Babak Shahbaba

Biostatistics with R

An Introduction to Statistics Through Biological Data

EXTRA MATERIALS
extras.springer.com

Springer

Use R!

Babak Shahbaba

Biostatistics with R

An Introduction to Statistics
Through Biological Data

Springer

Prof. Babak Shahbaba
Department of Statistics
University of California, Irvine
Irvine, CA 92697-1250
USA
babaks@uci.edu

Series Editors:
Robert Gentleman
Program in Computational Biology
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue, N. M2-B876
Seattle, WA 98109
USA

Kurt Hornik
Department of Statistik and Mathematik
Wirtschaftsuniversität Wien
Augasse 2-6
1090 Wien
Austria

Giovanni Parmigiani
The Sidney Kimmel Comprehensive
Cancer Center at Johns Hopkins University
550 North Broadway
Baltimore, MD 21205-2011
USA

Additional material to this book can be downloaded from http://extras.springer.com

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Rezy and Ryan

Preface

This book will discuss basic statistical analysis methods through a series of biological examples using R and R-Commander as computational tools. The book is intended for a wide range of readers, from people with relatively strong analytical background who want to learn about statistics and its application in biology, to nonstatistician scientists who use statistical methods in their research.

While the theoretical aspects of statistics are intriguing and interesting on their own, we believe that what separates statistics from other branches of mathematics is its intimate relationship with other fields, such as biology, economics, and social sciences, and its widespread application in these areas. In statistics, a theoretical work is usually inspired by applied problems, and new theories usually find immediate applications in real-world problems. This interweaving of theory and application has put statistics in a special place in the scientific world.

In this book, most topics are motivated by real examples first. We believe that learning a new topic becomes easier if it is motivated by interesting and engaging applied problems. We also hope that this approach helps students to improve their critical thinking and problem-solving skills for situations where they are presented with new problems. To this end, we motivate each new topic with a relevant problem from biology. We then try to reach the solution intuitively before discussing the related statistical methods. For example, when discussing Bayes' theorem, we first present a biological problem (finding the probability of lung cancer for smokers) and find the answer to that problem intuitively based on what we already know. Then, we introduce Bayes' theorem as a general form of our solution for this type of problem.

While discussing statistical methods and their applications, our goal is to keep a balance between mathematical rigor and readability. To accomplish this, we have moved concepts that tend to be more complex with limited applications in everyday analysis to the end of each chapter in "Advanced" sections. For the most part, these sections could be skipped in the first reading of this book.

Throughout the book, we use R-Commander, a free and publicly available computer program, to show how statistical methods can be used in practice. We believe that using these methods while learning them could help with the learning process.

Most of the examples discussed in this book are based on scientific studies whose data are publicly available. For each example, we provide the step-by-step application of R-Commander. Readers are encouraged to follow these steps while reading the book so that they can learn statistics through hands-on experience.

For some examples, the data are available through R and R-Commander. For these examples, we provide the steps required to obtain the data. For some other examples, the data are available online and can be downloaded from http://extras.springer.com. Appendix A shows the steps for installing and using R-Commander. Before reading the chapters, readers should follow these steps to install R-Commander on their computer.

The chapters are arranged according to what a typical statistical analysis involves. We usually start with some specific scientific questions in mind. Then, we design a scientific study to answer those questions. In Chap. 1, we very briefly discuss different types of studies and their objectives. We also present an overview of typical steps we take from raising a scientific question to answering it through statistical methods. These steps always involve identifying a *target population*, which is the group of individuals we want to study (e.g., population of humans, orange trees, cells).

Because the target populations are usually very large, we conduct our studies on a relatively small number of individuals randomly *sampled* (i.e., selected) from the population. From these individuals we collect information in the form of measurements of some specific characteristics such as age, size, and counts. We refer to the information obtained from these individuals as *data* collectively. In Chaps. 2 and 3, we discuss several *data exploration* techniques, which involve summarizing and visualizing data to obtain a high-level understanding of the data and the target population.

We want to generalize what we learn from the individuals participating in our study (i.e., the randomly selected individuals) to the whole population. This generalization should always be accompanied by our acknowledgement that we are not completely certain about our findings since our knowledge of the population is based on a relatively small sample of individuals from that population. Specifically, we always present our findings along with some measurements that reflect the extent of our *uncertainty*. To this end, we use *probability* as a powerful mathematical tool to measure uncertainty. We discuss probability in Chaps. 4 and 5.

