.3 + 2 \cdot 0.3 = 0.9$$

$$\mathbb{E}(U^2) = 0.3 \cdot (-1)^2 + \dots + 0.1 \cdot 4^2 = 3.4$$

$$\mathbb{E}(X^2) = 0.7 \cdot (-1)^2 + 0.3 \cdot 2^2 = 1.9$$

$$\mathbb{E}(Y^2) = 0.4 \cdot 0^2 + 0.3 \cdot 1^2 + 0.3 \cdot 2^2 = 1.5$$

$$\text{Var}(U) = \mathbb{E}(U^2) - [\mathbb{E}(U)]^2 = 3.4 - (0.8)^2 = 2.76$$

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = 1.9 - (-0.1)^2 = 1.89$$

$$\text{Var}(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = 1.5 - (0.9)^2 = 0.69.$$

It can be seen that  $\mathbb{E}(X) + \mathbb{E}(Y) = -0.1 + 0.9 = 0.8 = \mathbb{E}(U)$ . This makes sense because we know from (7.31) that  $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ . However,  $\text{Var}(U) = 2.76 \neq \text{Var}(X) + \text{Var}(Y) = 1.89 + 0.69$ . This follows from (7.7.1) which says that  $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\text{Cov}(X, Y)$  and therefore,  $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$  only if the covariance is 0. We know from (c) that  $X$  and  $Y$  are not independent and thus  $\text{Cov}(X, Y) \neq 0$ .

### Solution to Exercise 7.8

- (a) The random variables  $X$  and  $Y$  are independent if the balls are drawn with replacement. This becomes clear by understanding that drawing with replacement implies that for both the draws, the same balls are in the urn and the conditions in each draw remain the same. The first draw has no implications for the second draw.

If we were drawing the balls without replacement, then the first draw could possibly have implications for the second draw: for instance, if the first ball drawn was red, then the second one could not be red because there is only one red ball in the urn. This means that drawing without replacement implies dependency of  $X$  and  $Y$ . This can also be seen by evaluating the independence assumption (7.27):

$$P(X = 2, Y = 2) = 0 \neq P(X = 2) \cdot P(Y = 2) = \frac{1}{8} \cdot \frac{1}{8}.$$

- (b) The marginal probabilities  $P(X = x_i)$  can be obtained from the given information. For example, 3 out of 8 balls are black and thus  $P(X = 1) = 3/8$ . The conditional distributions  $P(Y|X = x_i)$  can be calculated easily by realizing that under the assumed dependency of  $X$  and  $Y$ , the second draw is always based on 7 balls (8 balls minus the one drawn under the condition  $X = x_i$ )—e.g. if the first ball drawn is black, then 7 balls, 2 of which are black, remain in the urn and  $P(Y = 1|X = 1) = 2/7$ . We thus calculate

$$\begin{aligned} P(Y = 1, X = 1) &= P(Y = 1|X = 1)P(X = 1) = \frac{2}{7} \cdot \frac{3}{8} = \frac{6}{56} \\ P(Y = 1, X = 2) &= P(Y = 1|X = 2)P(X = 2) = \frac{3}{7} \cdot \frac{1}{8} = \frac{3}{56} \\ &\dots \\ P(Y = 3, X = 3) &= P(Y = 3|X = 3)P(X = 3) = \frac{3}{7} \cdot \frac{4}{8} = \frac{12}{56} \end{aligned}$$

and obtain

		Y		
		1	2	3
X	1	$\frac{6}{56}$	$\frac{3}{56}$	$\frac{12}{56}$
	2	$\frac{3}{56}$	0	$\frac{4}{56}$
	3	$\frac{12}{56}$	$\frac{4}{56}$	$\frac{12}{56}$

- (c) The expectations are

$$\begin{aligned} E(X) &= 1 \cdot \frac{3}{8} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{4}{8} = \frac{17}{8} \\ E(Y) &= E(X) = \frac{17}{8}. \end{aligned}$$

To estimate  $\rho(X, Y)$ , we need  $\text{Cov}(X, Y)$  as well as  $\text{Var}(X)$  and  $\text{Var}(Y)$ :

$$\begin{aligned} E(XY) &= 1 \cdot \frac{6}{56} + 2 \cdot \frac{3}{56} + 3 \cdot \frac{12}{56} + 2 \cdot \frac{3}{56} + 4 \cdot 0 + 6 \cdot \frac{4}{56} + 3 \cdot \frac{12}{56} + 6 \cdot \frac{4}{56} + 9 \cdot \frac{12}{56} \\ &= \frac{246}{56} \\ E(X^2) &= E(Y^2) = 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{1}{8} + 3^2 \cdot \frac{4}{8} = \frac{43}{8} \\ \text{Var}(X) &= E(X^2) - [E(X)]^2 = \frac{43}{8} - \left(\frac{17}{8}\right)^2 = \frac{55}{64} \\ \text{Var}(Y) &= \text{Var}(X) = \frac{55}{64}. \end{aligned}$$

Using (7.38) and  $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$ , we obtain

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{\frac{246}{56} - \frac{289}{64}}{\sqrt{\frac{55}{64} \cdot \frac{55}{64}}} = -0.143.$$

*Solution to Exercise 7.9*

(a) The constant  $c$  must satisfy

$$\int_{40}^{100} \int_{10}^{100} c \left( \frac{100-x}{x} \right) dx dy = \int_{40}^{100} \int_{10}^{100} \frac{100c}{x} - c dx dy \stackrel{!}{=} 1$$

and therefore

$$\int_{40}^{100} [100c \ln(x) - cx]_{10}^{100} dy = \int_{40}^{100} 100c \ln 100 - 100c - 100c \ln(10) + 10c dy$$

which is

$$\begin{aligned} \int_{40}^{100} 100c \left( \ln \frac{100}{10} - 1 + \frac{1}{10} \right) dy &= [100cy (\ln 10 - 9/10)]_{40}^{100} \\ &= 600c(10 \ln 10 - 9) \rightarrow c \approx 0.00012. \end{aligned}$$

(b) The marginal distribution is

$$\begin{aligned} f_X(x) &= \int_{40}^{100} c \left( \frac{100-x}{x} \right) dy = \left[ c \left( \frac{100-x}{x} \right) y \right]_{40}^{100} \\ &= 100c \left( \frac{100-x}{x} \right) - 40c \left( \frac{100-x}{x} \right) \approx 0.00713 \left( \frac{100-x}{x} \right) \end{aligned}$$

for  $10 \leq x \leq 100$ .

(c) To determine  $P(X > 75)$ , we need the cumulative marginal distribution of  $X$ :

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f_X(t) dt = \int_{10}^x 0.00713 \left( \frac{100-t}{t} \right) dt \\ &= \int_{10}^x \frac{0.00713}{t} - 0.00713 dt = [0.713 \ln(t) - 0.00713 t]_{10}^x \\ &= 0.713 \ln(x) - 0.00713x - 0.00713 \ln(10) + 0.00713 \cdot 10. \end{aligned}$$

Now we can calculate

$$P(X > 75) = 1 - P(X \leq 75) = 1 - F_X(75) = 1 - 0.973 \approx 2.7\%.$$

(d) The conditional distribution is

$$f_{Y|X}(x, y) = \frac{f(x, y)}{f(x)} = \frac{c \left( \frac{100-x}{x} \right)}{60c \left( \frac{100-x}{x} \right)} = \frac{1}{60}.$$

*Solution to Exercise 7.10* If we evaluate the expectation with respect to  $Y$ , then both  $\mu$  and  $\sigma$  can be considered to be constants. We can therefore write

$$\mathbb{E}(Y) = \mathbb{E}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}(E(X) - \mu).$$

Since  $E(X) = \mu$ , it follows that  $E(Y) = 0$ . The variance is

$$\text{Var}(Y) = \text{Var}\left(\frac{X - \mu}{\sigma}\right).$$

Applying  $\text{Var}(a + bX) = b^2 \text{Var}(X)$  to this equation yields  $a = \mu$ ,  $b = \frac{1}{\sigma}$  and therefore

$$\text{Var}(Y) = \frac{1}{\sigma^2} \text{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1.$$

## Chapter 8

*Solution to Exercise 8.1* The random variable  $X$ : “number of packages with a toy” is binomially distributed. In each of  $n = 20$  “trials”, a toy can be found with probability  $p = \frac{1}{6}$ .

(a) We thus get

$$P(X = 4) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{20}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^{16} \approx 0.20.$$

(b) Similarly, we calculate

$$P(X = 0) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{20}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{20} = 0.026.$$

(c) This question relates to a hypergeometric distribution: there are  $N = 20$  packages with  $M = 3$  packages with toys and  $N - M = 17$  packages without a toy. The daughter gets  $n = 5$  packages and we are interested in  $P(X = 2)$ . Hence, we get

$$P(X = 2) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{3}{2} \binom{17}{3}}{\binom{20}{5}} \approx 0.13.$$

*Solution to Exercise 8.2* Given  $X \sim N(42.1, 20.8^2)$ , we get:

(a)

$$\begin{aligned} P(X \geq 50) &= 1 - P(X \leq 50) = 1 - \phi\left(\frac{x - \mu}{\sigma}\right) = 1 - \phi\left(\frac{50 - 42.1}{20.8}\right) \\ &= 1 - \phi(0.37) \approx 0.35. \end{aligned}$$

We obtain the same results in R as follows:

`1-pnorm(50,42.1,20.8)`



(b)

$$\begin{aligned} P(30 \leq X \leq 40) &= P(X \leq 40) - P(X \leq 30) \\ &= \phi\left(\frac{40 - 42.1}{20.8}\right) - \phi\left(\frac{30 - 42.1}{20.8}\right) \\ &= \phi(-0.096) - \phi(-0.577) = 1 - 0.538 - 1 + 0.718 \\ &\approx 18\%. \end{aligned}$$

We would have obtained the same results in R using:

`pnorm(40,42.1,20.8)-pnorm(30,42.1,20.8)`



*Solution to Exercise 8.3* The random variable  $X$  follows a discrete uniform distribution because  $p_i = \frac{1}{12}$  for each  $x_i$ . The expectation and variance are therefore

$$\begin{aligned} E(X) &= \frac{k+1}{2} = \frac{12+1}{2} = 6.5, \\ \text{Var}(X) &= \frac{1}{12}(12^2 - 1) \approx 11.92. \end{aligned}$$

*Solution to Exercise 8.4* Each guess is a Bernoulli experiment where the right answer is given with a probability of 50 %. The number of correct guesses therefore follows a binomial distribution, i.e.  $X \sim B(10; 0.5)$ . The probability of giving the right answer at least 8 times is identical to the probability of not being wrong more

than 2 times. We can thus calculate  $P(X \geq 8)$  as  $P(X \leq 2)$ :

$$\begin{aligned} P(X = 0) &= \binom{10}{0} 0.5^0 (1 - 0.5)^{10} \approx 0.000977 \\ P(X = 1) &= \binom{10}{1} 0.5^1 (1 - 0.5)^9 \approx 0.009766 \\ P(X = 2) &= \binom{10}{2} 0.5^2 (1 - 0.5)^8 \approx 0.043945. \end{aligned}$$

This relates to

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.000977 + 0.009766 + 0.043945 \approx 0.0547. \end{aligned}$$

We would have obtained the same results in R using:

```
pbisom(2, 10, 0.5)
1-pbisom(7, 10, 0.5)
```



### Solution to Exercise 8.5

- (a) It seems appropriate to model the number of fused bulbs with a Poisson distribution. We assume, however, that the probabilities of fused bulbs on two consecutive days are independent of each other; i.e. they only depend on  $\lambda$  but not on the time  $t$ .
- (b) The arithmetic mean is

$$\bar{x} = \frac{1}{30}(0 + 1 \cdot 8 + 2 \cdot 8 + \cdots + 5 \cdot 1) = \frac{52}{30} = 1.7333$$

which means that, on an average, 1.73 bulbs are fused per day. The variance is

$$\begin{aligned} s^2 &= \frac{1}{30}(0 + 1^2 \cdot 8 + 2^2 \cdot 8 + \cdots + 5^2 \cdot 1) - 1.7333^2 \\ &= \frac{142}{30} - 3.0044 = 1.72889. \end{aligned}$$

We see that mean and variance are similar, which is an indication that the choice of a Poisson distribution is appropriate since we assume  $E(X) = \lambda$  and  $\text{Var}(X) = \lambda$ .

- (c) The following table lists the proportions (i.e. relative frequencies  $f_j$ ) together with the probabilities  $P(X = x)$  from a  $Po(1.73)$ -distribution. As a reference, we also list the probabilities from a  $Po(2)$ -distribution since it is not practically possible that 1.73 bulbs stop working and it may hence be an option to round the mean.

	$f_i$	$Po(1.73)$	$Po(2)$
$P(X = 0)$	0.2	0.177	0.135
$P(X = 1)$	0.267	0.307	0.27
$P(X = 2)$	0.267	0.265	0.27
$P(X = 3)$	0.167	0.153	0.18
$P(X = 4)$	0.067	0.067	0.09
$P(X = 5)$	0.033	0.023	0.036

One can see that observed proportions and expected probabilities are close together which indicates again that the choice of a Poisson distribution was appropriate. Chapter 9 gives more details on how to estimate parameters, such as  $\lambda$ , from data if it is unknown.

- (d) Using  $\lambda = 1.73$ , we calculate

$$\begin{aligned} P(X > 5) &= 1 - P(X \leq 5) = 1 - \sum_{i=0}^5 \frac{\lambda^i}{i!} \exp(-\lambda) \\ &= 1 - \exp(-1.73) \left( \frac{1.73^0}{0!} + \frac{1.73^1}{1!} + \cdots + \frac{1.73^5}{5!} \right) \\ &= 1 - 0.99 = 0.01. \end{aligned}$$

Thus, the bulbs are replaced on only 1 % of the days.

- (e) If  $X$  follows a Poisson distribution then, given Theorem 8.2.1,  $Y$  follows an exponential distribution with  $\lambda = 1.73$ .
- (f) The expectation of an exponentially distributed variable is

$$E(Y) = \frac{1}{\lambda} = \frac{1}{1.73} = 0.578.$$

This means that, on average, it takes more than half a day until one of the bulbs gets fused.

### Solution to Exercise 8.6

- (a) Let  $X$  be a random variable describing “the number  $x$  of winning tickets among  $n$  bought tickets”; then  $X$  follows the hypergeometric distribution  $X \sim H(n, 500, 4000)$ . We need to determine  $n$  for the conditions specified. We are interested in

$$P(X \geq 3) = 1 - P(X = 2) - P(X = 1) - P(X = 0).$$

Using the PMF of the hypergeometric distribution

$$P(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}},$$

this equates to

$$P(X \geq 3) = 1 - \frac{\binom{500}{2} \binom{4000-500}{n-2}}{\binom{4000}{n}} - \frac{\binom{500}{1} \binom{4000-500}{n-1}}{\binom{4000}{1}} - \frac{\binom{500}{0} \binom{4000-500}{n}}{\binom{4000}{n}}.$$

We have the following requirement:

$$1 - \frac{\binom{500}{2} \binom{4000-500}{n-2}}{\binom{4000}{n}} - \frac{\binom{500}{1} \binom{4000-500}{n-1}}{\binom{4000}{1}} - \frac{\binom{500}{0} \binom{4000-500}{n}}{\binom{4000}{n}} \stackrel{!}{\geq} 0.99.$$

To solve this equation, we can program this function for  $P(X > 3; n)$  in R and evaluate it for different numbers of tickets sold, e.g. between 50 and 100 tickets:

```
raffle <- function(n){
  p <- 1 - ((choose(500,2)*choose(3500,n-2))/(choose(4000,n)))
  -((choose(500,1)*choose(3500,n-1))/(choose(4000,n)))
  -((choose(500,0)*choose(3500,n))/(choose(4000,n)))
  return(p)
}
raffle(50:100)
raffle(63:64)
```



The output shows that at least 64 tickets need to be bought to have a 99% guarantee that at least three tickets win. This equates to spending € 96.

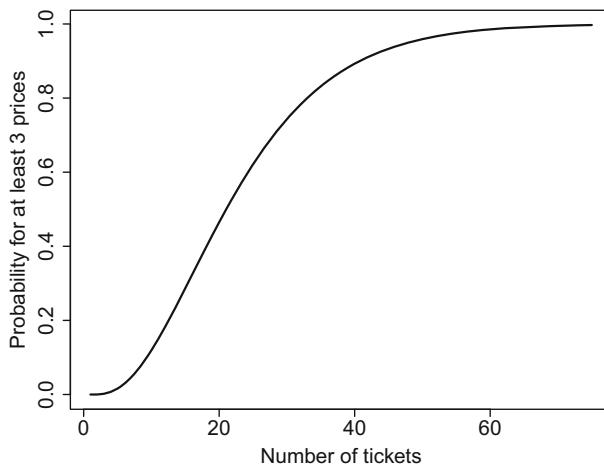
(b) We can plot the function as follows:

```
nb <- seq(1:75)
plot(nb,tombola(nb),type='l')
```



Figure B.16 shows the relationship between the number of tickets bought and the probability of having at least three winning tickets.

(c) The solution of (a) shows that it is well worth taking part in the raffle: Marco pays €96 and with a probability of 99 % and he wins at least three prizes which are worth  $€142 \cdot 3 = 426$ . More generally, the money generated by the raffle is  $€1.50 \times 4000 = 6000$ , but the prizes are worth  $€142 \cdot 500 = 71,000$ . One may suspect that the company produces the appliances much more cheaply than they are sold for and is thus so generous.



**Fig. B.16** Probability to have at least three winning tickets given the number of tickets bought

*Solution to Exercise 8.7* The probability of getting a girl is  $p = 1 - 0.5122 = 0.4878$ .

- (a) We are dealing with a geometric distribution here. Since we are interested in  $P(X \leq 3)$ , we can calculate:

$$P(X = 1) = 0.4878$$

$$P(X = 2) = 0.4878(1 - 0.4878) = 0.2498512$$

$$P(X = 3) = 0.4878(1 - 0.4878)^2 = 0.1279738$$

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.865625.$$

We would have obtained the same result in R using:

`pgem(2, 0.4878)`



Note that we have to specify “2” rather than “3” for  $x$  because R takes the number of unsuccessful trials rather than the number of trials until success.

- (b) Here we deal with a binomial distribution with  $k = 2$  and  $n = 4$ . We can calculate  $P(X = 2)$  as follows:

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k (1 - p)^{n-k} \\ &= \binom{4}{2} 0.4878^2 \cdot (1 - 0.4878)^2 = 0.3745536. \end{aligned}$$

R would have given us the same result using the `dbinom(2, 4, 0.4878)` command.

*Solution to Exercise 8.8*

- (a) The random variable  $Y$  follows a Poisson distribution, see Theorem 8.2.1 for more details.
- (b) The fisherman catches, on average, 3 fish an hour. We can thus assume that the rate  $\lambda$  is 3 and thus  $E(Y) = \lambda = 3$ . Similarly,  $E(X) = \frac{1}{\lambda} = \frac{1}{3}$  which means that it takes, on average, 20 min to catch another fish.
- (c) Using the PDF of the Poisson distribution, we get:

$$P(Y = 5) = \frac{3^5}{5!} \exp(-3) = 0.1 = 10\%$$

$$P(Y < 1) = P(Y = 0) = \frac{3^0}{0!} \exp(-3) \approx 0.0498 \approx 5\%.$$

We would have obtained the same results in *R* using the `dpois(5,3)` and `dpois(0,3)` commands.

*Solution to Exercise 8.9* The random variable  $\mathbf{X}$  = “choice of dessert” follows a multinomial distribution. More precisely,  $X_1$  describes whether chocolate brownies were chosen,  $X_2$  describes whether yoghurt was chosen,  $X_3$  describes whether lemon tart was chosen and  $\mathbf{X} = \{X_1, X_2, X_3\}$ .

- (a) Using the PMF of the multinomial distribution, we get

$$P(X_1 = n_1, X_2 = n_2, \dots, X_k = n_k) = \frac{n!}{n_1!n_2!\dots n_k!} \cdot p_1^{n_1} \cdots p_k^{n_k}$$

$$P(X_1 = 2, X_2 = 1, X_3 = 2) = \frac{5!}{2!1!2!} \cdot 0.2^2 \cdot 0.3^1 \cdot 0.5^2$$

$$= 9\%.$$

We would have obtained the same results in *R* as follows:

```
dmultinom(c(2,1,2),prob=c(0.2,0.3,0.5))
```



- (b) The probability of choosing lemon tart for the first two guests is 1. We thus need to determine the probability that 3 out of the remaining 3 guests order lemon tart:

$$P(X_1 = 0, X_2 = 0, X_3 = 3) = \frac{3!}{0!0!3!} \cdot 0.2^0 \cdot 0.3^0 \cdot 0.5^3$$

$$= 12.5\%.$$

Using `dmultinom(c(0,0,3),prob=c(0.2,0.3,0.5))` in *R*, we get the same result.