The process of analyzing data to learn about the whole population is referred to as *statistical inference*. This usually involves guessing some unknown values, drawing conclusions, and making decisions. Chapters 6, 7, and 8 discuss some basic methods of inferential statistics. Chapters 9, 10, and 11 provide slightly more advanced statistical inference methods, which for the most part could be considered as the generalization of topics covered in Chap. 8.

Finally, Chap. 12 discusses *clustering* methods, and Chap. 13 discusses *Bayesian analysis* very briefly. These topics are not traditionally included in introductory books on statistics. We decided to include these topics due to their immense importance in scientific studies. While this book does not do justice to these two topics, we hope that it serves as an introduction for interested readers.

As mentioned above, we use R-Commander to show how statistical methods can be used for real problems. Using R-Commander does not require any computer programming. For readers who are comfortable with learning a programming language, we discuss the equivalent R programs at the end of each chapter in Advanced sections. These readers should start from Appendix B, where we provide a brief introduction to R programming.

The methods discussed in this book have been developed by many researchers over many years. To avoid overburdening the reader, we provide only a small number of references, mainly for related books that go beyond what we have covered here, and also for real problems that we have used as examples.

Writing this book has helped me to improve my teaching, and the feedback I have received from my students has helped me to improve the book during the past several years. I would like to thank all my students who challenged me with their questions; they have been my toughest critics.

I would like to thank John Fox for developing R-Commander. This is an extremely useful tool for teaching basic statistics to students without programming background.

I would also like to thank Jessica Utts, Michael Phelan, Sam Behseta, and Wesley Johnson for reviewing the book and providing thoughtful comments and constructive criticisms to improve the quality of this book. I am very grateful to have such good friends and supportive colleagues.

A very special thanks goes to Laura Balzer, who is currently a graduate student at UC Berkeley. She has been extremely helpful in the process of preparing the initial draft and editing the book.

Finally, I would like to thank my family for being patient and supportive throughout the process of writing this book; it would not have been possible without their love and support.

Irvine, CA   Babak Shahbaba
November 11, 2011

Contents

Chapter 1
Introduction

1.1 Statistical Methods in the Context of Scientific Studies

This book discusses statistical methods from the application point of view. More specifically, we focus on biostatistical methods, which involve applying statistical methods to biological and health-related problems. Each section poses one or more practical problems and then presents the statistical tools related to solving these problems. The materials presented in this book cover basic and essential steps involved in analysis of biological and health-related data.

The overall objective of statistical methods is to use **empirical evidence** in order to improve our knowledge about the **target population**, which includes the entire group of individuals and objects (e.g., people, plants, cells) we want to study. As a result, statistics helps us to make more informed **decisions**. We study the population of interest by measuring a set of characteristics (e.g., age, size, weight) that are related to our study. We refer to these characteristics, whose values can change from one member of the population to another one, as **variables**. The objective of many scientific studies is to learn about the **variation** of a specific characteristic (variable) in the population of interest. For example, we might be interested in the range of normal body temperature among healthy people, or tumor size in breast cancer patients, or growth rate of walnut tress, or BMI (body mass index) in the US population. In many studies, we want to explain or predict how a variable changes with respect to some other variables. That is, we want to identify possible **relationships** among different variables. For example, we might want to study the effects of different diets on early growth of chicks, or ask how heart rate changes with body temperature, or whether a higher BMI is associated with higher blood pressure, or whether survival of breast cancer patients depends on the type of treatments (mastectomy vs. breast conservation therapy) they receive. We refer to the variables that are the main focus of our study as the **response** (or target) variables. In contrast, we call variables that explain or predict the variation in the response variable as **explanatory** variables or **predictors**.

Statistical analysis begins with a scientific problem usually presented in the form of a **hypothesis testing** or a **prediction** problem. Hypothesis testing refers to the

process of examining a scientific statement that explains a phenomenon. In general, hypothesis testing problems can be regarded as **decision** problems, where we need to decide to accept or reject the proposed explanation for the phenomenon. For example, Mackowiak et al. (1992) [19] asked whether the average normal body temperature is the widely accepted value of 98.6°F. Their hypothesis was that the average normal body temperature is less than the accepted value. A hypothesis might also be expressed in terms of possible relationships between two or more characteristics. For example, we might hypothesize that the normal body temperature is different between men and women. This means we believe that the body temperature and gender are related. For breast cancer patients, we might hypothesize that mastectomy leads to longer survival of patients compared to those who are treated with breast conservation therapy (lumpectomy, nodal dissection, and radiation).

Statistical methods are used to evaluate a hypothesis based on empirical data. Using these methods, we can decide whether we should reject a hypothesis or not. Such decisions in turn help us to make more informed decisions with respect to the scientific problem that inspired our study. For example, at the conclusion of their study, Mackowiak et al. argue that the average normal body temperature seems to be lower than previously believed, and a new upper limit for the range of normal body temperature should be considered. This recommendation has important consequences for deciding the body temperature set point and whether someone has a fever that requires medication. For treating breast cancer patients, several studies [27] have shown that there is no evidence of difference in survival between mastectomy and breast conservation therapy, at least for patients with less severe situations (e.g., small tumors, node negative). Based on these results, The US National Cancer Institute (NCI) recommended breast conservation operations, especially for the type of patients who participated in these studies (i.e., with less sever cancer), instead of mastectomy, which was the standard treatment in the 1960s.

In recent years, high-throughput scientific studies without any clear hypothesis have become very common. For example, scientists may examine thousands of genes with respect to their relationship to a disease without hypothesizing that any specific gene is responsible for the disease. In these studies, the objective is to explore a large number of possible factors (e.g., genes) in order to identify a small number of them for follow-up studies that tend to be more thorough with much smaller scales. Therefore, the initial large-scale studies are not designed for hypothesis testing rather generating a small number of hypotheses, which can be the focus of follow-up studies and tested properly in future.

Scientific problems are sometimes presented as prediction problems. Prediction refers to the process of guessing the value of the response variable using a set of predictors. For example, we might want to predict percent body fat using abdomen circumference, or predict the survival time for cancer patients using tumor size. A large body of the literature in biostatistics is devoted to developing statistical methods for predicting the risk of different diseases such as cancer, Alzheimer's disease, diabetes, and Parkinson's disease. Kahn et al. (2009) [13] developed statistical models for finding the risk of diabetes mellitus in US adults age 45 to 64 years using demographic, anthropometric, and clinical risk factors. Little et al. (2008) [17] showed

that statistical methods can be used to identify patients with Parkinson's disease by detecting dysphonia (an impairment in the normal production of vocal sounds). Predicting unknown outcomes and future events using statistical methods can help us with making better decisions. For example, people with high risk of diabetes might decide to follow preventing measures (e.g., diet).

1.2 Sampling

To answer our scientific questions, we would, ideally, study the entire population of interest (e.g., all breast cancer patients). However, this is usually impossible either physically, ethically, or economically. For example, to test the hypothesis about the average normal body temperature, it is not feasible to record the temperature of all healthy people. Instead, a **sample** of representative members is selected from the population. Then with the methods of **statistical inference**, the conclusions based on the sample can cautiously be attributed to the whole population. Mackowiak et al. (1992) selected $n = 148$ people, took their oral temperature, and then made conclusions about the body temperature of the whole population. To compare the effects of different treatments, one of the studies discussed in [27] includes 74 women treated by breast conservation therapy and 67 women treated by mastectomy.

Note that we should generalize our findings only to the population from which the sample is obtained. Mackowiak et al. recruited healthy individuals only. Therefore, their findings can only be applied to healthy people. In the study of different breast cancer treatments discussed above, patients with sever conditions were excluded. Therefore, we should not generalize our findings to this group of patients.

The samples are selected **randomly** (i.e., with some probability) from the population. Unless stated otherwise, these randomly selected members of populations are assumed to be **independent**. Informally, this means that the selected members are not related to each other, and selecting one of them does not affect the selection of another one. In the study by Mackowiak et al., the 148 randomly selected people were unrelated and selected independently from each other.

The selected members (e.g., people, households, cells) are called **sampling units**. In our sample, the individual entities from which we collect information are called **observation units**, or simply **observations**. (We sometimes refer to an observed value of a variable as *observation* and refer to the collection of these observations as *sample*.) In the above example, the sampling units are the same as the observational units. These are the individuals Mackowiak et al. selected from the population in order to measure their body temperature. In some cases, the sampling units are not the same as the observational units; rather, each sampling unit includes multiple observation units. For example, we might take a sample of households and measure some characteristics of the individuals in those households. In this case, the sampling units are households, whereas the observational units are the individuals.