(c) The expectation vector is

$$E(\mathbf{X}) = (np_1, \dots, np_k) = (20 \cdot 0.2, 20 \cdot 3, 20 \cdot 0.5) = (4, 6, 10).$$

This means we expect 4 guests to order brownies, 6 to order yoghurt, and 10 to order lemon tart. The covariance matrix can be determined as follows:

$$\text{Cov}(X_i, X_j) = \begin{cases} np_i(1 - p_i) & \text{if } i = j \\ -np_i p_j & \text{if } i \neq j. \end{cases}$$

Using  $n = 20$ ,  $p_1 = 0.2$ ,  $p_2 = 0.3$  and  $p_3 = 0.5$ , we obtain the covariance matrix as:

$$\begin{pmatrix} 3.2 & -1.2 & -2 \\ -1.2 & 4.2 & -3 \\ -2 & -3 & 5 \end{pmatrix}$$

### Solution to Exercise 8.10

$$\begin{aligned} P(S \geq 1, W \geq 1) &\stackrel{\text{indep.}}{=} P(S \geq 1) \cdot P(W \geq 1) \\ &= (1 - P(S = 0)) \cdot (1 - P(W = 0)) \\ &= (1 - e^{-3} \frac{3^0}{0!}) \cdot (1 - e^{-4} \frac{4^0}{0!}) \\ &\approx 0.93. \end{aligned}$$

### Solution to Exercise 8.11

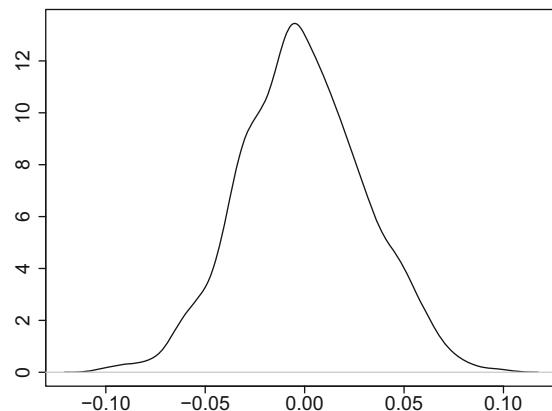
(a) Random numbers of a normal distribution can be generated using the `rnorm` command. By default  $\mu = 0$  and  $\sigma = 1$  (see `?rnorm`), so we do not need to specify these parameters. We simply need to set  $n = 1000$ . The mean of the 1000 realizations can thus be obtained using `mean(rnorm(1000))`. We can, for example, write a `for` loop to repeat this process 1000 times. An empty (`=NA`) vector of size 1000 can be used to store and evaluate the results:

```
set.seed(24121980)
R <- 1000
means <- c(rep(NA, R))
for(i in 1:R){means[i] <- mean(rnorm(1000))}
mean(means)
[1] -0.0007616465
var(means)
[1] 0.0009671311
plot(density(means))
```



We see that the mean of the arithmetic means is close to zero, but not exactly zero. The variance is approximately  $\sigma^2/n = 1/1000 = 0.001$ , as one would expect

**Fig. B.17** Kernel density plot of the distribution simulated in Exercise 8.11a



from the Central Limit Theorem. The distribution is symmetric, similar to a normal distribution, see Fig. B.17. It follows that  $\bar{X}_n$  is approximately  $N(\mu, \frac{\sigma^2}{n})$  distributed, as one could expect from the Theorem of Large Numbers and the Central Limit Theorem. It is important to understand that  $\bar{X}$  is not fixed but a random variable which follows a distribution, i.e. the normal distribution.

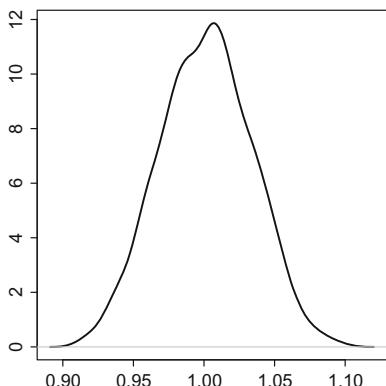
- (b) We can use the same code as above, except we use the exponential instead of the normal distribution:

```
means2 <- c(rep(NA,R))
for(i in 1:R){means2[i] <- mean(rexp(1000))}
mean(means2)
[1] 1.001321
var(means2)
[1] 0.001056113
plot(density(means))
```

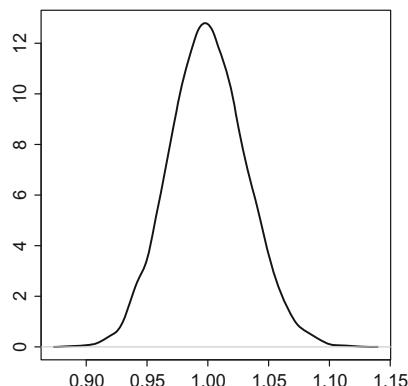


The realizations are i.i.d. observations. Once can see that, as in a),  $\bar{X}_n$  is approximately  $N(\mu, \frac{\sigma^2}{n}) = N(1, 1/1000)$  distributed. It is evident that the  $X_i$  do not necessarily need to follow a normal distribution for  $\bar{X}$  to follow a normal distribution, see also Fig. B.18a.

- (c) Increasing the number of repetitions makes the distribution look closer to a normal distribution, see Fig. B.18b. This visualizes that as  $n$  tends to infinity  $\bar{X}_n$  gets closer to a  $N(\mu, \frac{\sigma^2}{n})$ -distribution.



(a) 1,000 repetitions



(b) 10,000 repetitions

**Fig. B.18** Kernel density plots for Exercises 8.11b and 8.11c

## Chapter 9

### Solution to Exercise 9.1

- (a) The exercise tells us that  $X_i \stackrel{iid}{\sim} Po(\lambda)$ ,  $i = 1, 2, \dots, n$ . Let us look at the realizations  $x_1, x_2, \dots, x_n$ : under the assumption of independence, which we know is fulfilled because the  $X_i$ 's are i.i.d., and we can write the likelihood function as the product of the  $n$  PMF's of the Poisson distribution:

$$L(\theta; x) = \prod_{i=1}^n f(x_i; \theta) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum x_i}}{\prod x_i!} e^{-n\lambda}.$$

It is better to work on a log-scale because it is easy to differentiate. The results are identical no matter whether we use the likelihood function or the log-likelihood function because the log transformation is monotone in nature. The log-likelihood function is:

$$\ln L = \sum x_i \ln \lambda - \ln(x_1! \cdots x_n!) - n\lambda.$$

Differentiating with respect to  $\lambda$  yields

$$\frac{\partial \ln L}{\partial \lambda} = \frac{1}{\lambda} \sum x_i - n = 0$$

which gives us the ML estimate:

$$\hat{\lambda} = \frac{1}{n} \sum x_i = \bar{x}.$$

We need to confirm that the second derivative is  $< 0$  at  $\hat{\lambda} = \bar{x}$ ; otherwise, the solution would be a minimum rather than a maximum. We get

$$\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{1}{\hat{\lambda}^2} \sum x_i = -\frac{n}{\hat{\lambda}} < 0.$$

It follows that the arithmetic mean  $\bar{x} = \hat{\lambda}$  is the maximum likelihood estimator for the parameter  $\lambda$  of a Poisson distribution.

- (b) Using the results from (a) we can write the log-likelihood function for  $x_1 = 4, x_2 = 3, x_3 = 8, x_4 = 6, x_5 = 6$  as:

$$\ln L = 27 \ln \lambda - \ln(4! 3! 8! 6! 6!) - 5\lambda.$$

because  $\sum x_i = 27$ . We can write down this function in R as follows:

```
MLP <- function(lambda){  
  27*log(lambda) - log(factorial(4)*...*factorial(6)) -  
  5*lambda  
}
```



The function can be plotted using the `curve` command:

```
curve(MLP, from=0, to=10)
```

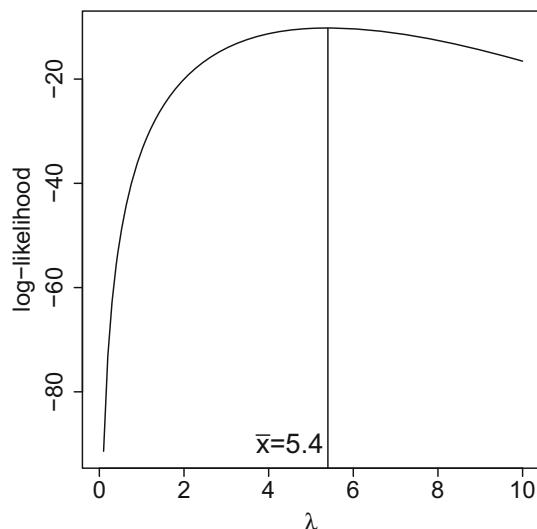


Figure B.19 shows the log-likelihood function. It can be seen that the function reaches its maximum at  $\bar{x} = 5.4$ .

- (c) Using (a) we can write the likelihood function as

$$L(\theta; x) = \prod_{i=1}^n f(x_i; \theta) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum x_i}}{\prod x_i!} e^{-n\lambda} = \underbrace{\lambda^{\sum x_i} e^{-n\lambda}}_{g(t, \lambda)} \underbrace{\frac{1}{\prod x_i!}}_{h(x_1, \dots, x_n)}.$$

**Fig. B.19** Illustration of the log-likelihood function



This means  $T = \sum_{i=1}^n x_i$  is sufficient for  $\lambda$ . The arithmetic mean, which is the maximum likelihood estimate, is a one-to-one function of  $T$  and therefore sufficient too.

### Solution to Exercise 9.2

- (a) The probability density function of a normal distribution equates to

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

with  $\infty < x < \infty, -\infty < \mu < \infty, \sigma^2 > 0$ . The likelihood function is therefore

$$L(x_1, x_2, \dots, x_n | \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

To find the maximum of  $L(x_1, x_2, \dots, x_n | \mu, \sigma^2)$ , it is again easier to work with the *log*-likelihood function, which is

$$l = \ln L(x_1, x_2, \dots, x_n | \mu, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \sum_{i=1}^n \left(\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Assuming  $\sigma^2$  to be 1, differentiating the log-likelihood function with respect to  $\mu$ , and equating it to zero gives us

$$\frac{\partial l}{\partial \mu} = 2 \sum_{i=1}^n \left(\frac{x_i - \mu}{1^2}\right) = 0 \quad \Leftrightarrow \quad n\mu = \sum_{i=1}^n x_i.$$

The ML estimate is therefore  $\hat{\mu} = \bar{x}$ .

- (b) Looking at the differentiated log-likelihood function in (a) shows us that the ML estimate of  $\mu$  is always the arithmetic mean, no matter whether  $\sigma^2$  is 1 or any other number.
- (c) Differentiating the log-likelihood function from (a) with respect to  $\sigma^2$  yields

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0.$$

Using  $\hat{\mu} = \bar{x}$  we calculate  $\frac{\partial l}{\partial \sigma^2} = 0$  as

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Since the parameter  $\theta = (\mu, \sigma^2)$  is two-dimensional, it follows that one needs to solve two ML equations, where  $\hat{\mu}_{\text{ML}}$  is estimated first, and  $\hat{\sigma}_{\text{ML}}^2$  second (as we did above). It follows further that one needs to look at the positive definiteness

of a matrix (the so-called information matrix) when checking that the second-order derivatives of the estimates are positive and therefore the estimates yield a maximum rather than a minimum. However, we omit this lengthy and time-consuming task here.

*Solution to Exercise 9.3* The probability density function of  $U(0, \theta)$  is

$$f(x) = \frac{1}{\theta} \text{ if } 0 < x < \theta \text{ and } 0 \text{ otherwise.}$$

Note that this equates to the PDF from Definition 8.2.1 for  $a = 0$  and  $b = \theta$ . The likelihood function is therefore

$$L(x_1, x_2, \dots, x_n | \theta) = \left(\frac{1}{\theta}\right)^n \text{ if } 0 < x_i < \theta \text{ and } 0 \text{ otherwise.}$$

One can see that  $L(x_1, x_2, \dots, x_n | \theta)$  increases as  $\theta$  decreases. The maximum of the likelihood function is therefore achieved for the smallest valid  $\theta$ . In particular,  $\theta$  is minimized when  $\theta \geq \max(x_1, x_2, \dots, x_n) = x_{(n)}$ . This follows from the definition of the PDF which requires that  $0 < x_i < \theta$  and therefore  $\theta > x_i$ . Thus, the maximum likelihood estimate of  $\theta$  is  $x_{(n)}$ , the greatest observed value in the sample.

*Solution to Exercise 9.4*

(a)  $T_n(X)$  is unbiased, and therefore also asymptotically unbiased, because

$$\mathbb{E}(T_n(X)) = \mathbb{E}(nX_{\min}) \stackrel{(7.29)}{=} n \frac{1}{n\lambda} = \frac{1}{\lambda} = \mu.$$

Similarly,  $V_n(X)$  is unbiased, and therefore also asymptotically unbiased, because

$$\mathbb{E}(V_n(X)) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \stackrel{(7.29)}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \frac{1}{n} n \frac{1}{\lambda} = \mu.$$

(b) To calculate the MSE we need to determine the bias and the variance of the estimators as we know from Eq. (9.5). It follows from (a) that both estimators are unbiased and hence the bias is 0. For the variances we get:

$$\text{Var}(T_n(X)) = \text{Var}(nX_{\min}) \stackrel{(7.33)}{=} n^2 \text{Var}(X_{\min}) = n^2 \frac{1}{n^2 \lambda^2} = \mu^2.$$

$$\text{Var}(V_n(X)) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \stackrel{(7.33)}{=} \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} n \frac{1}{\lambda^2} = \frac{1}{n} \mu^2.$$

Since the mean squared error consists of the sum of the variance and squared bias, the MSE for  $T_n(X)$  and  $V_n(X)$  are  $\mu^2$  and  $n^{-1}\mu^2$ , respectively. One can see that the larger  $n$ , the more superior  $V_n(X)$  over  $T_n(X)$  in terms of the mean squared error. In other words,  $V_n(X)$  is more efficient than  $T_n(X)$  because its variance is lower for any  $n > 1$ .

- (c) Using the results from (b), we get

$$\lim_{n \rightarrow \infty} \text{MSE}(V_n(X)) = \lim_{n \rightarrow \infty} \frac{1}{n} \mu^2 = 0.$$

This means the MSE approaches 0 as  $n$  tends to infinity. Therefore,  $V_n(X)$  is MSE consistent for  $\mu$ . Since  $V_n(X)$  is MSE consistent, it is also weakly consistent.

### *Solution to Exercise 9.5*

- (a) The point estimate of  $\mu$  is  $\bar{x}$  which is

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{24}(450 + \dots + 790) = 667.92.$$

The variance of  $\sigma^2$  can be estimated unbiasedly using  $s^2$ :

$$\begin{aligned} \hat{\sigma}^2 = s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{23}((450 - 667.92)^2 + \dots + (790 - 667.92)^2) \approx 18,035. \end{aligned}$$

- (b) The variance is unknown and needs to be estimated. We thus need the  $t$ -distribution to construct the confidence interval. We can determine  $t_{23;0.975} \approx 2.07$  using `qt(0.975, 23)` or Table C.2 (though the latter is not detailed enough),  $\alpha = 0.05$ ,  $\bar{x} = 667.97$  and  $\hat{\sigma}^2 = 18,035$ . This yields

$$\begin{aligned} I_l(X) &= \bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 667.92 - t_{23;0.975} \cdot \frac{\sqrt{18,035}}{\sqrt{24}} \approx 611.17, \\ I_u(X) &= \bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 667.92 + t_{23;0.975} \cdot \frac{\sqrt{18,035}}{\sqrt{24}} \approx 724.66. \end{aligned}$$

The confidence interval for  $\mu$  is thus [611.17; 724.66].

- (c) We can reproduce these results in R as follows:

```
eland <- c(450, 730, 700, 600, 620, , 790)
t.test(eland)$conf.int
```



*Solution to Exercise 9.6*

- Let us start with the confidence interval for the “Brose Baskets Bamberg”. Using  $t_{15;0.975} = 2.1314$  (`qt(0.975, 15)` or Table C.2) and  $\alpha = 0.05$ , we can determine the confidence interval as follows:

$$I_l(Ba) = \bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 199.06 - t_{15;0.975} \cdot \frac{7.047}{\sqrt{16}} = 195.305,$$

$$I_u(Ba) = \bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 199.06 + t_{15;0.975} \cdot \frac{7.047}{\sqrt{16}} = 202.815.$$

Thus, we get [195.305; 202.815].

- For the “Bayer Giants Leverkusen”, we use  $t_{13;0.975} = 2.1604$  to get

$$I_l(L) = \bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 196 - t_{13;0.975} \cdot \frac{9.782}{\sqrt{14}} = 190.352,$$

$$I_u(L) = \bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 196 + t_{13;0.975} \cdot \frac{9.782}{\sqrt{14}} = 201.648.$$

This leads to a confidence interval of [190.352; 201.648].

- For “Werder Bremen”, we need to use the quantile  $t_{22;0.975} = 2.0739$  which yields a confidence interval of

$$I_l(Br) = \bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 187.52 - t_{22;0.975} \cdot \frac{5.239}{\sqrt{23}} = 185.255,$$

$$I_u(Br) = \bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 187.25 + t_{22;0.975} \cdot \frac{5.239}{\sqrt{23}} = 189.786.$$

The interval is therefore [185.255; 189.786].

- The mean heights of the basketball teams are obviously larger than the mean height of the football team. The two confidence intervals of the basketball teams overlap, whereas the intervals of the football team with the two basketball teams do not overlap. It is evident that this indicates that the height of football players is substantially less than the height of basketball players. In Chap. 10, we will learn that confidence intervals can be used to test hypotheses about mean differences.

*Solution to Exercise 9.7*

- (a) Using  $n = 98$ , we calculate an unbiased point estimate for  $p$  using  $\bar{x} = \hat{p}$ . Note that  $x_i = 1$  if the wife has to wash the dishes.

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{98} \cdot 59 = \frac{59}{98} \approx 0.602.$$

- (b) Since  $n\hat{p}(1 - \hat{p}) = 98 \cdot 0.602 \cdot 0.398 = 23.48 > 9$  and  $p$  is sufficiently large, we can use the normal approximation to calculate a confidence interval for  $p$ . Using  $z_{1-\alpha/2} = z_{0.975} = 1.96$ , we obtain

$$I_l(X) = 0.602 - 1.96 \sqrt{\frac{0.602 \cdot 0.398}{98}} = 0.505,$$

$$I_u(X) = 0.602 + 1.96 \sqrt{\frac{0.602 \cdot 0.398}{98}} = 0.699.$$

This yields a confidence interval of  $[0.505, 0.699]$ . Note that the confidence interval does not contain  $p = 0.5$  which is the probability that would be expected if the coin was fair. This is a clear sign that the coin might be unfair.

- (c) If the coin is fair, we can use  $\hat{p} = 0.5$  as our prior judgement. We would then need

$$\begin{aligned} n &\geq \left[ \frac{z_{1-\alpha/2}}{\Delta} \right]^2 \hat{p}(1 - \hat{p}) \\ &\geq \left[ \frac{1.96}{0.005} \right]^2 0.5^2 = 38,416 \end{aligned}$$

dinners to get the desired precision—which is not possible as this would constitute a time period of more than 100 years. This shows that the expectation of such a high precision is unrealistic and needs to be modified. However, as the results of (b) show, the coin may not be fair. If the coin is indeed unfair, we may have, for example,  $p = 0.6$  and  $1 - p = 0.4$  which gives a smaller sample size. We can thus interpret the sample size as a conservative estimate of the highest number of dinners needed.

*Solution to Exercise 9.8* If a student fails then  $x_i = 1$ . We know that  $\sum x_i = 11$  and  $n = 104$ .

- (a) Using  $\bar{x} = p$  as point estimate, we get  $\hat{p} = \frac{11}{104} \approx 0.106 = 10.6\%$ . Using  $z_{1-\alpha/2} = z_{0.975} = 1.96$ , we can calculate the confidence interval as

$$0.106 \pm 1.96 \cdot \sqrt{\frac{0.106 \cdot 0.894}{104}} = [0.047; 0.165].$$

Using R we get:

```
binom.test(11,104)$conf.int
[1] 0.05399514 0.18137316
```



This result is different because the above command does not use the normal approximation. In addition,  $p$  is rather small which means that care must be exercised when using the results of the confidence interval with normal approximation in this context.

- (b) The point estimate of 10.6 % is substantially higher than 3.2 %. The lower bound confidence interval is still larger than the proportion of failures at county level. This indicates that the school is doing worse than most other schools in the county.

*Solution to Exercise 9.9*

- (a) Whether the  $i$ th household has switched on the TV and watches “Germany’s next top model” (GNTM) relates to a random variable  $X_i$  with

$$X_i = 1 : \text{if TV switched on and household watching GNTM}$$

$$X_i = 0 : \text{if TV switched off or household watches another show.}$$

It follows that  $X = \sum_{i=1}^{2500} X_i$  is the random variable describing the number of TVs, out of 2500 TVs, which are switched on and show GNTM. Since the  $X_i$ ’s can be assumed to be i.i.d., we can say that  $X$  follows a binomial distribution, i.e.  $X \sim B(2500; p)$  with  $p$  unknown. The length of the confidence interval for  $p$ ,

$$\left[ \hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right],$$

is

$$L = 2z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Unfortunately,  $\hat{p}(1 - \hat{p})$  is unknown but  $\hat{p}(1 - \hat{p}) \leq \frac{1}{4}$  because the maximum value for  $\hat{p}(1 - \hat{p})$  is 0.25 if  $\hat{p} = 0.5$ . Hence

$$L \leq 2z_{1-\alpha/2} \sqrt{\frac{\frac{1}{4}}{n}} = \frac{1.96}{\sqrt{2500}} = 0.0392.$$

This means the precision is half the length, i.e.  $\pm 0.0196 = \pm 1.96\%$ , in the worst case.

- (b) There is the danger of selection bias. A total of 500 households refused to take part in the study. It may be that their preferences regarding TV shows are different from the other 2500 households. For example, it may well be possible that those watching almost no TV refuse to be included; or that those watching TV shows which are considered embarrassing by society are refusing as well. In general, missing data may cause point estimates to be biased, depending on the underlying mechanism causing the absence.

*Solution to Exercise 9.10*

- (a) Using  $z_{1-\alpha/2} = 1.96$  and  $\hat{\sigma} = 0.233$ , we obtain an optimal sample size of at least

$$n_{opt} \geq \left[ 2 \frac{z_{1-\alpha/2} \sigma_0}{\Delta} \right]^2 = \left[ 2 \cdot \frac{1.96 \cdot 0.233}{0.2} \right]^2 = 20.85.$$

To calculate a confidence interval with a width of not more than 0.2 s, the results of at least 21 athletes are needed.