Our sample must be representative of the population, and their environments should be comparable to that of the whole population. For example, to study normal

body temperature of healthy people in the US, our sample would not be a good representative of the population if 80% of people participating in our study are men, or some of the participants are ill, or all the measurements are taken early in the morning. (Body temperature fluctuates over the day.) Using the appropriate sampling techniques (i.e., sampling design) is crucial to making valid conclusions.

1.3 Observational Studies and Experiments

After obtaining the sample, we collect information relevant to our study from the selected members. We typically do this either through an **observational study** or an **experiment**. In observational studies, researchers are passive examiners trying to have the least impact on the events and data collection process. They may simply measure the current values of all relevant characteristics (e.g., body temperature, heart rate, gender) for the sample, or observe how these characteristics change over time.

The study conducted by Mackowiak et al. is an observational study. Not only this study helped them to evaluate their hypothesis (i.e., normal body temperature is less than the commonly accepted value), but it also helped them to detect relationships among characteristics. For example, they found that body temperature and heart rate tend to increase and decrease together. In this case, we say that the two variables are **associated** with each other.

In general, observational studies can help us to discover association, which refers to situations where changes in one characteristic tend to coincide with changes in another one. However, we should not interpret the observed association as **causation**. The relationship is causal if one characteristic *influences* the other one. Unfortunately, it is difficult to establish causality based on observational studies. There is always the possibility that the observed relationship could be due to the effect of some **confounding** (lurking) factors. That is, the effect of the exploratory variable on the response variable is confounded with the effect of other factors, which may or may not be known, so we cannot distinguish the effect of the confounding factor from the effect of the explanatory variable. This usually happens when a confounding factor influences both the explanatory variable and the response variable. For example, we might observe that high consumption of soft drinks is associated with heart disease, and may be tempted to conclude that consumption of soft drinks causes heart disease. However, it is possible that high consumption of soft drinks is associated with a poor diet and eating of fatty foods, which contribute to heart disease. Factors such as diet, age, gender, ethnicity, and genetics are typical confounding factors in many scientific studies.

In general, we attempt to establish causal relationships by using **randomized experiments**, where researchers try to control the process as much as possible. In a randomized experiment, the sampling units (also referred to as **experimental units** or **subjects**) are *randomly* assigned to different *treatments*. For instance, to investigate the effect of dietary consumption on blood pressure, Sacks [28] studied 412 subjects who were randomly assigned to eat either a control diet typical of intake in

the US or the DASH (Dietary Approaches to Stop Hypertension) diet. They found that the DASH diet lowers blood pressure substantially.

Randomization, which refers to the random assignment of subjects to different treatments, is a key concept in designing experiments. It helps control the influence of confounding factors. The assumption is that randomization makes the groups as similar as possible with respect to any possible confounding factor. Then the only difference between these groups is the type of treatment imposed by researchers.

1.4 Data Exploration and Analysis

After selecting the sample and collecting the data, the next step toward statistical inference and decision making is to perform **data exploration**, which involves visualizing and summarizing the data. The objective of data visualization is to obtain a high-level understanding of the observed data. For example, we might realize that the observed values of heart rate in our sample are clustered around 75, and most of them fall within 60 and 90. Using data visualization techniques, we can learn about the **distribution** of a variable. Informally, the distribution of a variable tells us the possible values it can take, the chance of observing those values, and how often we expect to see them in a random sample from the population. Using data visualization, we can also learn about possible relationships between variables. For example, we might find that heart rate increases with high body temperature.

Through data visualization, we might detect previously unknown patterns and relationships that are worth further investigation. Visualization can also help us to identify possible data issues, such as unexpected or unusual measurements, known as **outliers**.

While visualization makes the task of understanding the data easier, the amount of information might still be overwhelming. To make the data more manageable, we need to further reduce the amount of information in some meaningful ways so that we can focus on the key aspects of the data. **Summary statistics** are used for this purpose. For example, the **average** (mean) of observed values is a statistic which is commonly used as a single value representation for the entire sample. It represents typical values we expect to see for a specific variable. Note that the mean of a variable is not the same as its typical values. It is merely a representation of such values. If we find that the average heart rate in our sample is 73.7, we expect most of the observed values for heart rate to be close to (but in general not equal to) 73.7. We will discuss data exploration in more detail in Chaps. 2 and 3.

1.5 Statistical Inference

We collect data on a sample from the population in order to learn about the whole population. For example, Mackowiak et al. (1992) measure the normal body temperature for 148 people to learn about the normal body temperature for the entire

population. More specifically, they wanted to make comments about the average normal body temperature in the whole population. However, since we do not have access to the whole population, the best we can do is to guess its average using the observed data. We say we are **estimating** the unknown population average. We discuss estimation in Chap. 6. Note that the exact value of the population average remains unknown and our estimate of it can change depending on our sample. Therefore, there is always some **uncertainty** associated with our estimations. The same is true when we are estimating the strength of association between two variables (e.g., body temperature and heart rate) or the effect of a treatment on the response variable (e.g., the effect of the DASH diet on blood pressure). In statistics, the mathematical tool to address uncertainty is **probability**. We discuss probability in Chaps. 4 and 5.

The process of using the data to draw conclusions about the whole population, while acknowledging the extent of our uncertainty about our findings, is called **statistical inference**. Our conclusions (the knowledge we acquire from data through statistical inference) allow us to make decisions with respect to the scientific problem that motivated our study and our data analysis. As discussed above, decision problems are sometimes presented in the form of hypothesis testing problems or prediction problems. Chapters 7 to 13 are mainly devoted to statistical inference methods.

1.6 Computation

To perform statistical inference, we usually rely on computer programs to prepare, explore, and analyze the data. Frequently used statistical programs include MINITAB, MATLAB, R, SAS, SPSS, and STATA. Because R [34] (http://www.r-project.org/) is free and arguably the most common software among statisticians, we will be using it in this book. R is readily available for all operating systems and can be installed from http://www.r-project.org/. While R provides a powerful tool for statistical analyses, it requires programming skills. Instead of using R directly, throughout this book, we mainly focus on using R-Commander [7], which is a user-friendly interface created by John Fox for basic practice of statistical methods in R without any programming. However, you still need to install R in order to use R-Commander. Appendix A provides step-by-step instructions for installing R and R-Commander.

1.6.1 Using R-Commander

R-Commander allows us to run basic statistical analysis without necessarily learning the programming language. It can be installed by opening R (double clicking on the R icon to open the *R Console*) and entering the following command:

```
install.packages("Rcmdr", dependencies=TRUE)
```
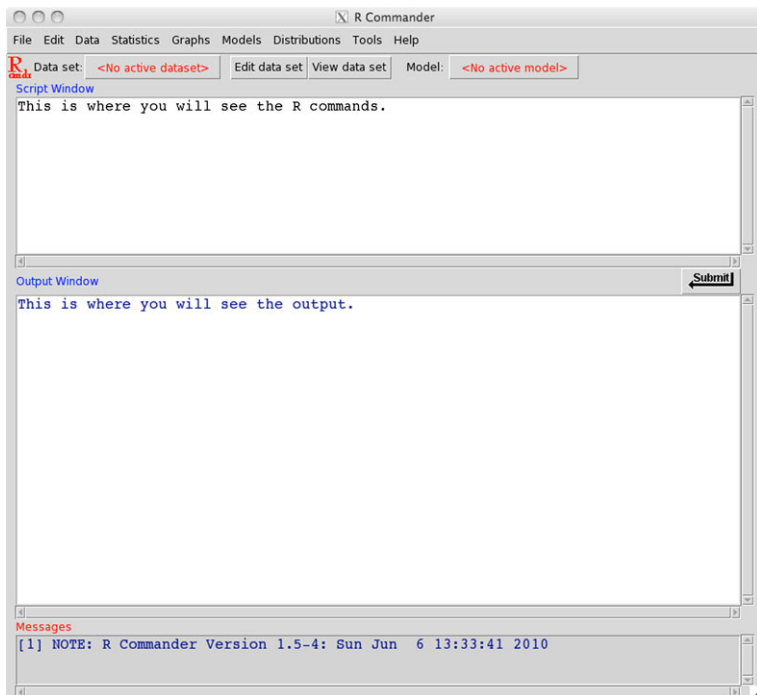
Fig. 1.1   The R-Commander window. Notice the menu bar, `Data set` box, `Edit data set` button, `View data set` button, *Script* window, `Submit` button, and *Output* window

Once R-Commander is installed, it can be used by typing `library(Rcmdr)` in the command line. (More information on how to install R and R-Commander are provided in Appendix A.)