- (b) The sample size is 30. Thus, the confidence interval width should be smaller than 0.2 s. This is indeed true as the calculations show:

$$\left[ 10.93 \pm \underbrace{t_{0.975; 29}}_{2.045} \cdot \frac{0.233}{\sqrt{30}} \right] = [10.84; 11.02].$$

The width is only 0.18 s.

- (c) If we calculate a 80 % confidence interval, then the lower confidence limit corresponds to the running time which is achieved by only 10 % of the athletes (of the population). Using  $t_{0.9; 29} \approx 1.31$  we get

$$\left[ 10.93 \pm 1.31 \cdot \frac{0.233}{\sqrt{30}} \right] = [10.87; 10.99].$$

The athlete's best time is thus below the lower confidence limit. He is among the top 10 % of all athletes, using the results of the confidence interval.

*Solution to Exercise 9.11*

- (a) The odds ratio is

$$OR = \frac{163 \cdot 477}{475 \cdot 151} \approx 1.08.$$

This means the chances that a pizza arrives in time are about 1.08 times higher for Laura compared with Melissa. To calculate the 95 % confidence interval, we need  $\hat{\theta}_0 = \ln(1.08) \approx 0.077$ ,  $z_{1-\alpha/2} \approx 1.96$ , and

$$\hat{\sigma}_{\hat{\theta}_0} = \left( \frac{1}{163} + \frac{1}{475} + \frac{1}{151} + \frac{1}{477} \right)^{\frac{1}{2}} = 0.13.$$

The interval for the log odds ratio is

$$[\ln(1.08) \pm 1.96 \cdot 0.13] \approx [-0.18; 0.33].$$

Exponentiating the interval gives us the 95 % confidence interval for the odds ratio which is  $[0.84; 1.39]$ . This indicates that the odds of Laura's pizzas arriving earlier than Melissa's are not much different from one. While the point estimate tells us that Laura's pizzas are delivered 1.08 times faster, the confidence interval tells us that there is uncertainty around this estimate in the sense that it could also be smaller than 1 and Melissa may not necessarily work more slowly than Laura.

- (b) We can reproduce the results in *R* by attaching the pizza data, creating a categorical delivery time variable (using `cut`) and then applying the `oddsratio` command from the library `epitools` onto the contingency table:

```
attach(pizza)
timecat <- cut(time, breaks=c(-1,30,150))
library(epitools)
oddsratio(table(timecat,operator), method='wald')
```



## Chapter 10

*Solution to Exercise 10.1* A type I error is defined as the probability of rejecting  $H_0$  if  $H_0$  is true. This error occurs if  $A$  thinks that  $B$  does confess, but  $B$  does not. In this scenario,  $A$  confesses, goes free, and  $B$  serves a three-year sentence. A type II error is defined as the probability of accepting  $H_0$ , despite the fact that  $H_0$  is wrong. In this case,  $B$  does confess, but  $A$  does not. In this scenario,  $B$  goes free and  $A$  serves a three-year sentence. A type II error is therefore worse for  $A$ . With a statistical test, we always control the type I error, but not the type II error.

### *Solution to Exercise 10.2*

- (a) The hypotheses are

$$H_0 : \mu = 100 \text{ versus } H_1 : \mu \neq 100.$$

- (b) It is a one-sample problem for  $\mu$ : thus, for known variance, the Gauss test can be used; the  $t$ -test otherwise, see also Appendix D. Since, in this exercise,  $\sigma$  is assumed to be known we can use the Gauss test; i.e. we can compare the test statistic with the normal distribution (as opposed to the  $t$ -distribution when using the  $t$ -test). The sample size is small: we must therefore assume a normal distribution for the data.
- (c) To calculate the realized test statistic, we need the arithmetic mean  $\bar{x} = 98.08$ . Then we obtain

$$t(x) = \frac{\bar{x} - \mu}{\sigma} \cdot \sqrt{n} = \frac{98.08 - 100}{2} \cdot \sqrt{15} = \frac{-1.92}{2} \cdot \sqrt{15} = -3.72.$$

We reject  $H_0$  if  $|t(x)| > z_{1-\frac{\alpha}{2}} = 1.96$ . Since  $|t(x)| = 3.72 > 1.96$ , the null hypothesis can be rejected. The test is significant. There is enough evidence to suggest that the observed weights do not come from a normal distribution with mean 100 g.

- (d) To show that  $\mu < 100$ , we need to conduct a one-sided test. Most importantly, the research hypothesis of interest needs to be stated as the *alternative hypothesis*:

$$H_0 : \mu \geq 100 \quad \text{versus} \quad H_1 : \mu < 100.$$

- (e) The test statistic is the same as in (b):  $t(x) = -3.72$ . However, the critical region changes.  $H_0$  gets rejected if  $t(x) < -z_{1-\alpha} = -1.64$ . Again, the null hypothesis is rejected. The producer was right in hypothesizing that the average weight of his chocolate bars was lower than 100 g.

*Solution to Exercise 10.3*

- (a) We calculate

	No auction	Auction
$\bar{x}$	16.949	10.995
$s^2$	2.948	2.461
$s$	1.717	1.569
$v$	0.101	0.143

Note that we use the unbiased estimates for the variance and the standard deviation as explained in Chap. 9; i.e. we use  $1/(n - 1)$  rather than  $1/n$ . It is evident that the mean price of the auctions ( $\mu_a$ ) is lower than the mean non-auction price ( $\mu_{na}$ ), and also lower than the price from the online book store. There is, however, a higher variability in relation to the mean for the auction prices. One may speculate that the best offers are found in the auctions, but that there are no major differences between the online store and the internet book store, i.e.

- $\mu_{na} \neq €16.95$ ,
- $\mu_a < €16.95$ ,
- $\mu_{na} > \mu_a$ .

- (b) We can use the  $t$ -test to test the hypotheses

$$H_0 : \mu_{na} = 16.95 \quad \text{versus} \quad H_1 : \mu_{na} \neq 16.95.$$

The test statistic is

$$t(x) = \frac{\bar{x} - \mu}{\sigma} \cdot \sqrt{n} = \frac{16.949 - 16.95}{1.717} \cdot \sqrt{14} = -0.002.$$

Using the decision rules from Table 10.2, we conclude that  $H_0$  gets rejected if  $|t(x)| > t_{13,0.975} = 2.16$ . We can calculate  $t_{13,0.975}$  either by using Table C.2 or by using  $R(qt(0.975, 13))$ . Since  $|t(x)| = 0.002 \not> 2.16$ , we keep the null hypothesis. There is not enough evidence to suggest that the prices of the online store differ from €16.95 (which is the price from the internet book store).

- (c) Using (9.6) we can calculate the upper and lower limits of the confidence interval as,

$$\bar{x} + t_{n-1; 1-\alpha/2} \cdot \frac{s_X}{\sqrt{n}} = 16.949 + \underbrace{t_{13; 0.975}}_{=2.16} \cdot \frac{1.717}{\sqrt{14}} = 17.94$$

$$\bar{x} - t_{n-1; 1-\alpha/2} \cdot \frac{s_X}{\sqrt{n}} = 16.949 - \underbrace{t_{13; 0.975}}_{=2.16} \cdot \frac{1.717}{\sqrt{14}} = 15.96$$

respectively. The confidence interval *does* contain  $\mu_0 = 16.95$ ; hence, the null hypothesis  $\mu = \mu_0 = 16.95$  cannot be rejected. This is the same conclusion as obtained from the one-sample  $t$ -test from above.

- (d) We test the following hypotheses:

$$H_0 : \mu_a \geq 16.95 \quad \text{versus} \quad H_1 : \mu_a < 16.95.$$

The realized test statistic is

$$t(x) = \frac{\bar{x} - \mu}{\sigma} \cdot \sqrt{n} = \frac{10.995 - 16.95}{1.569} \cdot \sqrt{14} = -14.201.$$

Table 10.2 tells us that  $H_0$  gets rejected if  $t(x) < -t_{13, 0.95}$ . Since  $-14.201 < -2.16$  we reject  $H_0$ . The test confirms our hypothesis that the mean auction prices are lower than €16.95, i.e. the price from the book store.

- (e) We need to conduct two tests: the two-sample  $t$ -test for the assumption of equal variances and the Welch test for the assumption of different variances.

- (i) *Two-sample  $t$ -test:*

$$H_0 : \mu_{na} \leq \mu_a \quad \text{versus} \quad H_1 : \mu_{na} > \mu_a.$$

To calculate the test statistic (10.4), we need to determine the pooled sample variance:

$$\begin{aligned} s^2 &= \frac{(n_{na} - 1)s_{na}^2 + (n_a - 1)s_a^2}{n_{na} + n_a - 2} \\ &= \frac{(14 - 1)2.948 + (14 - 1)2.461}{14 + 14 - 2} \approx 2.705 \end{aligned}$$

The test statistic is

$$\begin{aligned} t(x) &= \frac{\bar{x}_{na} - \bar{x}_a}{s} \cdot \sqrt{\frac{n_{na} \cdot n_a}{n_{na} + n_a}} = \frac{16.949 - 10.995}{\sqrt{2.705}} \cdot \sqrt{\frac{14 \cdot 14}{14 + 14}} \\ &= \frac{5.954}{1.645} \cdot \sqrt{\frac{196}{28}} = 3.621 \cdot \sqrt{7} = 9.578. \end{aligned}$$

We reject  $H_0$  if  $t(x) > t_{n_{na} + n_a - 2, 0.95} = t_{26, 0.95} = 1.71$ . Table C.2 does not list the quantile; thus, one uses  $R(\text{qt}(0.95, 26))$  to determine the quantile. Since  $9.578 > 1.71$ , we reject the null hypothesis. The test is significant. There is enough evidence to conclude that the mean auction prices are lower than the mean non-auction prices.

- (ii) *Welch test:* For the Welch test, we calculate the test statistic, using (10.6) as:

$$t(x) = \frac{|\bar{x}_{na} - \bar{x}_a|}{\sqrt{\frac{s_{na}^2}{n_{na}} + \frac{s_a^2}{n_a}}} = \frac{|16.949 - 10.995|}{\sqrt{\frac{2.948}{14} + \frac{2.461}{14}}} = 9.578.$$

To calculate the critical value, we need the degrees of freedom:

$$\begin{aligned} v &= \left( \frac{s_{na}^2}{n_{na}} + \frac{s_a^2}{n_a} \right)^2 / \left( \frac{(s_{na}^2/n_{na})^2}{n_{na}-1} + \frac{(s_a^2/n_a)^2}{n_a-1} \right) \\ &= \left( \frac{2.948}{14} + \frac{2.461}{14} \right)^2 / \left( \frac{(2.948/14)^2}{13} + \frac{(2.461/14)^2}{13} \right) \\ &\approx 25.79 \end{aligned}$$

We reject  $H_0$  if  $t(x) > t_{v,0.95}$ . Using  $t_{v,0.95} = 1.706$  (in R obtained as  $qt(0.95, 25.79)$ ) and  $t(x) = 9.578$ , we reject the null hypothesis. The conclusions are identical to using the two-sample *t*-test (in this example). Interestingly, the two test statistics are similar which indicates that the assumption of equal variances in case (i) was not unreasonable.

- (f) The *F*-test relies on the assumption of the normal distribution and tests the hypotheses:

$$H_0 : \sigma_{na}^2 = \sigma_a^2 \quad \text{versus} \quad H_1 : \sigma_{na}^2 \neq \sigma_a^2.$$

We can calculate the test statistic as described in Appendix C:

$$t(x) = \frac{s_{na}^2}{s_a^2} = \frac{2.949}{2.461} = 1.198.$$

The *F*-distribution is not symmetric around 0; thus, we need to calculate two critical values:  $f_{n_1-1; n_2-1; 1-\alpha/2}$  and  $f_{n_1-1; n_2-1; \alpha/2}$ . Using R we get  $f_{13,13,0.975} = 3.115$  ( $\text{qf}(0.975, 13, 13)$ ) and  $f_{13,13,0.025} = \frac{1}{3.115} = 0.321$  ( $\text{qf}(0.025, 13, 13)$ ).  $H_0$  is rejected if  $t(x) > f_{n_1-1; n_2-1; 1-\alpha/2}$  or  $t(x) < f_{n_1-1; n_2-1; \alpha/2}$ . Since  $0.321 < 1.198 < 3.115$  we do not reject  $H_0$ . This means there is strong evidence that the two variances are equal. This is also reassuring with respect to using the two-sample *t*-test above (which assumes equal variances). However, note that testing the equality of variances with the *F*-test and then using the two-sample *t*-test or Welch test based on the outcome of the *F*-test is not ideal and not necessarily correct. In practice, it is best to simply use the Welch test (rather than the *t*-test), which is also the default option in the R function `t.t.test`.

(g) We need the following table to calculate the rank sums:

Value	9.3	9.52	9.54	10.01	10.5	10.5	10.55	10.59	11.02	11.03
Sample	a	a	a	a	a	a	a	a	a	a
Rank	1	2	3	4	5	6	7	8	9	10
Value	11.89	11.99	12	13.79	15.49	15.9	15.9	15.9	15.9	15.9
Sample	a	a	a	na	a	na	na	na	na	na
Rank	11	12	13	14	15	16	17	18	19	20
Value	15.99	16.98	16.98	17.72	18.19	18.19	19.97	19.97		
Sample	na									
Rank	21	22	23	24	25	26	27	28		

We can calculate the rank sums as  $R_{na+} = 13 + 15 + \dots + 28 = 300$  and  $R_{a+} = 1 + 2 + \dots + 13 + 15 = 106$ , respectively. Thus

$$U_1 = 14^2 + \frac{14 \cdot 15}{2} - 106 = 195; \quad U_2 = 14^2 + \frac{14 \cdot 15}{2} - 300 = 1.$$

With  $U = \min(195, 1) = 1$  we can calculate the test statistic, which is approximately normally distributed, as:

$$t(x, y) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} = \frac{1 - \frac{14^2}{2}}{\sqrt{\frac{14 \cdot 14 \cdot 29}{12}}} \approx -4.46$$

Since  $|t(x, y)| = 4.46 > z_{1-\alpha/2} = 1.96$ , the null hypothesis can be rejected. We conclude that the locations of the two distributions are shifted. Obviously the prices of the auction are smaller in general, and so is the median.

(h) We can type in the data and evaluate the summary statistics using the `mean`, `var`, and `sd` commands:

```
na <- c(18.19, 16.98, 19.97, ..., 17.72)
a <- c(10.5, 12.0, 9.54, ..., 11.02)
mean(na)
mean(a)
var(na)
...

```



The `t.test` command can be used to answer questions (b)–(e). For (b) and (c) we use

```
t.test(na, mu=16.95)
```



## One-sample t-test

```
data: na
t = -0.0031129, df = 13, p-value = 0.9976
alternative hypothesis: true mean is not equal to 16.95
95 percent confidence interval:
15.95714 17.94001
sample estimates:
mean of x
16.94857
```

The test decision can be made by means of either the *p*-value ( $= 0.9976 > 0.05$ ) or the confidence interval ([15.95; 17.94], which covers 16.95). To answer (d) and (e) we need to make use of the option `alternative` which specifies the alternative hypothesis:

```
t.test(a,mu=16.95,alternative='less')
t.test(na,a, alternative='greater')
```



Note that the two-sample test provides a confidence interval for the difference of the means. Questions (f) and (g) can be easily solved by using the `var.test` and `wilcox.test` commands:

```
var.test(na,a)
wilcox.test(na,a)
```



## F-test to compare two variances

```
data: na and a
F = 1.198, num df = 13, denom df = 13, p-value = 0.7496
alternative hypothesis: true ratio of variances not equal to 1
95 percent confidence interval:
0.3845785 3.7317371
sample estimates:
ratio of variances
1.197976
```

## Wilcoxon rank sum test with continuity correction

```
data: na and a
W = 195, p-value = 8.644e-06
alternative hypothesis: true location shift is not equal to 0
```

The test decision is best made by using the  $p$ -value.

*Solution to Exercise 10.4* Since the data before and after the diet is dependent (weight measured on the same subjects), we need to use the paired  $t$ -test. To calculate the test statistic, we need to first determine the weight differences:

Person $i$	1	2	3	4	5	6	7	8	9	10
Before diet	80	95	70	82	71	70	120	105	111	90
After diet	78	94	69	83	65	69	118	103	112	88
Differences $d$	2	1	1	-1	6	1	2	2	-1	2

Using  $\bar{d} = 1.5$  and

$$s_d^2 = \frac{1}{10-1} \cdot (0.5^2 + 0.5^2 + 0.5^2 + 2.5^2 + 4.5^2 + \dots + 0.5^2) = 3.83,$$

we calculate the test statistic as

$$t(d) = \frac{\bar{d}}{s_d} \sqrt{n} = \frac{1.5}{\sqrt{3.83}} \sqrt{10} = 2.42.$$

The null hypothesis is rejected because  $|t(d)| = 2.42 > t_{9;0.975} = 2.26$ . This means the mean weight before and after the diet is different. We would have obtained the same results by calculating the confidence interval for the differences:

$$\left[ \bar{d} \pm t_{n-1;1-\alpha/2} \frac{s_d}{\sqrt{n}} \right] \Leftrightarrow \left[ 1.5 \pm 2.26 \cdot \frac{\sqrt{3.83}}{\sqrt{10}} \right] = [0.1; 2.9].$$

Since the confidence interval does not overlap with zero, we reject the null hypothesis; there is enough evidence that the mean difference is different (i.e. greater) from zero. While the test is significant and suggests a weight difference, it is worth noting that the mean benefit from the diet is only 1.5 kg. Whether this is a relevant reduction in weight has to be decided by the ten people experimenting with the diet.

*Solution to Exercise 10.5*

- (a) The production is no longer profitable if the probability of finding a deficient shirt is greater than 10 %. This equates to the hypotheses:

$$H_0 : p \leq 0.1 \quad \text{versus} \quad H_1 : p > 0.1.$$

The sample proportion of deficient shirts is  $\hat{p} = \frac{35}{230} = \frac{7}{46}$ . Since  $np(1 - p) = 230 \cdot \frac{1}{10} \cdot \frac{9}{10} > 9$  we can use the test statistic

$$\begin{aligned} t(x) &= \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} = \frac{\frac{7}{46} - \frac{1}{10}}{\sqrt{\frac{1}{10} \cdot \frac{9}{10}}} \cdot \sqrt{230} \\ &= \frac{6}{115} \cdot \frac{10}{3} \cdot \sqrt{230} = \frac{4}{23} \cdot \sqrt{230} = 2.638. \end{aligned}$$

The null hypothesis gets rejected because  $t(x) = 2.638 > z_{0.95} = 1.64$ . It seems the production is no longer profitable.

- (b) The test statistic is  $t(x) = \sum x_i = 30$ . The critical region can be determined by calculating

$$P_{H_0}(Y \leq c_l) \leq 0.025 \quad \text{and} \quad P_{H_0}(Y \geq c_r) \leq 0.975.$$

Using *R* we get

```
qbinom(p=0.975,prob=0.1,size=230)
[1] 32
qbinom(p=0.025,prob=0.1,size=230)
[1] 14
```



The test statistic ( $t(x) = 30$ ) does not fall outside the critical region ([14; 32]); therefore, the null hypothesis is *not* rejected. The same result is obtained by using the binomial test in *R*: `binom.test(c(30, 200), p=0.1)`. This yields a *p*-value of 0.11239 and a confidence interval covering 10 % ([0.09; 0.18]). Interestingly, the approximate binomial test rejects the null hypothesis, whereas the exact test keeps it. Since the latter test is more precise, it is recommended to follow its outcome.

- (c) The research hypothesis is that the new machine produces fewer deficient shirts:

$$H_0 : p_{\text{new}} \geq p_{\text{old}} \quad \text{versus} \quad H_1 : p_{\text{new}} < p_{\text{old}}.$$

To calculate the test statistic, we need the following:

$$\begin{aligned} \hat{d} &= \frac{x_{\text{new}} - x_{\text{old}}}{n_{\text{new}} + n_{\text{old}}} = \frac{7}{115} - \frac{7}{46} = -\frac{21}{230} \\ \hat{p} &= \frac{x_{\text{new}} + x_{\text{old}}}{n_{\text{new}} + n_{\text{old}}} = \frac{7 + 35}{230 + 115} = \frac{42}{345} = \frac{14}{115}. \end{aligned}$$

This yields:

$$\begin{aligned} t(x_{\text{new}}, x_{\text{old}}) &= \frac{\hat{d}}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_{\text{new}}} + \frac{1}{n_{\text{old}}}\right)}} = \frac{-\frac{21}{230}}{\sqrt{\frac{14}{115} \cdot \frac{101}{115} \left(\frac{1}{115} + \frac{1}{230}\right)}} \\ &= \frac{-\frac{21}{230}}{\sqrt{0.1069 \cdot 0.013}} = -\frac{0.0913}{0.0373} = -2.448. \end{aligned}$$

The null hypothesis is rejected because  $t(x_{\text{new}}, x_{\text{old}}) = -2.448 < z_{0.05} = -z_{0.95} = -1.64$ . This means we accept the alternative hypothesis that the new machine is better than the old one.

- (d) The data can be summarized in the following table:

	Machine 1	Machine 2
Deficient	30	7
Fine	200	112

To test the hypotheses established in (c), we apply the `fisher.test` command onto the contingency table:

```
fisher.test(matrix(c(30,200,7,112),ncol=2))
```



This yields a  $p$ -value of 0.0438 suggesting, as the test in (c), that the null hypothesis should be rejected. Note that the confidence interval for the odds ratio, also reported by  $R$ , does not necessarily yield the same conclusion as the test of Fisher.

### Solution to Exercise 10.6

- (a) To compare the two means, we should use the Welch test (since the variances cannot be assumed to be equal). The test statistic is

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}}} = \frac{|103 - 101.8|}{\sqrt{\frac{12599.56}{10} + \frac{62.84}{10}}} \approx 0.0337.$$

The alternative hypothesis is  $H_1 : \mu_2 > \mu_1$ ; i.e. we deal with a one-sided hypothesis. The null hypothesis is rejected if  $t(x, y) > t_{v; 1-\alpha}$ . Calculating

$$\begin{aligned} v &= \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left( \frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1} \right) \\ &= \left( \frac{62.844}{10} + \frac{12599}{10} \right)^2 / \left( \frac{(62.844/10)^2}{9} + \frac{(12599/10)^2}{9} \right) \approx 9.09 \end{aligned}$$

yields  $t_{9.09; 0.95} = 1.831$  (using `qt(0.95, 9.09)` in  $R$ ; or looking at Table C.2 for 9 degrees of freedom). Therefore,  $t(x, y) < t_{9.09; 0.95}$  and the null hypothesis is not rejected. This means player 2 could not prove that he scores higher on average.

(b) We have to first rank the merged data:

Value	6	29	40	47	47	64	87	88	91	98
Sample	2	2	2	2	2	2	2	1	1	2
Rank	1	2	3	4	5	6	7	8	9	10
Value	99	99	101	104	105	108	111	112	261	351
Sample	1	1	1	1	1	1	1	1	2	2
Rank	11	12	13	14	15	16	17	18	19	20

This gives us  $R_{1+} = 8 + 9 + \dots + 18 = 133$ ,  $R_{2+} = 1 + 2 + \dots + 20 = 77$ ,  $U_1 = 10^2 + (10 \cdot 11)/2 - 133 = 22$ ,  $U_2 = 10^2 + (10 \cdot 11)/2 - 77 = 78$ , and therefore  $U = 22$ . The test statistic can thus be calculated as

$$t(x, y) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} = \frac{22 - \frac{10^2}{2}}{\sqrt{\frac{10 \cdot 10 \cdot 21}{12}}} \approx -2.12.$$

Since  $|t(x, y)| = 2.12 > z_{1-\alpha} = 1.64$ , the null hypothesis is rejected. The test supports the alternative hypothesis of higher points for player 1. The  $U$ -test has the advantage of not being focused on the expectation  $\mu$ . The two samples are clearly different: the second player scores with much more variability and his distribution of scores is clearly not symmetric and normally distributed. Since the sample is small, and the assumption of a normal distribution is likely not met, it makes sense to *not* use the  $t$ -test. Moreover, because the distribution is skewed the mean may not be a sensible measure of comparison. The two tests yield different conclusions in this example which shows that one needs to be careful in framing the right hypotheses. A drawback of the  $U$ -test is that it uses only ranks and not the raw data: it thus uses less information than the  $t$ -test which would be preferred when comparing means of a reasonably sized sample.

*Solution to Exercise 10.7* Otto speculates that the probability of finding a bear of colour  $i$  is 0.2, i.e.  $p_{\text{white}} = 0.2$ ,  $p_{\text{red}} = 0.2$ ,  $p_{\text{orange}} = 0.2$ ,  $p_{\text{yellow}} = 0.2$ , and  $p_{\text{green}} = 0.2$ . This hypothesis can be tested by using the  $\chi^2$  goodness-of-fit test. The test statistic is

$$\begin{aligned} t(x) = \chi^2 &= \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} = \frac{1}{250}(222 - 250)^2 + (279 - 250)^2 \\ &\quad + (251 - 250)^2 + (232 - 250)^2 + (266 - 250)^2 = 8.824. \end{aligned}$$

The null hypothesis cannot be rejected because  $t(x) = 8.824 \not> c_{4-1-0;0.95} = 9.49$ . While the small number of white bears might be disappointing, the test suggests that there is not enough evidence to reject the hypothesis of equal probabilities.

*Solution to Exercise 10.8*

- (a) To answer this question, we need to conduct the  $\chi^2$ -independence test. The test statistic  $t(x, y)$  is identical to Pearson's  $\chi^2$  statistic, introduced in Chap. 4. In Exercise 4.4, we have calculated this statistic already,  $\chi^2 \approx 182$ , see p. 350 for the details. The null hypothesis of independence is rejected if  $t(x, y) = \chi^2 > c_{(I-1)(J-1);1-\alpha}$ . With  $I = 2$  (number of rows), and  $J = 4$  (number of columns) we get  $c_{3;0.95} = 7.81$  using Table C.3 (or `qchisq(0.95, 3)` in R). Since  $182 > 7.81$ , we reject the null hypothesis of independence.
- (b) The output refers to a  $\chi^2$  test of homogeneity: the null hypothesis is that the proportion of passengers rescued is identical for the different travel classes. This hypothesis is rejected because  $p$  is smaller than  $\alpha = 0.05$ . It is evident that the proportions of rescued passengers in the first two classes (60 %, 43.9 %) are much higher than in the other classes (25 %, 23.8 %). One can see that the test statistic (182.06) is identical to (a). This is not surprising: the  $\chi^2$ -independence test and the  $\chi^2$  test of homogeneity are technically identical, but the null hypotheses differ. In (a), we showed that “rescue status” and “travel class” are independent; in (b), we have seen that the conditional distributions of rescue status given travel class differ by travel class, i.e. that the proportions of those rescued differ by the categories 1. class/2. class/3. class/staff.
- (c) The summarized table is as follows:

	1. Class/2. Class	3. Class/Staff	Total
Rescued	327	391	718
Not rescued	295	1215	1510
Total	622	1606	2228

Using (4.7) we get

$$t(x, y) = \chi^2 = \frac{2228(327 \cdot 1215 - 295 \cdot 391)^2}{718 \cdot 1510 \cdot 622 \cdot 1606} = 163.55.$$

The  $\chi^2$ -independence test and the  $\chi^2$  test of homogeneity are technically identical:  $H_0$  is rejected if  $t(x, y) > c_{(I-1)(J-1);1-\alpha}$ . Since  $163.55 > c_{1;0.95} = 3.84$  the null hypothesis is rejected. As highlighted in (b) this has two interpretations: (i) “rescue status” and “travel class” are independent (independence test) and (ii) the conditional distributions of rescue status given travel class differ by travel class (homogeneity test). The second interpretation implies that the probability of being rescued differs by travel class. The null hypothesis of the same probabilities of being rescued is also rejected using the test of Fisher. Summarizing the data in a matrix and applying the `fisher.test` command yields a  $p$ -value smaller than  $\alpha = 0.05$ .

```
fisher.test(matrix(c(327, 295, 391, 1215), ncol=2, nrow=2))
```



### *Solution to Exercise 10.9*

- (a) The hypotheses are:

$$H_0 : \mu_X = \mu_{Y1} \quad \text{versus} \quad H_1 : \mu_X \neq \mu_{Y1}.$$

The pooled variance,

$$s^2 = \frac{19 \cdot 2.94 + 19 \cdot 2.46}{39} = \frac{102.6}{39} = 2.631,$$

is needed to calculate the test statistic:

$$t(x, y) = \frac{4.97 - 4.55}{1.622} \cdot \sqrt{\frac{400}{40}} = \frac{0.42}{1.622} \cdot \sqrt{10} = 0.8188.$$

$H_0$  is rejected if  $|t(x, y)| > t_{38,0.975} \approx 2.02$ . Since  $|t(x, y)| = 0.8188$  we do not reject the null hypothesis.

- (b) The hypotheses are:

$$H_0 : \mu_X = \mu_{Y2} \quad \text{versus} \quad H_1 : \mu_X \neq \mu_{Y2}.$$

The pooled variance,

$$s^2 = \frac{19 \cdot 2.94 + 19 \cdot 3.44}{39} = 3.108,$$

is needed to calculate the test statistic:

$$t(x, y) = \frac{4.97 - 3.27}{1.763} \cdot \sqrt{10} = 3.049.$$

$H_0$  is rejected because  $|t(x, y)| = 3.049 > t_{38,0.975} \approx 2.02$ .

- (c) In both (a) and (b), there exists a true difference in the mean. However, only in (b) is the *t*-test able to detect the difference. This highlights that smaller differences can only be detected if the sample size is sufficiently large. However, if the sample size is very large, it may well be that the test detects a difference where there is no difference.

### *Solution to Exercise 10.10*

- (a) After reading in and attaching the data, we can simply use the `t.test` command to compare the expenditure of the two samples:

```
theatre <- read.csv('theatre.csv')
attach(theatre)
t.test(Culture[Sex==1], Culture[Sex==0])
```



## Welch Two-Sample t-test

```
data: Culture[Sex == 1] and Culture[Sex == 0]
t = -1.3018, df = 667.43, p-value = 0.1934
alternative hypothesis: true difference not equal to 0
95 percent confidence interval:
-12.841554 2.602222
sample estimates:
mean of x mean of y
217.5923 222.7120
```

We see that the null hypothesis is *not* rejected because  $p = 0.1934 > \alpha$  (also, the confidence interval overlaps with “0”).

- (b) A two-sample  $t$ -test and a  $U$ -test yield the same conclusion. The  $p$ -values, obtained with

```
wilcox.test(Culture[Sex==1],Culture[Sex==0])
t.test(Culture[Sex==1],Culture[Sex==0],var.equal=TRUE)
```



are 0.1946 and 0.145, respectively. Interestingly, the test statistic of the two-sample  $t$ -test is almost identical to the one from the Welch test in (a) ( $-1.2983$ )—this indicates that the assumption of equal variances may not be unreasonable.

- (c) We can use the usual `t.test` command together with the option `alternative = 'greater'` to get the solution.

```
t.test(Theatre[Sex==1],Theatre[Sex==0],
alternative='greater')
```



Here,  $p$  is much smaller than 0.001; hence, the null hypothesis can be rejected. Women spend more on theatre visits than men.

- (d) We deal with dependent (paired) data because different variables (expenditure this year versus expenditure last year) are measured on the same observations. Thus, we need to use the paired  $t$ -test—which we can use in *R* by specifying the `paired` option:

```
t.test(Theatre/Theatre_ly,paired=TRUE)
```



### Paired t-test

```

data: Theatre and Theatre_ly
t = 1.0925, df = 698, p-value = 0.275
alternative hypothesis: true difference in means is != 0
95 percent confidence interval:
-2.481496 8.707533
sample estimates:
mean of the differences
3.113019

```

Both the  $p$ -value (which is greater than 0.05) and the confidence interval (overlapping with “0”) state that the null hypothesis should be kept. The mean difference in expenditure (3.1 SFR) is not large enough to suggest that this difference is not caused by chance. Visitors spend, on average, about the same in the two years.

### *Solution to Exercise 10.11*

- (a) We can use the one-sample  $t$ -test to answer both questions:

```
t.test(temperature, mu=65, alternative='greater')
t.test(time, mu=30, alternative='less')
```



#### One-sample t-test

```

data: temperature
t = -11.006, df = 1265, p-value = 1
alternative hypothesis: true mean is greater than 65
...

data: time
t = 23.291, df = 1265, p-value = 1
alternative hypothesis: true mean is less than 30

```

We cannot confirm that the mean delivery time is less than 30 min and that the mean temperature is greater than 65 °C. This is not surprising: we have already seen in Exercise 3.10 that the manager should be unsatisfied with the performance of his company.

- (b) We can use the exact binomial test to investigate  $H_0 : p \geq 0.15$  and  $H_1 : p < 0.15$ . For the `binom.test` command, we need to know the numbers of successes and failures, i.e. the number of deliveries where a free wine should have been

given to the customer. Applying the `table` commands yields 229 and 1037 deliveries, respectively.

```
table(free_wine)
binom.test(c(229,1037),p=0.15,alternative='less')
```



```
Exact binomial test

data: c(229, 1037)
number of successes = 229, number of trials = 1266,
p-value = 0.9988 alternative hypothesis: true probability
is less than 0.15
95 percent confidence interval:
0.0000000 0.1996186
sample estimates, probability of success: 0.1808847
```

The null hypothesis cannot be rejected because  $p = 0.9988 > \alpha = 0.05$  and because the confidence interval covers  $p = 0.15$ . We get the same results if we use the variable “got\_wine” instead of “free\_wine”. While the test says that we cannot exclude the possibility that the probability of receiving a free wine is less than 15 %, the point estimate of 18 % suggests that the manager still has to improve the timeliness of the deliveries or stop the offer of free wine.

- (c) We first need to create a new categorical variable (using `cut`) which divides the temperatures into two parts: below and above 65 °C. Then we can simply apply the test commands (`fisher.test`, `chisq.test`, `prop.test`) to the table of branch and temperature:

```
hot <- cut(temperature,breaks=c(-Inf,65,Inf))
fisher.test(table(hot,operator))
chisq.test(table(hot,operator))
prop.test(table(hot,operator))
```



We know that the two  $\chi^2$  tests lead to identical results. For both of them the  $p$ -value is 0.2283 which suggests that we should keep the null hypothesis. There is not enough evidence which would support that the proportion of hot pizzas differs by operator, i.e. that the two variables are independent! The test of Fisher yields the same result ( $p = 0.2227$ ).

- (d) The null hypothesis is that the proportion of deliveries is the same for each branch:  $H_0 : p_{\text{East}} = p_{\text{West}} = p_{\text{Centre}}$ . To test this hypothesis, we need a  $\chi^2$  goodness-of-fit test:

```
chisq.test(table(branch))
```



Chi-squared test for given probabilities

```
data: table(branch)
X-squared = 0.74408, df = 2, p-value = 0.6893
```

The null hypothesis of equal proportions is therefore not rejected ( $p > \alpha = 0.05$ ).

- (e) To compare three proportions we need to use the  $\chi^2$  homogeneity test:

```
prop.test(table(branch, operator))
```



```
X-squared = 0.15719, df = 2, p-value = 0.9244
alternative hypothesis: two.sided
sample estimates:
prop 1    prop 2    prop 3
0.5059382 0.5097561 0.4965517
```

We can see that the proportions are almost identical and that the null hypothesis is not rejected ( $p = 0.9244$ ).

- (f) To test this relationship, we use the  $\chi^2$ -independence test:

```
chisq.test(table(driver, branch))
```



```
X-squared = 56.856, df = 8, p-value = 1.921e-09
```

The null hypothesis of independence is rejected.

*Solution to Exercise 10.12* To test the hypothesis  $p_{\text{Shalabh}} = p_{\text{Heumann}} = p_{\text{Schomaker}} = 1/3$  we use the  $\chi^2$  goodness-of-fit test. The test statistic is

$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} = \frac{1}{111}[(110 - 111)^2 + (118 - 111)^2 + (105 - 111)^2] \\ &= (1 + 49 + 36)/111 \approx 0.77\end{aligned}$$

$H_0$  gets rejected if  $\chi^2 = 0.77 > c_{3-1-0,1-0.01} = 9.21$ . Thus, the null hypothesis is not rejected. We accept that all three authors took about one-third of the pictures.

## Chapter 11

### Solution to Exercise 11.1

- (a) Calculating  $\bar{x} = \frac{1}{6}(26 + 23 + 27 + 28 + 24 + 25) = 25.5$  and  $\bar{y} = \frac{1}{6}(170 + 150 + 160 + 175 + 155 + 150) = 160$ , we obtain the following table needed for the estimation of  $\hat{\alpha}$  and  $\hat{\beta}$ :

Body mass index			Systolic blood pressure			$v_i$
$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	
26	0.5	0.25	170	10	100	5
23	-2.5	6.25	150	-10	100	25
27	1.5	2.25	160	0	0	0
28	2.5	6.25	175	15	225	37.5
24	-1.5	2.25	155	-5	25	7.5
25	-0.5	0.25	150	-10	100	5
Total	153	17.5	960	550	80	

With  $\sum_i v_i = \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 80$ , it follows that  $S_{xy} = 80$ . Moreover, we get  $S_{xx} = \sum_i (x_i - \bar{x})^2 = 17.5$  and  $S_{yy} = \sum_i (y_i - \bar{y})^2 = 550$ . The parameter estimates are therefore

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{80}{17.5} \approx 4.57,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 160 - 4.57 \cdot 25.5 = 43.465.$$

A one-unit increase in the BMI therefore relates to a 4.57 unit increase in the blood pressure. The model suggests a positive association between BMI and systolic blood pressure. It is impossible to have a BMI of 0; therefore,  $\hat{\alpha}$  cannot be interpreted meaningfully here.

- (b) Using (11.14), we obtain  $R^2$  as

$$R^2 = r^2 = \left( \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \right)^2 = \left( \frac{80}{\sqrt{17.5 \cdot 550}} \right)^2 \approx 0.66.$$

Thus 66 % of the data's variability can be explained by the model. The goodness of fit is good, but not perfect.

### Solution to Exercise 11.2

- (a) To estimate  $\hat{\beta}$ , we use the second equality from (11.6):

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}.$$

Calculating  $\bar{x} = 1.333$ ,  $\bar{y} = 5.556$ ,  $\sum_{i=1}^n x_i y_i = 62.96$ ,  $\sum_{i=1}^n x_i^2 = 24.24$ , and  $\sum_{i=1}^n y_i^2 = 281.5$  leads to

$$\begin{aligned}\hat{\beta} &= \frac{62.91 - 9 \cdot 1.333 \cdot 5.556}{24.24 - 9 \cdot 1.333^2} \approx \frac{-3.695}{8.248} \approx -0.45, \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 5.556 + 0.45 \cdot 1.333 \approx 6.16, \\ \hat{y}_i &= 6.16 - 0.45x_i.\end{aligned}$$

For children who spend no time on the internet at all, this model predicts 6.16 h of deep sleep. Each hour spent on the internet decreases the time in deep sleep by 0.45 h which is 27 min.

- (b) Using the results from (a) and  $S_{yy} = \sum_{i=1}^n y_i^2 - n\bar{y}^2 \approx 3.678$  yields:

$$R^2 = r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{(-3.695)^2}{8.248 \cdot 3.678} \approx 0.45.$$

About 45 % of the variation can be explained by the model. The fit of the model to the data is neither very good nor very bad.

- (c) After collecting the data in two vectors (`c()`), printing a summary of the linear model (`summary(lm())`) reproduces the results. A scatter plot can be produced by plotting the two vectors against each other (`plot()`). The regression line can be added with `abline()`:

```
it <- c(0.3,2.2,...,2.3)
sleep <- c(5.8,4.4,...,6.1)
summary(lm(sleep~it))
plot(it,sleep)
abline(a=6.16,b=-0.45)
```

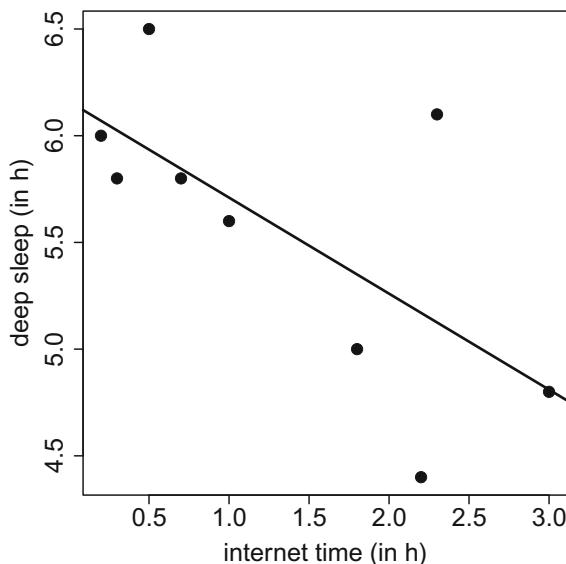


The plot is displayed in Fig. B.20.

- (d) Treating  $X$  as a binary variable yields the following values: 0, 1, 0, 0, 0, 1, 1, 0, 1. We therefore have  $\bar{x} = 0.444$ ,  $\sum x_i^2 = 4$ , and  $\sum x_i y_i = 4.4 + 5.0 + 4.8 + 6.1 = 20.3$ . Since the  $Y$ -values stay the same we calculate

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{20.3 - 9 \cdot 5.556 \cdot 0.444}{4 - 9 \cdot 0.444^2} \approx -0.85.$$

**Fig. B.20** Scatter plot and regression line for the association of internet use and deep sleep



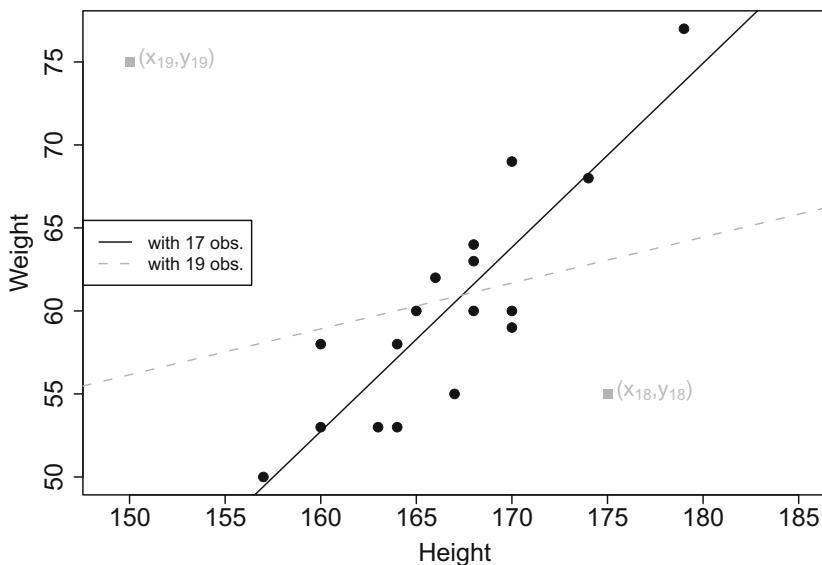
Thus, those children who are on the internet for a long time (i.e.  $> 1$  h) sleep on average 0.85 h ( $= 51$  min) less. If we change the coding of 1's and 0's,  $\hat{\beta}$  will just have a different sign:  $\hat{\beta} = 0.85$ . In this case, we can conclude that children who spend less time on the internet sleep on average 0.85 h longer than children who spend more time on the internet. This is the same conclusion and highlights that the choice of coding does not affect the interpretation of the results.