Now take the time to familiarize yourself with the R-Commander window (Fig. 1.1). In the menu bar, there is a `Data set` box that will display the name of the active (current) data set. The subsequent two buttons allow the data set to be edited and viewed, respectively. When commands are executed via the menu bar in R-Commander, their corresponding R codes appear in both the *Script* window and the *Output* window. You can enter these commands directly in the *Script* window or R Console itself and obtain the exact same results as using the menu bar. If you are not interested in programming, however, you can ignore these codes for the most part and focus on the outputs appearing in the *Output* window. In that case, the only commands you need to know are `install.packages` and `library`, which you already used to install and open R-Commander. In general, these commands are used to install and load *R packages*. These are user contributed programs that are created for specific statistical techniques. (For example, the `Rcmdr` package is created to provide a user-friendly interface for routine statistical analysis.)

Many of these packages include interesting data sets related to biology and health sciences. While the statistical techniques discussed in this book are all available in R-Commander, we will occasionally load some additional R packages in order to

Fig. 1.2 Importing the
`Pima.tr` data from the
`MASS` package. Select the
appropriate package and data
set
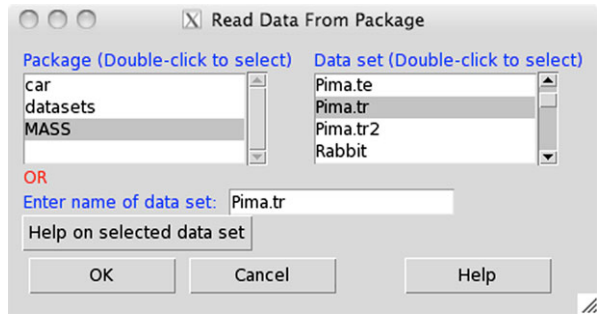
Fig. 1.3 Viewing the
`Pima.tr` data in
R-Commander. Each row
corresponds to a Pima Indian
woman in our sample, and
each column corresponds to a
variable

have access to their data sets. One of these packages is the `MASS` package created
by Venables and Ripley [37]. This package is installed automatically when you in-
stall R, and it is loaded automatically when you run R-Commander.

To see the list of data sets available in `MASS`, in the menu bar click on `Data →`
`Data in packages → Read data set from an attached pack-`
`age`. Under `Package`, you will see the name of several packages (e.g., `car`,
`datasets`, and `MASS`). These packages are loaded automatically when you open
R-Commander.

As an example, we will use the `Pima.tr` data set. Select `Pima.tr` under the
`Data set` option (Fig. 1.2) and click OK. Now `Pima.tr` is the active data set.
Click the `View data set` button to examine it (Fig. 1.3). The data, which are
collected by the US National Institute of Diabetes and Digestive and Kidney Dis-
eases, includes 200 women of Pima Indian heritage living near Phoenix, Arizona
[31]. The women are at least 21 years old and are tested for diabetes. You can learn
more about this data set by clicking on `Data → Active data set → Help`
`on active data set (if available)`.

In the `Pima.tr` data set, each row corresponds to an individual in the sample
and is regarded as one **observation**. Each column corresponds to a characteristic