### Solution to Exercise 11.3

(a) The correlation coefficient is

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{yy} S_{xx}}} = \frac{170,821 - 17 \cdot 166.65 \cdot 60.12}{\sqrt{(62,184 - 17 \cdot 60.12^2)(472,569 - 17 \cdot 166.65^2)}} \\ &= \frac{498.03}{\sqrt{738.955 \cdot 441.22}} = 0.87. \end{aligned}$$

This indicates strong positive correlation: the higher the height, the higher the weight. Since  $R^2 = r^2 = 0.87^2 \approx 0.76$ , we already know that the fit of a linear regression model will be good (no matter whether height or weight is treated as outcome). From (11.11), we also know that  $\hat{\beta}$  will be positive.



**Fig. B.21** Scatter plot and regression line for both 17 and 19 observations

- (b) We know from (a) that  $S_{xy} = 498.03$  and that  $S_{xx} = 441.22$ . The least squares estimates are therefore

$$\hat{\beta} = \frac{498.03}{441.22} = 1.129,$$

$$\hat{\alpha} = 60.12 - 166.65 \cdot 1.129 = -128.03.$$

Each centimetre difference in height therefore means a 1.129 kg difference in weight. It is not possible to interpret  $\hat{\alpha}$  meaningfully in this example.

- (c) The prediction is

$$-128.03 + 1.129 \cdot 175 = 69.545 \text{ kg}.$$

- (d)–(g) The black dots in Fig. B.21 show the scatter plot of the data. There is clearly a positive association in that greater height implies greater weight. This is also emphasized by the regression line estimated in (b). The two additional points appear in dark grey in the plot. It is obvious that they do not match the pattern observed in the original 17 data points. One may therefore speculate that with the inclusion of the two new points  $\hat{\beta}$  will be smaller. To estimate the new regression line we need

$$\bar{x} = \frac{1}{19}(17 \cdot 166.65 + 150 + 175) = 166.21,$$

$$\bar{y} = \frac{1}{19}(17 \cdot 60.12 + 75 + 55) = 60.63.$$

This yields

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{191696 - 19 \cdot 166.21 \cdot 60.63}{525694 - 19 \cdot 166.21^2} = \frac{227.0663}{804.4821} \approx 0.28.$$

This shows that the two added points shrink the estimate from 1.129 to 0.28. The association becomes less clear. This is an insightful example showing that least squares estimates are generally *sensitive to outliers* which can potentially affect the results.

### *Solution to Exercise 11.4*

- (a) The point estimate of  $\beta$  suggests a 0.077 % increase of hotel occupation for each one degree increase in temperature. However, the null hypothesis of  $\beta = 0$  cannot be rejected because  $p = 0.883 > 0.05$ . We therefore cannot show an association between temperature and hotel occupation.
- (b) The average hotel occupation is higher in Davos (7.9 %) and Polenca (0.9 %) compared with Basel (reference category). However, these differences are not significant. Both  $H_0 : \beta_{\text{Davos}} = 0$  and  $H_0 : \beta_{\text{Polenca}} = 0$  cannot be rejected. The model cannot show a significant difference in hotel occupation between Davos/Polenca and Basel.
- (c) The analysis of variance table tells us that the null hypothesis of equal average temperatures in the three cities ( $\beta_1 = \beta_2 = 0$ ) cannot be rejected. Note that in this example the overall  $F$ -test would have given us the same results.
- (d) In the multivariate model, the main conclusions of (a) and (b) do not change: testing  $H_0 : \beta_j = 0$  never leads to the rejection of the null hypothesis. We cannot show an association between temperature and hotel occupation (given the city); and we cannot show an association between city and hotel occupation (given the temperature).
- (e) Stratifying the data yields considerably different results compared to (a)–(c): In Davos, where tourists go for skiing, each increase of 1 °C relates to a drop in hotel occupation of 2.7 %. The estimate  $\hat{\beta} \approx -2.7$  is also significantly different from zero ( $p = 0.000231$ ). In Polenca, a summer holiday destination, an increase of 1 °C implies an increase of hotel occupation of almost 4 %. This estimate is also significantly different from zero ( $p = 0.00114 < 0.05$ ). In Basel, a business destination, there is a somewhat higher hotel occupation for higher temperatures ( $\hat{\beta} = 1.3$ ); however, the estimate is not significantly different from zero. While there is no overall association between temperature and hotel occupation (see (a) and (c)), there is an association between them if one looks at the different cities separately. This suggests that an interaction between temperature and city should be included in the model.

- (f) The design matrix contains a column of 1's (to model the intercept), the temperature and two dummies for the categorical variable “city” because it has three categories. The matrix also contains the interaction terms which are both the product of temperature and Davos and temperature and Polenca. The matrix has 36 rows because there are 36 observations: 12 for each city.

	Int.	Temp.	Davos	Polenca	Temp. × Davos	Temp. × Polenca
1	1	-6	1	0	-6	0
2	1	-5	1	0	-5	0
:	:	:	:	:	:	:
12	1	0	1	0	0	0
13	1	10	0	1	0	10
:	:	:	:	:	:	:
24	1	12	0	1	0	12
25	1	1	0	0	0	0
:	:	:	:	:	:	:
36	1	4	0	0	0	0

- (g) Both interaction terms are significantly different from zero ( $p = 0.000375$  and  $p = 0.033388$ ). The estimate of temperature therefore differs by city, and the estimate of city differs by temperature. For the reference city of Basel, the association between temperature and hotel occupation is estimated as 1.31; for Davos it is  $1.31 - 4.00 = -2.69$  and for Polenca  $1.31 + 2.66 = 3.97$ . Note that these results are identical to (d) where we fitted three different regressions—they are just summarized in a different way.
- (h) From (f) it follows that the point estimates for  $\beta_{\text{temperature}}$  are 1.31 for Basel, -2.69 for Davos, and 3.97 for Polenca. Confidence intervals for these estimates can be obtained via (11.29):

$$(\hat{\beta}_i + \hat{\beta}_j) \pm t_{n-p-1; 1-\alpha/2} \cdot \hat{\sigma}_{(\hat{\beta}_i + \hat{\beta}_j)}.$$

We calculate  $t_{n-p-1; 1-\alpha/2} = t_{36-5-1, 0.975} = t_{30, 0.975} = 2.04$ . With  $\text{Var}(\hat{\beta}_{\text{temp}}) = 0.478$  (obtained via 0.6916<sup>2</sup> from the model output or from the second row and second column of the covariance matrix),  $\text{Var}(\hat{\beta}_{\text{temp:Davos}}) = 0.997$ ,  $\text{Var}(\hat{\beta}_{\text{Polenca}}) = 1.43$ ,  $\text{Cov}(\hat{\beta}_{\text{temp}}, \hat{\beta}_{\text{temp:Davos}}) = -0.48$ , and also  $\text{Cov}(\hat{\beta}_{\text{temp}}, \hat{\beta}_{\text{temp:Polenca}}) = -0.48$  we obtain:

$$\begin{aligned}\hat{\sigma}_{(\hat{\beta}_{\text{temp}} + \hat{\beta}_{\text{Davos}})} &= \sqrt{0.478 + 0.997 - 2 \cdot 0.48} \approx 0.72, \\ \hat{\sigma}_{(\hat{\beta}_{\text{temp}} + \hat{\beta}_{\text{Polenca}})} &= \sqrt{0.478 + 1.43 - 2 \cdot 0.48} \approx 0.97, \\ \hat{\sigma}_{(\hat{\beta}_{\text{temp}} + \hat{\beta}_{\text{Basel}})} &= \sqrt{0.478 + 0 + 0} \approx 0.69.\end{aligned}$$

The 95 % confidence intervals are therefore:

$$\begin{aligned} \text{Davos: } & [-2.69 \pm 2.04 \cdot 0.72] \approx [-4.2; -1.2], \\ \text{Polenca: } & [3.97 \pm 2.04 \cdot 0.97] \approx [2.0; 5.9], \\ \text{Basel: } & [1.31 \pm 2.04 \cdot 0.69] \approx [-0.1; 2.7]. \end{aligned}$$

### *Solution to Exercise 11.5*

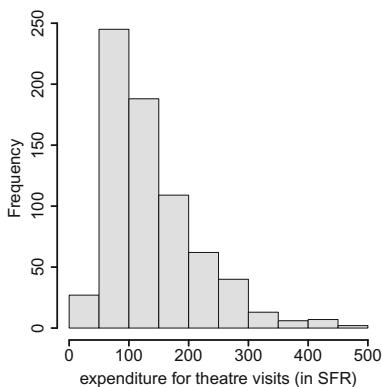
- (a) The missing value [1] can be calculated as

$$T = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} = \frac{0.39757 - 0}{0.19689} = 2.019.$$

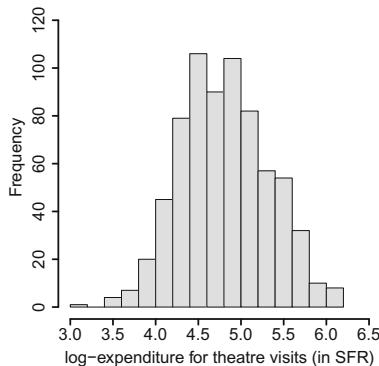
Since  $t_{699-5-1,0.975} = 1.96$  and  $2.019 > 1.96$ , it is clear that the  $p$ -value from [2] is smaller than 0.05. The exact  $p$ -value can be calculated in R via  $(1-\text{pt}(2.019, 693)) * 2$  which yields 0.0439. The `pt` command gives the probability value for the quantile of 2.019 (with 693 degrees of freedom): 0.978. Therefore, with probability  $(1 - 0.978) \%$  a value is right of 2.019 in the respective  $t$ -distribution which gives, multiplied by two to account for a two-sided test, the  $p$ -value.

- (b)–(c) The plot on the left shows that the residuals are certainly not normally distributed as required by the model assumptions. The dots do not approximately match the bisecting line. There are too many high positive residuals which means that we are likely dealing with a right-skewed distribution of residuals. The plot on the right looks alright: no systematic pattern can be seen; it is a random plot. The histogram of both theatre expenditure and `log(theatre_expenditure)` suggests that a log-transformation may improve the model, see Fig. B.22. Log-transformations are often helpful when the outcome's distribution is skewed to the right.
- (d) Since the outcome is log-transformed, we can apply the interpretation of a log-linear model:

- Each year's increase in age yields an  $\exp(0.0038) = 1.0038$  times higher ( $=0.38 \%$ ) expenditure on theatre visits. Therefore, a 10-year age difference relates to an  $\exp(10 \cdot 0.0038) = 1.038$  times higher expenditure ( $=3.8 \%$ ).
- Women (gender = 1) spend on average (given the other variables)  $\exp(0.179) \approx 1.20$  times more money on theatre visits.
- Each 1000 SFR more yearly income relates to an  $\exp(0.0088) = 1.0088$  times higher expenditure on theatre visits. A difference in 10,000 SFR per year therefore amounts to an 8.8 % difference in expenditure.
- Each extra Swiss Franc spent on cultural activities is associated with an  $\exp(0.00353) = 1.0035$  times higher expenditure on theatre visits.



(a) Expenditure



(b) log(expenditure)

**Fig. B.22** Histogram of the outcome

- Except for theatre expenditure from the preceding year, all  $\beta_j$  are significantly different from zero.
- (e) While in (b) the residuals were clearly not normally distributed, this assumption seems to be fulfilled now: the QQ-plot shows dots which lie approximately on the bisecting line. The fitted values versus residuals plot remains a chaos plot. In conclusion, the log-transformation of the outcome helped to improve the quality of the model.

*Solution to Exercise 11.6*

- (a) The multivariate model is obtained by using the `lm()` command and separating the covariates with the `+` sign. Applying `summary()` to the model returns the comprehensive summary.

```
mp <- lm(time ~ temperature + branch + day + operator + driver
+ bill + pizzas + discount_customer)
summary(mp)
```



```

Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.42270 2.00446 20.166 < 2e-16 ***
temperature -0.20823 0.02594 -8.027 2.28e-15 ***
branchEast -1.60263 0.42331 -3.786 0.000160 ***
branchWest -0.11912 0.37330 -0.319 0.749708
dayMonday -1.15858 0.63300 -1.830 0.067443 .
daySaturday 0.88163 0.50161 1.758 0.079061 .
daySunday 1.01655 0.56103 1.812 0.070238 .
dayThursday 0.78895 0.53006 1.488 0.136895
dayTuesday 0.79284 0.62538 1.268 0.205117
dayWednesday 0.25814 0.60651 0.426 0.670468
operatorMelissa -0.15791 0.34311 -0.460 0.645435
driverDomenico -2.59296 0.73434 -3.531 0.000429 ***
driverLuigi -0.80863 0.58724 -1.377 0.168760
driverMario -0.39501 0.43678 -0.904 0.365973
driverSalvatore -0.50410 0.43480 -1.159 0.246519
bill 0.14102 0.01600 8.811 < 2e-16 ***
pizzas 0.55618 0.11718 4.746 2.31e-06 ***
discount_customer -0.28321 0.36848 -0.769 0.442291
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1     1

Residual standard error: 5.373 on 1248 degrees of freedom
Multiple R-squared: 0.3178,   Adjusted R-squared: 0.3085
F-statistic: 34.2 on 17 and 1248 DF, p-value: < 2.2e-16

```

The output shows that lower temperature, higher bills, and more ordered pizzas increase the delivery times. The branch in the East is the fastest, and so is the driver Domenico. While there are differences with respect to day, discount customers, and the operator, they are not significant at the 5 % level.

- (b) The confidence intervals are calculated as:  $\hat{\beta}_i \pm t_{n-p-1; 1-\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}}$ . We know from the model output from (a) that there are 1248 degrees of freedom (1266 observations – 18 estimated coefficients). The respective quantile from the  $t$ -distribution is obtained with the `qt()` function. The coefficients are accessed via the `coefficients` command (alternatively: `mp$coefficients`); the variances of the coefficients are either accessed via the diagonal elements of the covariance matrix (`diag(vcov(mp))`) or the model summary (`summary(mp)[[4]][,2]`)—both of which are laborious. The summary of coefficients, lower confidence limit (lcl), and upper confidence limit (ucl) may be summarized in a matrix, e.g. via merging the individual columns with the `cbind` command.

```
lcl <- coefficients(mp) - qt(0.975,1248)*sqrt(diag(vcov(mp)))
ucl <- coefficients(mp) + qt(0.975,1248)*sqrt(diag(vcov(mp)))
cbind(coefficients(mp),lcl,ucl)
```



	lcl	ucl
(Intercept)	40.4227014	36.4902223
temperature	-0.2082256	-0.2591146
branchEast	-1.6026299	-2.4331162
branchWest	-0.1191190	-0.8514880
dayMonday	-1.1585828	-2.4004457
...		

- (c) The variance is estimated as the residual sum of squares divided by the degrees of freedom, see also (11.27). Applying the `residuals` command to the model and using other basic operations yields an estimated variance of 28.86936.

```
sum(residuals(mp)^2)/(mp$df.residual)
```



Taking the square root of the result yields  $\sqrt{28.86936} = 5.37$  which is also reported in the model output from (a) under “Residual standard error”.

- (d) The sum of squares error is defined as  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ . The total sum of squares is  $\sum_{i=1}^n (y_i - \bar{y})^2$ . This can be easily calculated in R. The goodness of fit is then obtained as  $R^2 = 1 - SQ_{\text{Error}}/SQ_{\text{Total}} = 0.3178$ . Dividing  $SQ_{\text{Error}}$  by  $n - p - 1 = 1266 - 17 - 1 = 1248$  and  $SQ_{\text{Total}}$  by  $n - 1 = 1265$  yields  $R^2_{\text{adj}} = 0.3085$ . This corresponds to the model output from (a).

```
SQE <- sum(residuals(mp)^2)
SQT <- sum((time-mean(time))^2)
1-(SQE/SQT)
1-((SQE/1248)/(SQT/1265))
```



- (e) Applying `stepAIC` to the fitted model (with option “back” for backward selection) executes the model selection by means of AIC.

```
library(MASS)
stepAIC(mp, direction='back')
```



The output shows that the full model has an AIC of 4275.15. The smallest AIC is achieved by removing the operator variable from the model.

```
Step: AIC=4275.15
time ~ temperature + branch + day + operator + driver + bill +
      pizzas + discount_customer

          Df Sum of Sq   RSS   AIC
- operator      1     6.11 36035 4273.4
- discount_customer  1    17.05 36046 4273.8
<none>                   36029 4275.2
- day            6    448.79 36478 4278.8
- driver         4    363.91 36393 4279.9
- branch         2    511.10 36540 4289.0
- pizzas         1    650.39 36679 4295.8
- temperature    1   1860.36 37889 4336.9
- bill           1   2241.30 38270 4349.6
```

The reduced model has an AIC of 4273.37. Removing the discount customer variable from the model yields an improved AIC ( $4272.0 < 4273.37$ ).

```
Step: AIC=4273.37
time ~ temperature + branch + day + driver + bill + pizzas +
      discount_customer

          Df Sum of Sq   RSS   AIC
- discount_customer  1    17.57 36053 4272.0
<none>                   36035 4273.4
- day            6    452.00 36487 4277.1
- driver         4    364.61 36400 4278.1
- branch         2    508.57 36544 4287.1
- pizzas         1    649.54 36685 4294.0
- temperature    1   1869.98 37905 4335.4
- bill           1   2236.19 38271 4347.6
```

The model selection procedure stops here as removing any variable would only increase the AIC, not decrease it.

```
Step: AIC=4271.98
time ~ temperature + branch + day + driver + bill + pizzas

          Df Sum of Sq   RSS   AIC
<none>                   36053 4272.0
- day            6    455.62 36508 4275.9
- driver         4    368.18 36421 4276.8
- branch         2    513.17 36566 4285.9
- pizzas         1    657.07 36710 4292.8
- temperature    1   1878.24 37931 4334.3
- bill           1   2228.88 38282 4345.9
```

The final model, based on backward selection with AIC, includes day, driver, branch, number of pizzas ordered, temperature, and bill.

- (f) Fitting the linear model with the variables obtained from (e) and obtaining the summary of it yields an  $R^2_{\text{adj}}$  of 0.3092.

```
mps <- lm(time ~ temperature + branch + day + driver + bill +
pizzas)
summary(mps)
```



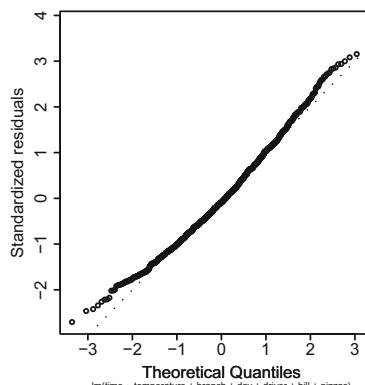
This is only marginally higher than the goodness of fit from the full model ( $0.3092 > 0.3085$ ). While the selected model is better than the model with all variables, both, with respect to AIC and  $R^2_{\text{adj}}$ , the results are very close and remind us of the possible instability of applying automated model selection procedures.

- (g) Both the normality assumption and heteroscedasticity can be checked by applying `plot()` to the model. From the many graphs provided we concentrate on the second and third of them:

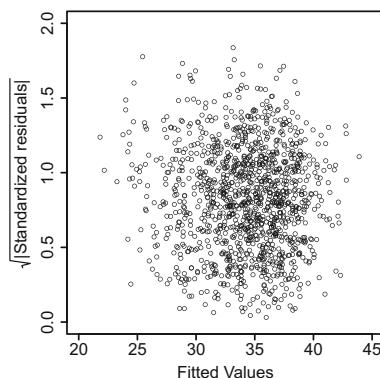
```
plot(mps, which=2)
plot(mps, which=3)
```



Figure B.23a shows that the residuals are approximately normally distributed because the dots lie approximately on the bisecting line. There are some smaller deviations at the tails but they are not severe. The plot of the fitted values versus



(a) QQ-plot



(b) Fitted values vs. residuals

**Fig. B.23** Checking the model assumptions

the square root of the absolute values of the residuals shows no pattern; it is a random plot (Fig. B.23b). There seems to be no heteroscedasticity.

- (h) Not all variables identified in (e) represent necessarily a “cause” for delayed or improved delivery time. It makes sense to speculate that *because* many pizzas are being delivered (and need to be made!) the delivery time increases. There might also be reasons why a certain driver is improving the delivery time: maybe he does not care about red lights. This could be investigated further given the results of the model above. However, high temperature does not *cause* the delivery time to be shorter; likely it is the other way around: the temperature is hotter because the delivery time is shorter. However, all of these considerations remain speculation. A regression model only exhibits associations. If there is a significant association, we know that given an accepted error (e.g. 5 %), values of  $\mathbf{x}$  are higher when values of  $\mathbf{y}$  are higher. This is useful but it does not say whether  $\mathbf{x}$  caused  $\mathbf{y}$  or vice versa.
- (i) To check whether it is worth to add a polynomial, we simply add the squared temperature to the model. To make  $R$  understand that we apply a transformation we need to use  $I()$ .

```
mps2 <- lm(time ~ temperature + I(temperature^2) +
I(temperature^3) + branch + day + driver + bill + pizzas)
summary(mps2)
```



	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-18.954965	8.795301	-2.155	0.03134 *
temperature	1.736692	0.282453	6.149	1.05e-09 ***
I(temperature^2)	-0.015544	0.002247	-6.917	7.36e-12 ***
branchEast	-1.429772	0.416107	-3.436	0.00061 ***
...				

It can be seen that the null hypothesis  $H_0 : \beta_{\text{temp}^2} = 0$  is rejected. This indicates that it is worthwhile to assume a quadratic relationship between temperature and delivery time.

- (j) The prediction can be obtained by the `predict` command as follows:

```
predict(mps,pizza[1266,])
```



The prediction is 36.5 min and therefore 0.8 min higher than the real delivery time.

## More details on Chap. 3

*Proof of equation (3.27).*

$$\begin{aligned}
 \tilde{s}^2 &= \frac{1}{n} \sum_{j=1}^k \sum_{x_i \in K_j} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{j=1}^k \sum_{x_i \in K_j} (x_i - \bar{x}_j + \bar{x}_j - \bar{x})^2 \\
 &= \underbrace{\frac{1}{n} \sum_{j=1}^k \sum_{x_i \in K_j} (x_i - \bar{x}_j)^2}_{[i]} + \underbrace{\frac{1}{n} \sum_{j=1}^k \sum_{x_i \in K_j} (\bar{x}_j - \bar{x})^2}_{[ii]} \\
 &\quad + \underbrace{\frac{2}{n} \sum_{j=1}^k \sum_{x_i \in K_j} (x_i - \bar{x}_j)(\bar{x}_j - \bar{x})}_{[iii]}
 \end{aligned}$$

We obtain the following expressions for [i]–[iii]:

$$\begin{aligned}
 [i] &= \frac{1}{n} \sum_{j=1}^k n_j \frac{1}{n_j} \sum_{x_i \in K_j} (x_i - \bar{x}_j)^2 = \frac{1}{n} \sum_{j=1}^k n_j \tilde{s}_j^2, \\
 [ii] &= \frac{1}{n} \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2, \\
 [iii] &= \frac{2}{n} \sum_{j=1}^k (\bar{x}_j - \bar{x}) \sum_{x_i \in K_j} (x_i - \bar{x}_j) = \frac{2}{n} \sum_{j=1}^k (\bar{x}_j - \bar{x}) 0 = 0.
 \end{aligned}$$

Since [i] is the within-class variance and [ii] is the between-class variance, Eq. (3.27) holds.

## More details on Chap. 7

*Proof of Theorem 7.2.3.* Consider the interval  $(x_0 - \delta, x_0]$  with  $\delta \geq 0$ . From (7.12) it follows that  $P(x_0 - \delta < X \leq x_0) = F(x_0) - F(x_0 - \delta)$  and therefore

$$\begin{aligned} P(X = x_0) &= \lim_{\delta \rightarrow 0} P(x_0 - \delta < X \leq x_0) \\ &= \lim_{\delta \rightarrow 0} [F(x_0) - F(x_0 - \delta)] \\ &= F(x_0) - F(x_0) = 0. \end{aligned}$$

*Proof of Theorem 7.3.1.*

$$\begin{aligned} \text{Var}(X) &\stackrel{(7.17)}{=} E(X - \mu)^2 \\ &= E(X^2 - 2\mu X + \mu^2) \\ &\stackrel{(7.28-7.31)}{=} E(X^2) - 2\mu E(X) + E(\mu^2) \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - [E(X)]^2. \end{aligned}$$

*Proof of Theorem 7.4.1.* We define a discrete random variable  $Y$  as

$$Y = \begin{cases} 0 & \text{if } |X - \mu| < c \\ c^2 & \text{if } |X - \mu| \geq c. \end{cases} \quad (\text{C.1})$$

The respective probabilities are  $P(|X - \mu| < c) = p_1$  and  $P(|X - \mu| \geq c) = p_2$ . The definition of  $Y$  in (C.1) implies that

$$Y \leq |X - \mu|^2$$

since for  $|X - \mu|^2 < c^2$   $Y$  takes the value  $y_1 = 0$  and therefore  $Y \leq |X - \mu|^2$ . If  $|X - \mu|^2 \geq c^2$   $Y$  takes the value  $y_2 = c^2$ , and therefore  $Y \leq |X - \mu|^2$ . Using this knowledge, we obtain

$$E(Y) \leq E(|X - \mu|^2) = \text{Var}(X).$$

However, for  $Y$  we also have

$$E(Y) = 0 \cdot p_1 + c^2 \cdot p_2 = c^2 P(|X - \mu| \geq c)$$

which means that we can summarize the above findings in the following inequality:

$$c^2 P(|X - \mu| \geq c) \leq \text{Var}(X).$$

This equates to Tschebyschev's inequality. Using  $P(\bar{A}) = 1 - P(A)$ , i.e.

$$P(|X - \mu| \geq c) = 1 - P(|X - \mu| < c),$$

we obtain the second formula of Tschebyschev's inequality:

$$P(|X - \mu| < c) \geq 1 - \frac{\text{Var}(X)}{c^2}.$$

*Proof of rule (7.30).* For discrete variables, we have

$$\mathbb{E}(a + bX) \stackrel{(7.16)}{=} \sum_i (a + bx_i)p_i = \sum_i ap_i + bx_i p_i = a \sum_i p_i + b \sum_i x_i p_i.$$

Since  $\sum_i p_i = 1$  and  $\sum_i x_i p_i = \mathbb{E}(X)$ , we obtain  $\mathbb{E}(a + bX) = a + b\mathbb{E}(X)$ , which is rule (7.30). In the continuous case, we have

$$\begin{aligned} \mathbb{E}(a + bX) &= \int_{-\infty}^{\infty} (a + bx)f(x)dx = \int_{-\infty}^{\infty} (af(x)dx + bx f(x)dx) \\ &= a \int_{-\infty}^{\infty} f(x)dx + b \int_{-\infty}^{\infty} xf(x)dx = a + b\mathbb{E}(X). \end{aligned}$$

*Proof of rule (7.33).* Using  $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ , we can write the variance of  $bX$  as

$$\text{Var}(bX) = \mathbb{E}([bX]^2) + \mathbb{E}(bX)^2.$$

Using (7.29), we get  $\mathbb{E}([bX]^2) = b^2\mathbb{E}(X^2)$  and  $\mathbb{E}(bX)^2 = (b\mathbb{E}(X))^2$ . Therefore

$$\text{Var}(bX) = b^2(\mathbb{E}(X^2) - \mathbb{E}(X)^2) = b^2\text{Var}(X).$$

*Proof of  $\rho = 1$  for a perfect linear relationship.* If  $Y = aX + b$  with  $a \neq 0$ , we get

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= a\mathbb{E}[(X - \mu_X)(X - \mu_X)] \end{aligned}$$

because  $\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$ , see p. 147. Then

$$\begin{aligned} \text{Cov}(X, Y) &= a\text{Var}(X), \\ \text{Var}(Y) &\stackrel{(7.33)}{=} a^2\text{Var}(X), \end{aligned}$$

and therefore

$$\rho(X, Y) = \frac{a\text{Var}(X)}{\sqrt{a^2\text{Var}(X)\text{Var}(X)}} = \frac{a}{|a|} = 1$$

if  $a > 0$ . Similarly, if  $Y = aX + b$  with  $a < 0$  we get  $\rho(X, Y) = -1$ .

## More details on Chap. 8

**Theorem of Large Numbers.** To explain the Theorem of Large Numbers, we first need to first define **stochastic convergence**.

**Definition C.1** A sequence of random variables,  $(X_n)_{n \in \mathbb{N}}$ , converges stochastically to 0, if for any  $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n| > \epsilon) = 0 \tag{C.2}$$

holds.

This is equivalent to  $\lim_{n \rightarrow \infty} P(|X_n| \leq \epsilon) = 1$ .

**Theorem C.1** (Theorem of large numbers) *Consider  $n$  i.i.d. random variables  $X_1, X_2, \dots, X_n$  with  $E(X_i) = \mu$ ,  $\text{Var}(X_i) = \sigma^2$  and  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . It holds that*

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < c) = 1, \quad \forall c \geq 0. \quad (\text{C.3})$$

This implies that  $\bar{X}_n$  converges stochastically to  $\mu$ . As another motivation, recall Definition 7.6.1 where we define random variables  $X_1, X_2, \dots, X_n$  to be i.i.d. (independently identically distributed) if all  $X_i$  follow the same distribution and are independent of each other. Under this assumption, we showed in (7.36) that  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ . It follows that the larger  $n$ , the smaller the variance. If we apply Tschebyschev's inequality (Theorem 7.4.1) to  $\bar{X}$ , we get the following equation for  $(\bar{X}_n - \mu)_{n \in \mathbb{N}}$ :

$$P(|\bar{X}_n - \mu| < c) \geq 1 - \frac{\text{Var}(\bar{X}_n)}{c^2} = 1 - \frac{\sigma^2}{nc^2}. \quad (\text{C.4})$$

This means that for each  $c \geq 0$ , the right-hand side of the above equation tends to 1 as  $n \rightarrow \infty$  which gives a similar interpretation as the Theorem of Large Numbers.

**Central Limit Theorem.** Let  $X_i$  ( $i = 1, 2, \dots, n$ ) be  $n$  i.i.d. random variables with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$ . If we consider the sum  $\sum_{i=1}^n X_i$ , we obtain  $E(\sum_{i=1}^n X_i) = n\mu$  and  $\text{Var}(\sum_{i=1}^n X_i) = n\sigma^2$ . If we want to standardize  $\sum_{i=1}^n X_i$  we can use Theorem 7.3.2 to obtain

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}}, \quad (\text{C.5})$$

i.e. it holds that  $E(Y_n) = 0$  and  $\text{Var}(Y_n) = 1$ .

**Theorem C.2** (Central Limit Theorem) *Let  $X_i$  ( $i = 1, 2, \dots, n$ ) be  $n$  i.i.d. random variables with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$ .  $Y_n$  denotes the standardized sum of  $X_i$ ,  $i = 1, 2, \dots, n$ . The CDF of  $Y_n$  is*

$$\lim_{n \rightarrow \infty} P(Y_n \leq y) = \Phi(y), \quad \forall y,$$

where  $\Phi(y)$  denotes the CDF of the standard normal distribution.

This theorem tells us that  $Y_n$  is approximately standard normal distributed if  $n$  is large, i.e.

$$Y_n \sim N(0, 1) \quad \text{as } n \rightarrow \infty.$$

This is equivalent to saying that  $\sum_{i=1}^n X_i$  is approximately  $N(n\mu, n\sigma^2)$  distributed if  $n$  is large, i.e.

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2). \quad (\text{C.6})$$

As a consequence  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  is  $N(\mu, \frac{\sigma^2}{n})$  distributed for large  $n$ , i.e.

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

In summary, the Theorem of Large Numbers tells us that  $\bar{X}_n$  converges stochastically to  $\mu$ , whereas the Central Limit Theorem tells us that  $\bar{X}_n$  converges to a  $N(\mu, \frac{\sigma^2}{n})$ -distribution as  $n$  tends to infinity.

**PDF of the  $\chi^2$ -Distribution.** The PDF of the  $\chi^2$ -distribution, with  $n$  degrees of freedom, is defined as

$$f(x) = \begin{cases} \frac{x^{n/2-1} \exp(-x/2)}{\Gamma(n/2) 2^{n/2}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (\text{C.7})$$

Note that  $\Gamma(n)$  is the Gamma function, defined as  $\Gamma(n) = (n - 1)!$  for positive integers and  $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$  otherwise.

**PDF of the  $t$ -Distribution.** The PDF of the  $t$ -distribution, with  $n$  degrees of freedom, is defined as

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(n/2)\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} \quad -\infty < x < \infty. \quad (\text{C.8})$$

Note that  $\Gamma(n)$  is the Gamma function, defined as  $\Gamma(n) = (n - 1)!$  for positive integers and  $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$  otherwise.

**PDF of the  $F$ -Distribution.** The PDF of the  $F$ -distribution, with  $n$  and  $m$  degrees of freedom, respectively, is defined as

$$f(x) = \frac{\Gamma(\frac{n+m}{2})(\frac{n}{m})^{n/2} x^{n/2-1}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})(1 + \frac{nx}{m})^{(n+m)/2}}, \quad x > 0. \quad (\text{C.9})$$

The PDF of the  $F$ -distribution with  $m$  and  $n$  degrees of freedom can be derived by interchanging the roles of  $m$  and  $n$ .

## More details on Chap. 9

**Another Example of Sufficiency.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$  where  $\mu$  and  $\sigma^2$  are both unknown. We attempt to find a sufficient statistic for  $\mu$  and  $\sigma^2$ .

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \mu, \sigma^2) &= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left[ -\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2} \right] \\ &= \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right) \right]. \end{aligned}$$

Here the joint density depends on  $x_1, x_2, \dots, x_n$  through two statistics  $t_1(x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i$  and  $t_2(x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i^2$  with  $h(x_1, x_2, \dots, x_n) = 1$ . Thus  $T_1 = \sum_{i=1}^n X_i$  and  $T_2 = \sum_{i=1}^n X_i^2$  are jointly sufficient for  $\mu$  and  $\sigma^2$ . Therefore,  $\bar{X}$  and  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  are jointly sufficient for  $\mu$  and  $\sigma^2$  as they are a one-to-one function of  $T_1$  and  $T_2$ . They are the maximum likelihood estimates for  $\mu$  and  $\sigma^2$ .

## More details on Chap. 10

**Exact Test of Fisher.** Similar to the approximate two-sample binomial test in Sect. 10.4.2, we assume two samples following a binomial distribution with parameters  $(n_1, p_1)$  and  $(n_2, p_2)$ , respectively.

$$\begin{aligned} \mathbf{X} &= (X_1, X_2, \dots, X_{n_1}), \quad X_i \sim B(1; p_1) \\ \mathbf{Y} &= (Y_1, Y_2, \dots, Y_{n_2}), \quad Y_i \sim B(1; p_2). \end{aligned}$$

For the sums of these random variables, we get:

$$X = \sum_{i=1}^{n_1} X_i \sim B(n_1; p_1), \quad Y = \sum_{i=1}^{n_2} Y_i \sim B(n_2; p_2).$$

Let  $Z = X + Y$ . The Exact Test of Fisher uses the fact that the row marginal frequencies  $n_1$  and  $n_2$  in the following table

	Success	Failure	Total
Population A	$X$	$n_1 - X$	$n_1$
Population B	$Z - X = Y$	$n_2 - (Z - X)$	$n_2$
Total	$Z$	$(n_1 + n_2 - Z)$	$n_1 + n_2$

are fixed by the sample sizes  $n_1$  and  $n_2$ . Conditional on the total number of successes  $Z = z$  (i.e. the column margins are assumed to be fixed), the only remaining random variable is  $X$  (since the other three entries of the table are then determined by the realization  $x$  of  $X$  and the margins). Under  $H_0 : p_1 = p_2 = p$ , it can be shown that

$$X \sim H(n, n_1, z),$$

i.e.

$$P(X = x | Z = z) = \frac{\binom{n_1}{x} \binom{n-n_1}{z-x}}{\binom{n}{z}}.$$

The proof is straightforward using the idea of conditioning on  $Z = z$  and the assumption of independence of the two samples:

$$\begin{aligned} P(X = x | Z = z) &= \frac{P(X = x, Z = z)}{P(Z = z)} = \frac{P(X = x, Y = z - x)}{P(Z = z)} \\ &= \frac{P(X = x)P(Y = z - x)}{P(Z = z)} \\ &= \frac{\binom{n_1}{x} p^x (1-p)^{n_1-x} \binom{n_2}{z-x} p^{z-x} (1-p)^{n_2-(z-x)}}{\binom{n}{z} p^z (1-p)^{n-z}} \\ &= \frac{\binom{n_1}{x} \binom{n_2}{z-x}}{\binom{n}{z}} = \frac{\binom{n_1}{x} \binom{n-n_1}{z-x}}{\binom{n}{z}}. \end{aligned}$$

Note that in the equation above we use the fact that under  $H_0$ ,  $Z = X + Y$  is  $B(n, p)$  distributed; see the additivity theorem for the binomial distribution, i.e. Theorem 8.1.1.

*Example C.1* Consider two competing lotteries A and B. Say we buy 10 tickets from each lottery and test the hypothesis of equal winning probabilities. The data can be summarized in a  $2 \times 2$  table:

	Winning	Not winning	Total
Lottery A	1	24	25
Lottery B	7	18	25
Total	8	42	50

In R, we can use the command `fisher.test` to perform the test. Using the example data and  $H_1 : p_1 \neq p_2$ , we get

```
ft <- matrix(nrow=2, ncol=2, data=cbind(c(1,7), c(24,18)))
fisher.test(x=ft)
```



with output

Fisher's Exact Test for Count Data

```
data: fisher.table
p-value = 0.0488
alternative hypothesis: true odds ratio is not equal to 1
```

```

95 percent confidence interval:
 0.002289885 0.992114690
sample estimates:
odds ratio
 0.1114886

```

For the example data and  $\alpha = 0.05$ , the null hypothesis is rejected, since the  $p$ -value is lower than  $\alpha$ . For the calculation of the  $p$ -value, the one-sided and two-sided cases have to be distinguished. The idea is that while fixing the margins at the observed values (i.e. 25, 25, 8, 42), we have to calculate the sum of the probabilities of all tables which have lower probability than the observed table. In  $R$ , one can use the functions `dhyper` and `phyper` for calculating (cumulative) probabilities. For example,  $P(X = 1|Z = 8)$  can be calculated as

`dhyper(1, 25, 25, 8)`



[1] 0.02238402

A more extreme table than the observed one is

0	25	25
8	17	25
8	42	50

with probability  $P(X = 0) = \text{dhyper}(0, 25, 25, 8) = 0.002$ , which is lower than  $P(X = 1)$ . The sum is  $0.0224 + 0.002 = 0.0244$  which is the (left) one-sided  $p$ -value. In this case (not generally true!), the two-sided  $p$ -value is simply two times the one-sided value, i.e.  $2 \cdot 0.0244 = 0.0488$ .

*Remark C.1* The two-sided version of the Exact Test of Fisher can also be used as a test of independence of two binary variables. It is equivalent to the test of equality of two proportions, see Example 10.8.2.

**One-Sample  $\chi^2$ -Test for Testing Hypothesis About the Variance.** We assume a normal population, i.e.  $X \sim N(\mu, \sigma^2)$  and an i.i.d. sample  $(X_1, X_2, \dots, X_n)$  distributed as  $X$ . We only introduce the test for two-sided hypotheses

$$H_0 : \sigma^2 = \sigma_0^2$$

versus

$$H_1 : \sigma^2 \neq \sigma_0^2.$$

The test statistic

$$T(\mathbf{X}) = \frac{(n - 1)S_X^2}{\sigma_0^2}$$

follows a  $\chi_{n-1}^2$ -distribution under  $H_0$ . The critical region is constructed by taking the  $\alpha/2$ - and  $(1 - \alpha/2)$  quantile as critical values; i.e.  $H_0$  is rejected, if

$$t(x) < c_{n-1;\alpha/2}$$

or if

$$t(x) > c_{n-1;1-\alpha/2},$$

where  $c_{n-1;\alpha/2}$  and  $c_{n-1;1-\alpha/2}$  are the desired quantiles of a  $\chi^2$ -distribution. In *R*, the test can be called by the `sigma.test` function in the `TeachingDemos` library or the `varTest` function in library `EnvStats`. Both functions also return a confidence interval for the desired confidence level. Note that the test is biased. An unbiased level  $\alpha$  test would not take  $\alpha/2$  at the tails but two different tail probabilities  $\alpha_1$  and  $\alpha_2$  with  $\alpha_1 + \alpha_2 = \alpha$ .

**F-Test for Comparing Two Variances.** Comparing variances can be of interest when comparing the variability, i.e. the “precision” of two industrial processes; or when comparing two medical treatments with respect to their reliability. Consider two populations characterized by two independent random variables  $X$  and  $Y$  which follow normal distributions:

$$X \sim N(\mu_X, \sigma_X^2), \quad Y \sim N(\mu_Y, \sigma_Y^2).$$

For now, we distinguish the following two hypotheses:

$$\begin{aligned} H_0 : \sigma_X^2 &= \sigma_Y^2 & \text{versus} & \quad H_1 : \sigma_X^2 \neq \sigma_Y^2, & \text{two-sided} \\ H_0 : \sigma_X^2 &\leq \sigma_Y^2 & \text{versus} & \quad H_1 : \sigma_X^2 > \sigma_Y^2, & \text{one-sided.} \end{aligned}$$

The third hypothesis with  $H_1 : \sigma_X^2 < \sigma_Y^2$  is similar to the second hypothesis where  $X$  and  $Y$  are replaced with each other.

## Test Statistic

Let  $(X_1, X_2, \dots, X_{n_1})$  and  $(Y_1, Y_2, \dots, Y_{n_2})$  be two independent random samples of size  $n_1$  and  $n_2$ . The test statistic is defined as the ratio of the two sample variances

$$T(\mathbf{X}, \mathbf{Y}) = \frac{S_X^2}{S_Y^2}, \quad (\text{C.10})$$

which is, under the null hypothesis,  $F$ -distributed with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom, see also Sect. 8.3.3.