Fig. 1.4 Importing the `BodyTemperature` data into R-Commander. Enter "BodyTemperature" as the name of the data set. Accept all other defaults
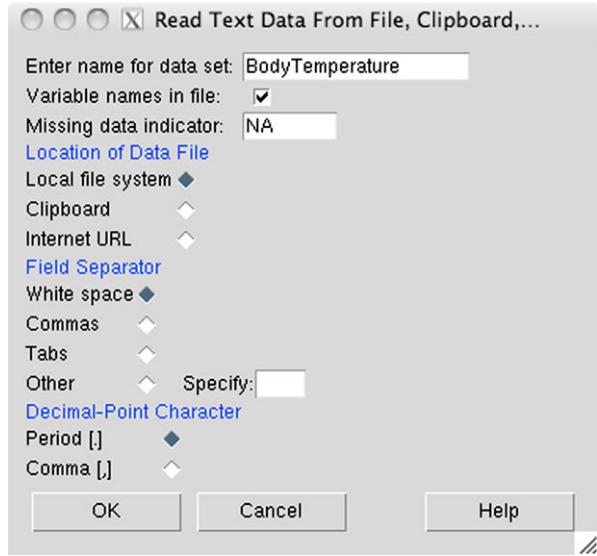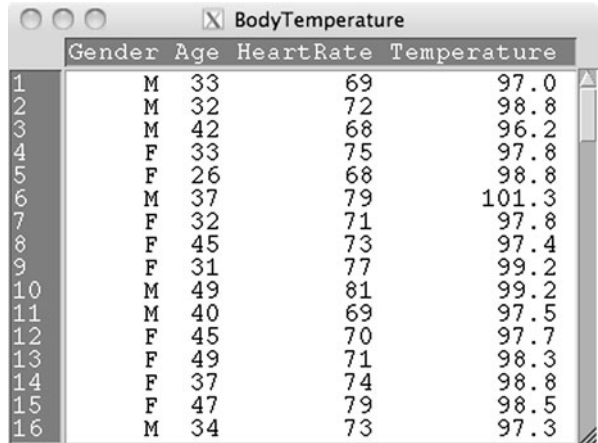
(variable) of interest measured for each individual. The list of these characteristics is as follows:

- `npreg`: number of pregnancies.
- `glu`: plasma glucose concentration in an oral glucose tolerance test.
- `bp`: diastolic blood pressure.
- `skin`: triceps skin fold thickness (mm).
- `bmi`: body mass index.
- `ped`: diabetes pedigree function.
- `age`: age in years.
- `type`: disease status; Yes for diabetic and No for nondiabetic.

In this study, `type` (i.e., disease status) is the variable of interest. Therefore, we refer to it as the response or target variable. The other variables in the study (e.g., `npreg` and `bmi`) are believed to be related to the response variable and may explain its variation (e.g., Yes or No), or they can be used to predict the response variable (e.g., whether a woman would be affected by diabetes). Consequently, we refer to them as the explanatory variables or predictors.

If the data set you would like to analyze is available as a text file, you first need to import it into R-Commander. As an example, download the `BodyTempera-ture.txt` file from book website (http://extras.springer.com) and save it in your local directory. This file includes gender, age, heart rate, and normal body temperature of 100 adults between the age of 20 and 50. In order to read the data set into R-Commander, click `Data → Import data → from text file, clip-board, or URL`. Name the data set "BodyTemperature" and accept all other defaults (Fig. 1.4). Click `OK` and then select the `BodyTemperature.txt` you saved in your local directory. Notice that the name of the active data set changes

Fig. 1.5 Viewing the
`BodyTemperature` data in
R-Commander

	Gender	Age	HeartRate	Temperature
1	M	33	69	97.0
2	M	32	72	98.8
3	M	42	68	96.2
4	F	33	75	97.8
5	F	26	68	98.8
6	M	37	79	101.3
7	F	32	71	97.8
8	F	45	73	97.4
9	F	31	77	99.2
10	M	49	81	99.2
11	M	40	69	97.5
12	F	45	70	97.7
13	F	49	71	98.3
14	F	37	74	98.8
15	F	47	79	98.5
16	M	34	73	97.3

to `BodyTemperature`. You can always switch back to `Pima.tr` by clicking on
the name of active data sets. For now, click the `View data set` button to exam-
ine `BodyTemperature` (Fig. 1.5). The data set includes 100 observations and 4
variables.

When importing `BodyTemperature`, we kept all the options (shown in
Fig. 1.4) as their default values except the name of the data set. We kept the option
`Variable names in file` checked since the first row of the data includes
the names of variables. If this is not the case for your data, you should uncheck this
option. This way, R-Commander will choose generic names for the variables. The
default value for `Missing data indicator` is "NA". This means that in the
data set, "NA" (Not Available) is recorded whenever the value of a variable is miss-
ing (not known). If missing values are identified by a different indicator, you should
change this option accordingly. Since the `BodyTemperature.txt` was saved as
a text file in our local directory, we kept the option `Local file system` under
`Location of Data file`. We always recommend this option. However, we
can also import the data after copying it to clipboard and choosing the `Clipboard`
option. If the data set is available online, we can import it directly by choosing
the `Internet URL` option and specifying the internet address when prompted.
When reading a data set into R-Commander, it is very important to specify the
`Field separator` correctly. This refers to the character that separates the vari-
ables (columns) in the data. For `BodyTemperature`, the variables are separated
by white spaces, which is the default value for this option.