## Critical Region

**Two-Sided Case.** The motivation behind the construction of the critical region for the two-sided case,  $H_0: \sigma_X^2 = \sigma_Y^2$  vs.  $H_1: \sigma_X^2 \neq \sigma_Y^2$ , is that if the null hypothesis is true (i.e. the two variances are equal) then the test statistic (C.10) would be 1; also,  $T(\mathbf{X}, \mathbf{Y}) > 0$ . Therefore, very small (but positive) and very large values of  $T(\mathbf{X}, \mathbf{Y})$  should lead to a rejection of  $H_0$ . The critical region can then be written as  $K = [0, k_1] \cup (k_2, \infty)$ , where  $k_1$  and  $k_2$  are critical values such that

$$P(T(\mathbf{X}, \mathbf{Y}) < k_1 | H_0) = \alpha/2$$

$$P(T(\mathbf{X}, \mathbf{Y}) > k_2 | H_0) = \alpha/2.$$

Here  $k_1$  and  $k_2$  can be calculated from the quantile function of the  $F$ -distribution as

$$k_1 = f_{n_1-1, n_2-1, \alpha/2}, \quad k_2 = f_{n_1-1, n_2-1, 1-\alpha/2}.$$

*Example C.2* Let  $n_1 = 50$ ,  $n_2 = 60$  and  $\alpha = 0.05$ . Using the `qf` command in *R*, we can determine the critical values as:

```
qf(q=0.025, df1=50-1, df2=60-1)
qf(q=0.975, df1=50-1, df2=60-1)
```



The results are  $k_1 = 0.5778867$  and  $k_2 = 1.706867$ .

*Remark C.2* There is the following relationship between quantiles of the  $F$ -distribution:

$$f_{n_1-1; n_2-1; \alpha/2} = \frac{1}{f_{n_2-1; n_1-1; 1-\alpha/2}}.$$

**One-Sided Case.** In the one-sided case, the alternative hypothesis is always formulated in such a way that the variance in the numerator is greater than the variance in the denominator. The hypotheses are  $H_0: \sigma_X^2 \leq \sigma_Y^2$  versus  $H_1: \sigma_X^2 > \sigma_Y^2$  or  $H_0: \sigma_X^2/\sigma_Y^2 \leq 1$  versus  $H_1: \sigma_X^2/\sigma_Y^2 > 1$ . This means it does not matter whether  $H_1: \sigma_X^2 > \sigma_Y^2$  or  $H_1: \sigma_X^2 < \sigma_Y^2$ ; by constructing the test statistic in the correct way, we implicitly specify the hypothesis. The critical region  $K$  consists of the largest values of  $T(\mathbf{X}, \mathbf{Y})$ , i.e.  $K = (k, \infty)$ , where  $k$  has to fulfil the condition

$$P(T(\mathbf{X}, \mathbf{Y}) > k | H_0) = \alpha.$$

The resulting critical value is denoted by  $k = f_{n_1-1; n_2-1; 1-\alpha}$ .

## Observed Test Statistic

Using the sample variances, the realized test statistic  $t$  is calculated as:

$$t(x, y) = \frac{s_x^2}{s_y^2}, \quad s_x^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2.$$

## Test Decisions

Case	$H_0$	$H_1$	Reject $H_0$ , if
(a)	$\sigma_X = \sigma_Y$	$\sigma_X \neq \sigma_Y$	$t(x, y) > f_{n_1-1; n_2-1; 1-\alpha/2}$ or $t(x, y) < f_{n_1-1; n_2-1; \alpha/2}$
(b)	$\sigma_X \leq \sigma_Y$	$\sigma_X > \sigma_Y$	$t(x, y) > f_{n_1-1; n_2-1; 1-\alpha}$

*Remark C.3* We have tacitly assumed that the expected values  $\mu_X$  and  $\mu_Y$  are unknown and have to be estimated. However, this happens rarely, if ever, in practice. When estimating the expected values by the arithmetic means, it would be appropriate to increase the degrees of freedom from  $n_1 - 1$  to  $n_1$  and  $n_2 - 1$  to  $n_2$ . Interestingly, standard software will not handle this case correctly.

*Example C.3* A company is putting baked beans into cans. Two independent machines at two sites are used. The filling weights are assumed to be normally distributed with mean 1000 g. It is speculated that one machine is more precise than the other. Two samples of the two machines give the following results:

Sample	$n$	$\bar{x}$	$s^2$
X	20	1000.49	72.38
Y	25	1000.26	45.42

With  $\alpha = 0.1$  and the hypotheses

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2,$$

we get the following quantiles

```
qf(0.05, 20-1, 25-1)
[1] 0.4730049
qf(0.95, 20-1, 25-1)
[1] 2.039858
```



The observed test statistic is

$$t = \frac{72.38}{45.42} = 1.59.$$

Therefore,  $H_0$  is not rejected, since  $k_1 \leq t \leq k_2$ . We cannot reject the hypothesis of equal variability of the two machines. In R, the  $F$ -test can be used using the command `var.test`.

*Remark C.4* For the  $t$ -test, we remarked that the assumption of normality is not crucial because the test statistic is approximately normally distributed, even for moderate sample sizes. However, the  $F$ -test relies heavily on the assumption of normality. This is why alternative tests are often used, for example the Levene's test.

## More details on Chap. 11

**Obtaining the Least Squares Estimates in the Linear Model.** The function  $S(a, b)$  describes our optimization problem of minimizing the residual sum of squares:

$$S(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2.$$

Minimizing  $S(a, b)$  is achieved using the principle of maxima and minima which involves taking the partial derivatives of  $S(a, b)$  with respect to both  $a$  and  $b$  and setting them equal to 0. The partial derivatives are

$$\frac{\partial}{\partial a} S(a, b) = \sum_{i=1}^n \frac{\partial}{\partial a} (y_i - a - bx_i)^2 = -2 \sum_{i=1}^n (y_i - a - bx_i), \quad (\text{C.11})$$

$$\frac{\partial}{\partial b} S(a, b) = \sum_{i=1}^n \frac{\partial}{\partial b} (y_i - a - bx_i)^2 = -2 \sum_{i=1}^n (y_i - a - bx_i)x_i. \quad (\text{C.12})$$

Now we set (C.11) and (C.12) as equal to zero, respectively:

$$\begin{aligned} \text{(I)} \quad & \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i) = 0, \\ \text{(II)} \quad & \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)x_i = 0. \end{aligned}$$

This equates to

$$\begin{aligned} (\text{I}') \quad n\hat{a} + \hat{b} \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i, \\ (\text{II}') \quad \hat{a} \sum_{i=1}^n x_i + \hat{b} \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i. \end{aligned}$$

Multiplying (I') by  $\frac{1}{n}$  yields

$$\hat{a} + \hat{b}\bar{x} = \bar{y}$$

which gives us the solution for  $a$ :

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Putting this solution into (II') gives us

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

Using  $\sum_{i=1}^n x_i = n\bar{x}$  leads to

$$\hat{b} \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}.$$

If we use

$$\sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = S_{xx}$$

and

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

we eventually obtain the least squares estimate of  $b$ :

$$\begin{aligned} \hat{b} S_{xx} &= S_{xy} \\ \hat{b} &= \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

*Remark C.5* To show that the above solutions really relate to a minimum, and not to a maximum, we would need to look at all the second-order partial derivatives of  $S(a, b)$  and prove that the bordered Hessian matrix containing these derivatives is always positive definite. We omit this proof however.

### Variance Decomposition.

We start with the following equation:

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}).$$

If we square both sides we obtain

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 - 2 \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}).$$

The last term on the right-hand side is

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}) &\stackrel{(11.8)}{=} \sum_{i=1}^n (y_i - \bar{y}) \hat{b}(x_i - \bar{x}) \\ &= \hat{b} S_{xy} \stackrel{(11.6)}{=} \hat{b}^2 S_{xx} \stackrel{(11.8)}{=} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \end{aligned}$$

We therefore obtain

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

which equates to

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

### The Relation between $R^2$ and $r$ .

$$\begin{aligned} S Q_{\text{Residual}} &= \sum_{i=1}^n (y_i - (\hat{a} + \hat{b}x_i))^2 \stackrel{(11.8)}{=} \sum_{i=1}^n [(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})]^2 \\ &= S_{yy} + \hat{b}^2 S_{xx} - 2\hat{b} S_{xy} \\ &= S_{yy} - \hat{b}^2 S_{xx} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \\ S Q_{\text{Regression}} &= S_{yy} - S Q_{\text{Residual}} = \frac{(S_{xy})^2}{S_{xx}}. \end{aligned}$$

We therefore obtain

$$R^2 = \frac{S_{\text{Regression}}}{S_{yy}} = \frac{(S_{xy})^2}{S_{xx} S_{yy}} = r^2.$$

### The Least Squares Estimators are Unbiased.

$$E(\hat{\beta}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})$$

Given that  $\mathbf{X}$  in the model is assumed to be fixed (i.e. non-stochastic and not following any distribution), we obtain

$$E(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}).$$

Since  $E(\epsilon) = \mathbf{0}$  it follows that  $E\mathbf{y} = \mathbf{X}\beta$  and therefore

$$E(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta.$$

**How to Obtain the Variance of the Least Squares Estimator.** With the same arguments as above (i.e  $\mathbf{X}$  is fixed and non-stochastic) and applying the rule  $\text{Var}(bX) = b^2\text{Var}(X)$  from the scalar case to matrices we obtain:

$$\text{Var}(\hat{\beta}) = \text{Var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

**Maximum Likelihood Estimation in the Linear Model.** The linear model follows a normal distribution:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}).$$

Therefore, the likelihood function of  $\mathbf{y}$  also follows a normal distribution:

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \right\}.$$

The log-likelihood function is given by

$$l(\beta, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta).$$

To obtain the maximum likelihood estimates of  $\beta$  and  $\sigma^2$ , one needs to obtain the maxima of the above function using the principle of maxima and minima that involves setting the partial derivatives equal to zero and finding the solution:

$$\begin{aligned}\frac{\partial l}{\partial \beta} &= \frac{1}{2\sigma^2} 2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}, \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = 0.\end{aligned}$$

We therefore have

$$\begin{aligned}\mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'\mathbf{y}, \text{ or } \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \hat{\sigma}^2 &= \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})\end{aligned}$$

which give us the ML estimates of  $\beta$  and  $\sigma^2$  in the linear regression model. Here, we also need to check the Hessian matrix of second-order partial derivatives to show that we really found a minimum and not a maximum. We omit this proof however.

## Distribution Tables

See Tables C.1, C.2 and C.3.

**Table C.1** CDF values for the standard normal distribution,  $\Phi(z)$ . These values can also be obtained in R by using the `pnorm(p)` command

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.879000	0.881000	0.882977
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965620	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691

(continued)

**Table C.1** (continued)

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999

**Table C.2**  $(1 - \alpha)$  quantiles for the  $t$ -distribution. These values can also be obtained in *R* using the `qt(p, df)` command.

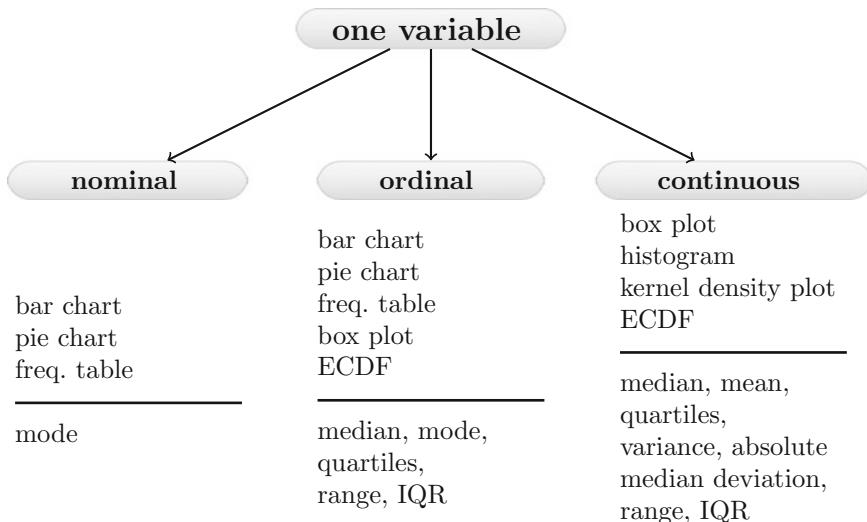
$df$	$1 - \alpha$			
	0.95	0.975	0.99	0.995
1	6.3138	12.706	31.821	63.657
2	2.9200	4.3027	6.9646	9.9248
3	2.3534	3.1824	4.5407	5.8409
4	2.1318	2.7764	3.7469	4.6041
5	2.0150	2.5706	3.3649	4.0321
6	1.9432	2.4469	3.1427	3.7074
7	1.8946	2.3646	2.9980	3.4995
8	1.8595	2.3060	2.8965	3.3554
9	1.8331	2.2622	2.8214	3.2498
10	1.8125	2.2281	2.7638	3.1693
11	1.7959	2.2010	2.7181	3.1058
12	1.7823	2.1788	2.6810	3.0545
13	1.7709	2.1604	2.6503	3.0123
14	1.7613	2.1448	2.6245	2.9768
15	1.7531	2.1314	2.6025	2.9467
16	1.7459	2.1199	2.5835	2.9208
17	1.7396	2.1098	2.5669	2.8982
18	1.7341	2.1009	2.5524	2.8784
19	1.7291	2.0930	2.5395	2.8609
20	1.7247	2.0860	2.5280	2.8453
30	1.6973	2.0423	2.4573	2.7500
40	1.6839	2.0211	2.4233	2.7045
50	1.6759	2.0086	2.4033	2.6778
60	1.6706	2.0003	2.3901	2.6603
70	1.6669	1.9944	2.3808	2.6479
80	1.6641	1.9901	2.3739	2.6387
90	1.6620	1.9867	2.3685	2.6316
100	1.6602	1.9840	2.3642	2.6259
200	1.6525	1.9719	2.3451	2.6006
300	1.6499	1.9679	2.3388	2.5923
400	1.6487	1.9659	2.3357	2.5882
500	1.6479	1.9647	2.3338	2.5857

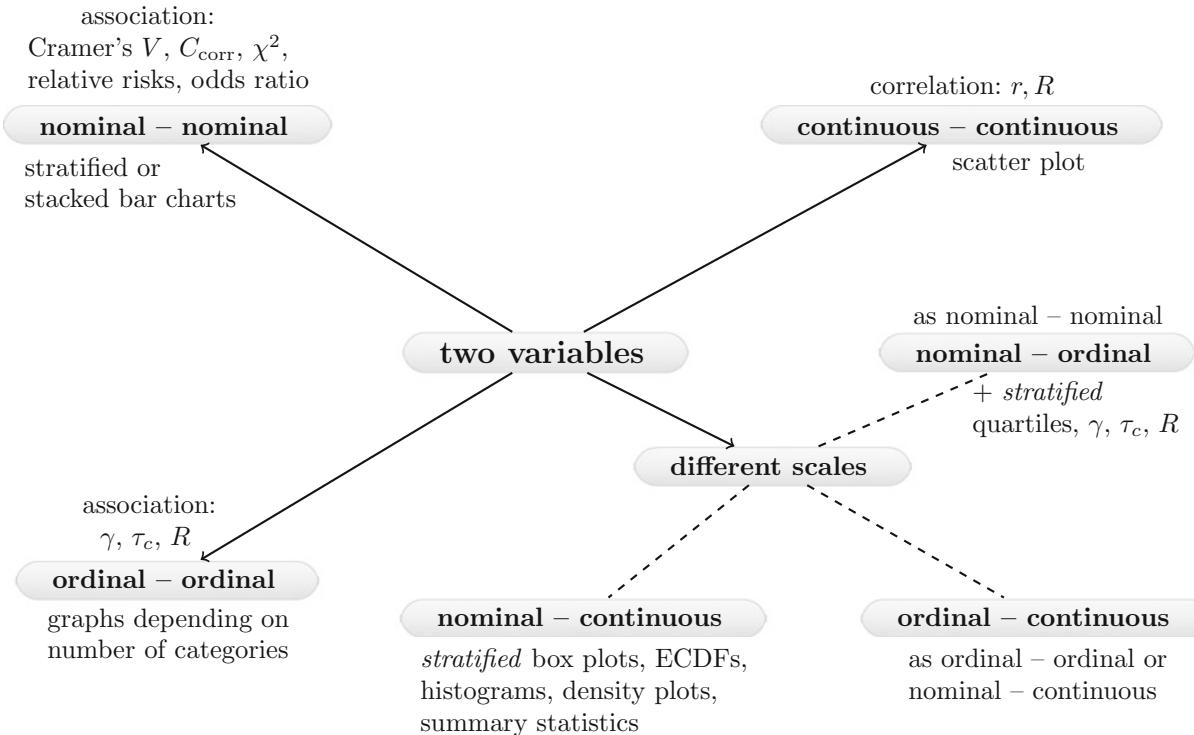
**Table C.3**  $(1 - \alpha)$  quantiles of the  $\chi^2$ -distribution. These values can also be obtained in *R* using the `qchisq(p,df)` command

$df$	$1 - \alpha$					
	0.01	0.025	0.05	0.95	0.975	0.99
1	0.0001	0.001	0.004	3.84	5.02	6.62
2	0.020	0.051	0.103	5.99	7.38	9.21
3	0.115	0.216	0.352	7.81	9.35	11.3
4	0.297	0.484	0.711	9.49	11.1	13.3
5	0.554	0.831	1.15	11.1	12.8	15.1
6	0.872	1.24	1.64	12.6	14.4	16.8
7	1.24	1.69	2.17	14.1	16.0	18.5
8	1.65	2.18	2.73	15.5	17.5	20.1
9	2.09	2.70	3.33	16.9	19.0	21.7
10	2.56	3.25	3.94	18.3	20.5	23.2
11	3.05	3.82	4.57	19.7	21.9	24.7
12	3.57	4.40	5.23	21.0	23.3	26.2
13	4.11	5.01	5.89	22.4	24.7	27.7
14	4.66	5.63	6.57	23.7	26.1	29.1
15	5.23	6.26	7.26	25.0	27.5	30.6
16	5.81	6.91	7.96	26.3	28.8	32.0
17	6.41	7.56	8.67	27.6	30.2	33.4
18	7.01	8.23	9.39	28.9	31.5	34.8
19	7.63	8.91	10.1	30.1	32.9	36.2
20	8.26	9.59	10.9	31.4	34.2	37.6
25	11.5	13.1	14.6	37.7	40.6	44.3
30	15.0	16.8	18.5	43.8	47.0	50.9
40	22.2	24.4	26.5	55.8	59.3	63.7
50	29.7	32.4	34.8	67.5	71.4	76.2
60	37.5	40.5	43.2	79.1	83.3	88.4
70	45.4	48.8	51.7	90.5	95.0	100.4
80	53.5	57.2	60.4	101.9	106.6	112.3
90	61.8	65.6	69.1	113.1	118.1	124.1
100	70.1	74.2	77.9	124.3	129.6	135.8

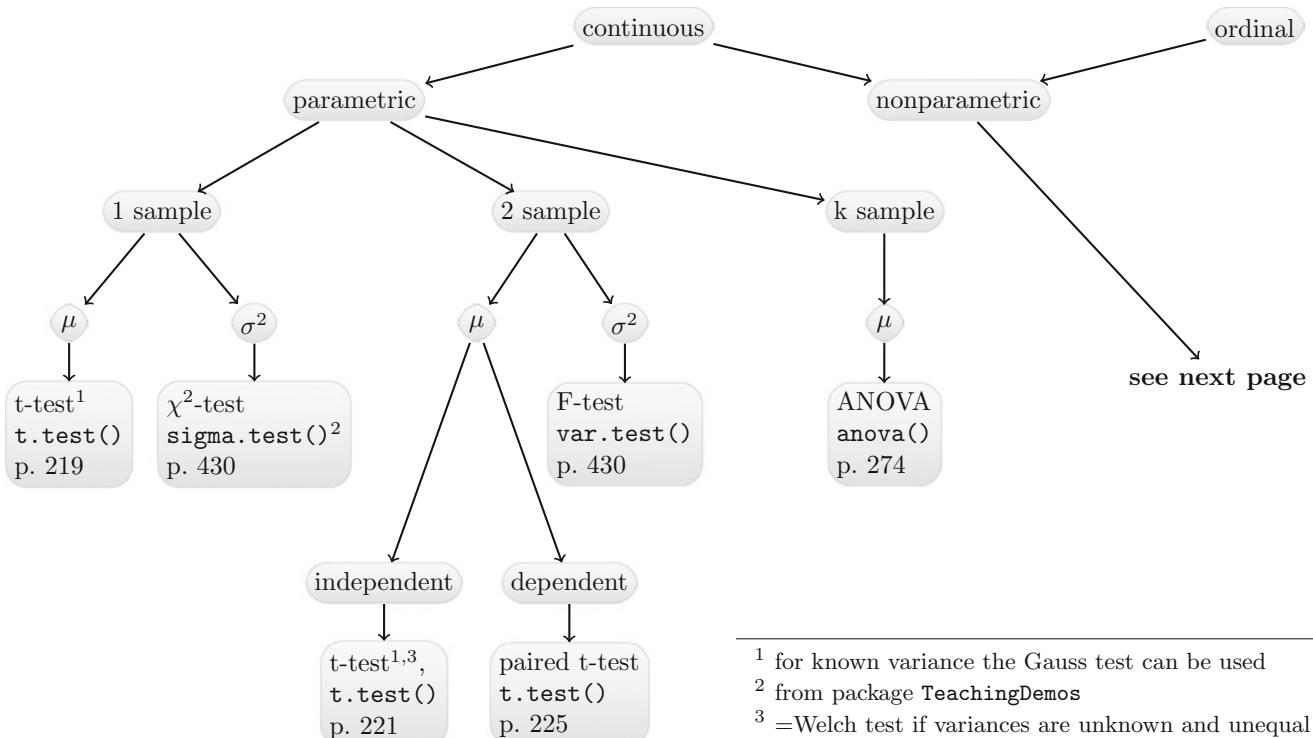
**Quantiles of the *F*-Distribution.** These quantiles can be obtained in *R* using the `qf(p,df1,df2)` command.

### Descriptive Data Analysis





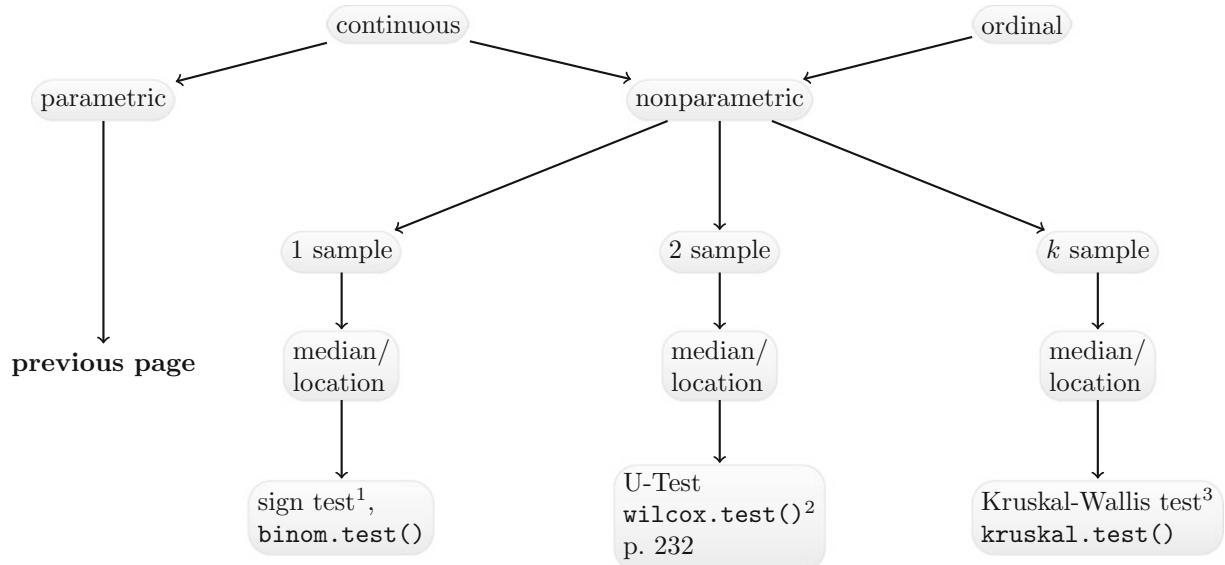
## Summary of Tests for Continuous and Ordinal Variables



<sup>1</sup> for known variance the Gauss test can be used

<sup>2</sup> from package **TeachingDemos**

<sup>3</sup> =Welch test if variances are unknown and unequal

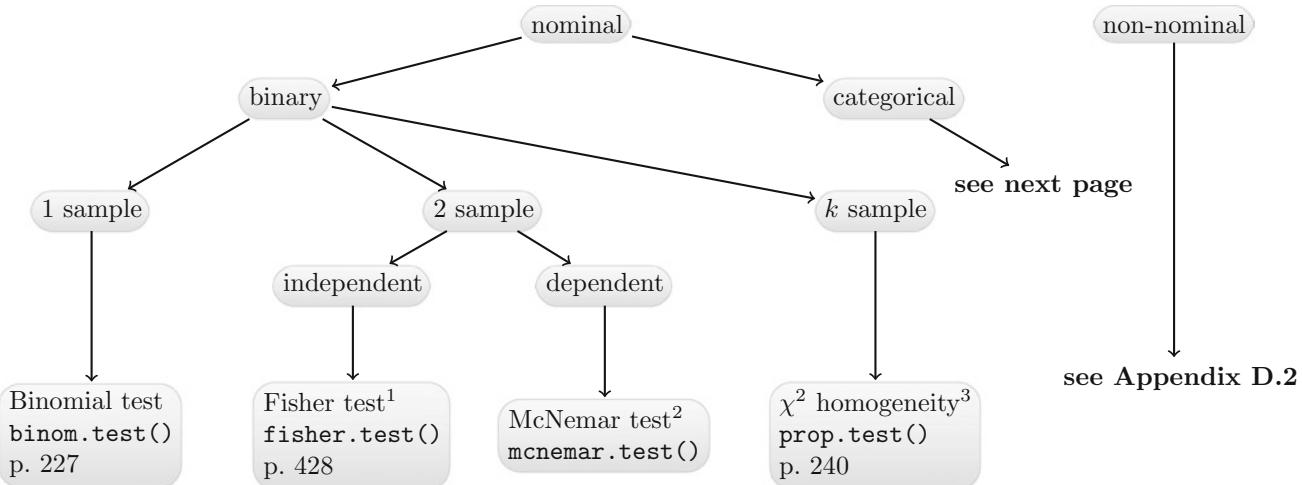


<sup>1</sup> not explained in this book; alternative: Mood's median test

<sup>2</sup> use option `paired=TRUE` for dependent data

<sup>3</sup> not explained in this book; use Friedman test for dependent data

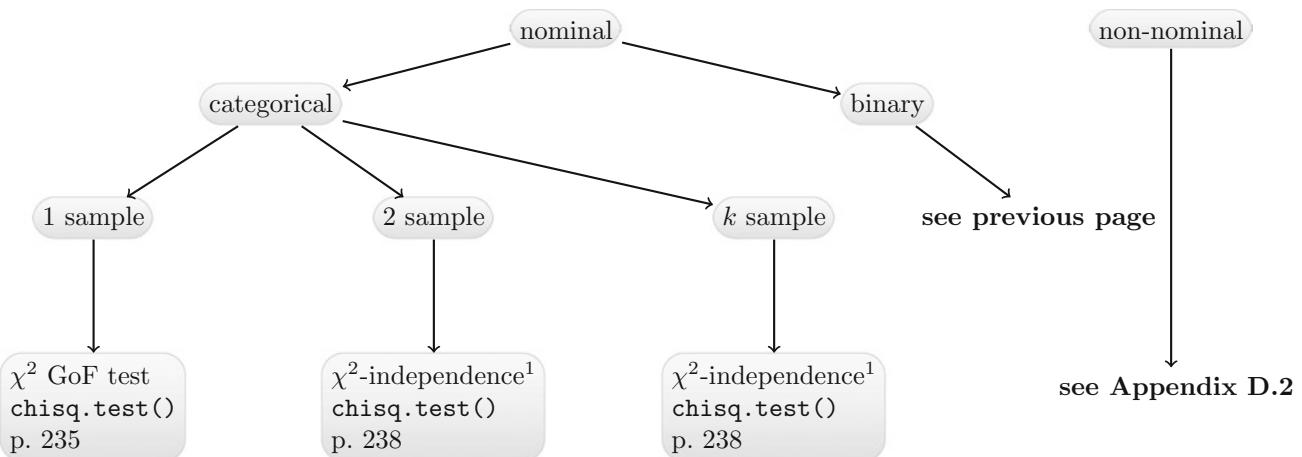
## Summary of Tests for Nominal Variables



<sup>1</sup> alternative:  $\chi^2$ -independence test (`chisq.test()`)  
(test decision equivalent to  $\chi^2$ -homogeneity test)

<sup>2</sup> not explained in this book

<sup>3</sup> test decision equivalent to  $\chi^2$ -independence test



<sup>1</sup> test decision equivalent to  $\chi^2$ -homogeneity test  
(`prop.test`)

---

## References

- Adler, J. (2012). *R in a Nutshell*. Boston: O'Reilly.
- Albert, J., & Rizzo, M. (2012). *R by Example*. New York: Springer.
- Bock, J. (1997). *Bestimmung des Stichprobenumfangs*. Munich: Oldenbourg Verlag. (in German).
- Casella, G., & Berger, R. (2002). *Statistical inference*. Boston, MA: Cengage Learning.
- Chow, S., Wang, H., & Shao, J. (2007). *Sample size calculations in clinical research*. London: Chapman and Hall.
- Crawley, M. (2013). *The R book*. London: Wiley.
- Dalgaard, P. (2008). *Introductory statistics with R*. Berlin: Springer.
- Everitt, B., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. New York: Springer.
- Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2009). *Survey methodology. Wiley series in survey methodology*. Hoboken, NJ: Wiley.
- Hernan, M., & Robins, J. (2017). *Causal inference*. Boca Raton: Chapman and Hall/CRC.
- Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50, 361–365.
- Kauermann, G., & Küchenhoff, H. (2011). *Stichproben - Methoden und praktische Umsetzung in R*. Heidelberg: Springer. (in German).
- Ligges, U. (2008). *Programmieren in R*. Heidelberg: Springer. (in German).
- R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.
- Young, G., & Smith, R. (2005). *Essentials of statistical inference*. Cambridge: Cambridge University Press.

---

# Index

## A

Absolute deviation, 51  
mean deviation, 51  
median deviation, 51  
Additivity theorem, 116, 147  
Akaike's information criterion (AIC), 282  
Analysis of variance, 274  
ANOVA, 274  
Arithmetic mean, 38  
properties, 40  
weighted, 38  
Association, 67, 249

## B

Backward selection, 282  
Bar chart, 24  
Behrens-Fisher problem, 222  
Bernoulli distribution, 156  
Bias, 184  
Binomial coefficient, 102  
distribution, 157  
Bivariate random variables, 140  
Box plot, 56

## C

Calculation rules  
CDF, 133  
expectation, 144  
normal random variables, 168  
probabilities, 117  
variance, 144  
Causation, 288  
CDF, 129  
calculation rules, 133

joint, 141  
quantile, 137  
quartile, 137  
Central limit theorem, 426  
Central tendency, 38  
Certain event, 110  
 $\chi^2$  distribution, 171, 427  
goodness-of-fit test, 235  
independence test, 238  
test of homogeneity, 240, 241  
variance test, 430  
Coding dummy, 266  
effect, 267  
Coefficient binomial, 102  
of variation, 55  
regression, 251  
Combinations, 102  
with order, 103, 104  
with replacement, 103, 104  
without order, 102, 103  
without replacement, 102, 103  
Combinatorics, 97  
Complementary event, 110  
Composite event, 110  
Conditional distribution, 141, 143  
frequency distribution, 70  
probability, 117  
relative frequency distribution, 70  
Confidence bound, 196  
interval, 196, 197  
interval for  $\mu; \sigma^2$  known, 197  
interval for  $\mu; \sigma^2$  unknown, 198

- interval for  $p$ , 199  
 interval for the odds ratio, 201  
 level, 196  
 Consistency, 189  
 Contingency  
   coefficient, 77  
   table, 68, 140  
 Continuous variable, 6  
 Convergence  
   stochastic, 425  
 Correlation coefficient, 148  
   of Bravais–Pearson, 82  
   of Spearman, 84  
   product moment, 82  
 Covariance, 146, 147  
 Covariate, 251  
 Cramer's V, 77  
 Cross tabulation, *see* contingency table  
 Cumulative  
   distribution function, 129  
   frequency, *see* frequency, cumulative  
   marginal distribution, 143
- D**
- Data  
   matrix, 3, 9  
   observation, 3  
   set, 3, 9  
   transformation, 11  
   unit, 3  
 Decomposition  
   complete, 113  
 Degenerate distribution, 156  
 Density, 129  
 Design matrix, 263, 270  
 Dispersion, 48  
   absolute deviation, 51  
   absolute mean deviation, 51  
   absolute median deviation, 51  
   mean squared error, 51  
   measure, 49  
   range, 49  
   standard deviation, 51  
 Distribution, 19  
   Bernoulli, 156  
   Binomial, 157  
    $\chi^2$ , 171, 427  
   conditional, 141, 143  
   conditional frequency, 70  
   conditional relative frequency, 70  
   continuous, 165  
   cumulative marginal, 143  
   degenerate, 156  
   exponential, 170  
   F, 427  
   Gauss, 166  
   geometric, 163  
   hypergeometric, 163  
   independent and identical, 145, 426  
   joint relative frequency, 70  
   marginal, 140, 142  
   marginal frequency, 70  
   marginal relative frequency, 70  
   multinomial, 161  
   normal, 166  
   Poisson, 160  
   standard, 154  
   Student, 172  
   t, 172, 427  
   uniform discrete, 154  
 Duality, 216  
 Dummy variable, 266
- E**
- Efficiency, 185  
 Elementary event, 110  
 Empirical cumulative distribution function  
   (ECDF), 19  
 Epitools, *see* R packages  
 Error  
   type I, 213  
   type II, 213  
 Estimation  
   interval, 195  
   least squares, 252, 253, 264  
   maximum likelihood, 192  
   method of moments, 195  
   nonparametric, 182  
   parametric, 182  
 Event, 110  
   additive theorem, 116  
   certain, 110  
   composite, 110  
   disjoint, 112  
   elementary, 110  
   impossible, 110  
   simple, 110  
   sure, 110  
   theorem of additivity, 115, 116  
 Expectation, 134  
   calculation rules, 144  
 Expected frequencies, 74

- Experiment  
    Laplace, 114  
    random, 109  
Exponential distribution, 170
- F**
- Factorial function, *see* function, factorial  
F-distribution, 173, 427
- Fisher  
    exact test, 230
- Foreign, *see* R packages
- Frequency  
    absolute, 18, 69, 113  
    cumulative, 19  
    expected, 74  
    relative, 18, 21, 113  
    table, 19, 69
- F-Test, 272, 431
- Function  
    cumulative distribution, *see* CDF  
    empirical cumulative distribution, *see*  
        ECDF  
    factorial, 100  
    joint cumulative distribution, 141  
    joint probability distribution, 140, 141  
    probability mass, *see* PMF  
    step, 133
- G**
- Gamma  
    of Goodman and Kruskal, 87
- Gauss test  
    one-sample, 216  
    two-sample, 221
- Generalized method of moments, 195
- Geometric distribution, 163
- Ggplot2, *see* R packages
- Gini coefficient, 60  
    standardized, 61
- Goodman and Kruskal's  $\gamma$ , 87
- Goodness of fit  
    adjusted measure, 281  
    measure, 258  
    test, 235
- Graph  
    bar chart, 24  
    box plot, 56  
    histogram, 27  
    kernel density plot, 29  
    Lorenz curve, 58  
    pie chart, 26
- QQ-plot, 44  
random plot, 286  
scatter plot, 80
- Growth  
    factor, 46  
    rate, 46
- H**
- Heteroscedasticity, 286  
Histogram, 27  
Homoscedasticity, 286  
Hypergeometric distribution, 163
- Hypothesis, 210  
    alternative, 211  
    linear, 280  
    null, 211  
    one-sided, 211  
    two-sided, 211
- I**
- i.i.d., 145, 153, 426  
Impossible event, 110  
Independence, 73, 121  
    pairwise, 122  
    random variables, 143  
    stochastic, 121, 144
- Ineq, *see* R packages
- Inequality  
    Tschebyschev, 139
- Inference, 181, 182  
    least squares, 252, 264  
    maximum likelihood, 192  
    method of moments, 195
- Interaction, 276
- Intercept, 251
- Interquartile range, 49
- Interval estimation, 195
- J**
- Joint  
    cumulative distribution function, 141  
    frequency distribution, 70  
    probability distribution function, 140,  
        141  
    relative frequency distribution, 70
- K**
- Kernel density plot, 29  
Kolmogorov–Smirnov test, 237

- L**
- Laplace
    - experiment, 114
    - probability, 114
  - Lattice, *see* R packages
  - Least squares, 252
  - Life time, 170
  - Likelihood, 192
  - Line of equality, 59
  - Linear
    - hypotheses, 280
  - Linear model, 251
    - residuals, 270
  - Linear regression, 249
    - interaction, 276
  - Location parameter, 38
  - Log-linear model, 269
  - Lorenz curve, 58
- M**
- Mann-Whitney U-test, 232
  - Marginal
    - distribution, 140, 142
    - frequency distribution, 70
    - relative frequency distribution, 70
  - MASS, *see* R packages
  - Matrix
    - covariance, 147
    - design, 263
  - Maximum likelihood estimation (MLE), 192
  - Mean
    - arithmetic, 38
    - properties, 40
    - weighted arithmetic, 38
  - Mean squared error (MSE), 51, 184
  - Measure
    - dispersion, 49
    - symmetric, 76
  - Measure of association
    - $\chi^2$  coefficient, 76
    - contingency coefficient  $C$ , 77
    - correlation coefficient, 82
    - Cramer's  $V$ , 77
    - odds ratio, 78
    - rank correlation coefficient, 84
    - relative risk, 78
  - Memorylessness, 170
  - Method of moments, 195
  - Model
    - fitted regression model, 253
    - fitted value, 253
    - linear, 251
    - log-linear, 269
    - nonlinear, 251
  - Multinomial distribution, 161
  - Multiple linear regression, 262
  - Multiplication theorem of probability, 119
  - Multivariate, 249
  - Mvtnorm, *see* R packages
- N**
- Namibia, 206
  - Newton-Raphson, 193
  - Nominal variable, 6
  - Normal distribution, 166
- O**
- Observation, 3
  - Odds ratio, 78
  - One-sample problem, 209, 210
  - Ordered
    - set, 99
    - values, 20
  - Ordinal variable, 6
  - Outcome, 251
- P**
- Parameter
    - location, 38
    - regression, 251
    - space, 184
  - PDF, 129
    - joint, 140, 141
  - Percentile, 43
  - Permutation, 101
    - with replacement, 101
    - without replacement, 101
  - Pie chart, 26
  - Plot
    - kernel density, 29
    - QQ, 44
    - random, 286
    - scatter, 79
    - trumpet, 286, 288
  - Poisson distribution, 160
  - Polynomial regression, 267
  - Population, 4
  - Power, 213
  - Probability
    - calculation rules, 117
    - conditional, 117, 119

- density function, 129  
Laplace, 114  
mass function, 132  
posterior, 120  
prior, 120  
Probability theory  
  axioms, 115  
*p*-value, 215
- Q**  
QQ-plot, 44  
Quantile, 42, 137  
Quartile, 43, 137  
Quintile, 43
- R**  
 $R^2$   
  adjusted, 281  
Random variables, 127  
  bivariate, 140  
  continuous, 129  
  discrete, 131  
  i.i.d., 153  
  independence, 144  
  standardization, 138  
Range, 49  
Real stories, 63, 124, 176, 244, 245  
Realization, 128  
Reference category, 266  
Regression  
  line, 251  
  linear, 249  
  multiple linear, 262  
  polynomial, 267  
Regressor, 251  
Relationship, 249  
Relative risk, 78  
Residuals, 253, 270  
  standardized, 285  
Response, 251  
R packages  
  compositions, 219  
  epitools, 202  
  foreign, 12  
  ggplot2, 35, 73, 199  
  ineq, 59, 62  
  lattice, 73  
  MASS, 28, 283  
  mvtnorm, 316  
  ryouready, 87, 88  
TeachingDemos, 445
- vcd, 77
- S**  
Sample  
  estimate, 182  
  pooled variance, 222  
  space, 110  
  variance, 51  
Sampling  
  without replacement, 163  
Scale  
  absolute, 7  
  continuous, 6  
  interval, 7  
  nominal, 6  
  ordinal, 6  
  ratio, 7  
Scatter plot, 80  
Set  
  ordered, 100  
  unordered, 100  
Significance level, 213  
Simple event, 110  
Slope, 251  
Standard deviation, 51, 136  
Standard error, 189, 205  
Standardization, 54, 138  
Standard normal distribution, 166  
Statistic, 146, 183  
Step function, 20, 133  
Stochastic convergence, 425  
Stuart's  $\tau_c$ , 87  
Sufficiency, 190  
Sure event, 110
- T**  
Table  
  contingency, 68, 69, 140  
  frequency, 19, 69  
 $\tau_c$   
  of Stuart, 87  
T-distribution, 172, 427  
Test  
  ANOVA, 274  
  Binomial, 227  
   $\chi^2$  for variance, 232, 430  
   $\chi^2$  goodness of fit, 235  
   $\chi^2$  independence, 238  
   $\chi^2$  of homogeneity, 240  
  duality, 216  
  equivalence, 213

- 
- F, 272, 431  
 Fisher, 230, 428  
 Friedman, 446  
 Kolmogorov–Smirnov, 237  
 Kruskal–Wallis, 446  
 Mann–Whitney U, 232  
 McNemar, 447  
 Mood, 446  
 one-sample Gauss test, 216  
 one-sample  $t$ -test, 219  
 one-sided, 211  
 overall F, 272  
 paired  $t$ -test, 225  
 sign, 446  
 significance, 213  
 two-sample binomial, 230  
 two-sample Gauss test, 221  
 two-sample  $t$ -test, 222  
 two-sided, 211  
 U, 232  
 Welch, 222  
 Wilcoxon rank sum, 232  
 Wilcoxon–Mann–Whitney, 232  
 Test distributions, 317  
 Theorem  
     additivity, 116, 147  
     additivity of  $\chi^2$  variables, 172  
     additivity of disjoint events, 115  
 Bayes, 120  
 central limit, 426  
 i.i.d., 185  
 large numbers, 425, 426  
 law of total probability, 119  
 multiplication for probabilities, 119  
 Neyman–Fisher Factorization, 191  
 PDF, 129  
 standardization, 138  
 Student, 172  
 Tschebyschev, 139  
 variance decomposition, 136  
 Tie, 85, 86  
 Transformation, 11  
 T-test  
 one-sample, 219  
 paired, 225  
 two-sample, 222  
 Two-sample problem, 210
- U**  
 Uniform distribution  
     continuous, 165  
     discrete, 154  
 Unit, 3  
 Unordered set, 100  
 U test, 232
- V**  
 Variable, 4  
     binary, 7  
     bivariate random, 140  
     categorical, 7  
     continuous, 6  
     dependent, 251  
     discrete, 6  
     dummy, 266  
     grouped, 7  
     independent, 251  
     nominal, 6  
     ordinal, 6  
     random, 127  
     response, 251  
     standardized, 54  
 Variance, 51, 135  
     additivity theorem, 147  
     between classes, 53  
     calculation rules, 144  
     decomposition, 257  
     dispersion, 48  
     pooled, 222  
     within classes, 53  
 Vcd, *see* R packages
- W**  
 Welch test, 222  
 Whiskers, 56  
 Wilcoxon–Mann–Whitney test, 232