1.6.2 Using R

While R-Commander can be conveniently used for basic application of statistical
methods, using R directly (i.e., using the R programming language) allows for more
control over the analysis and leads to a deeper understanding of statistical methods.

In Appendix B, we provide a brief introduction to R programming for those who are interested in learning R. Also, in the "Advanced" section of most chapters, we discuss some R programming techniques related to the topics covered in the corresponding chapter.

1.7 Advanced

In the Advanced section at the end of each chapter, we discuss some topics that are intended for more advanced readers and can be disregarded in the first reading of this book. Here, we provide more discussion on sampling, observational studies, and experiments.

1.7.1 More on Sampling

The sampling strategy (a.k.a. sampling design) is an important factor affecting the results of scientific studies. Here, we briefly discuss some of the most widely used sampling designs.

Simple Random Sampling   **Simple random sampling** (SRS) is the most straight forward sampling procedure. Suppose that the population of interest has N members, and we want to select n of them for our sample. If the chance of being selected is the same for any group of n members in the population, we refer to the sampling strategy as simple random sampling. For example, we could assign a unique number $1, \ldots, N$ to each member of the population and randomly select n of these numbers. For this, we could write the numbers on pieces of paper (equal sizes), put them in a hat and select n without looking, or use random number generator computer programs to generate n distinct numbers from 1 through N. (The latter approach is obviously more feasible when the population size N is large.)

Stratified Sampling   Suppose that we want to find average normal body temperature in the US population. Some studies have shown that body temperature varies between different races. In the study conducted by McGann et al. [20], African-Americans had higher average body temperature compared to Caucasians. To make sure our findings based on the sample are generalizable to the whole population, we should make sure that our sample is comparable to the whole population with respect to the key subpopulations. In this case, we should make sure that neither of these two subpopulations, African-Americans and Caucasians, is overrepresented in our sample. Suppose that 72% of the US population are Caucasian, 12% are African-American, and the remaining 16% belong to other races. If we intend to select $n = 150$ people for our study, we can randomly sample $150 \times 0.72 = 108$ from the Caucasian population, $150 \times 0.12 = 18$ from the African-American population,

and $150 \times 0.16 = 24$ people from other races. This way, our sample would be comparable to the whole population with respect to the proportions of different races. This is an example of stratified random sampling. In this approach, the population is first partitioned into subpopulations, a.k.a. **strata**, and sampling (usually simple random sampling) is performed separately within each subpopulation. In the above example, we stratify the population by race.

Cluster Sampling   Suppose that we want to find the average length of stay (LOS) in hospital for patients suffering from acute appendicitis from 2009 to 2010 in the US. Sampling directly from the population of patients could be difficult; we might not have access to the list of all patients treated during 2009–2010. Instead, we can sample from the population of all hospitals in the US, and for each hospital, we can subsample some or all (e.g., using SRS) of the appendicitis patients admitted to that hospital. Note that in this case, the observation units (patients) are different from the sampling units (hospitals). This sampling design is known as **cluster sampling**. We start the sampling process by first grouping observations units into **clusters**. Then, we sample from these clusters and subsample some or all members of the selected clusters. When analyzing data from clustering sample, we should take the clustering of the observed data into account. In the above example, it is perceivable that observations coming from the same hospital are more similar compared to observations from different hospitals. For example, a hospital may tend to keep patients longer. As a result, patients sampled from that hospital would have relatively higher LOS compared to patients sampled from a hospital whose policy is to release patients as soon as possible.

1.7.2 More on Observational Studies

Observational studies can be classified into **retrospective** and **prospective** studies. In retrospective studies, researchers look into the histories of the participants. For example, to investigate the effect of smoking on lung cancer, a group of patients with lung cancer can be surveyed to determine if they smoked in the past. In **prospective** studies, the researchers identify different groups and observe them over time (without disturbing and influencing the natural processes of the event). For example, we might observe a sample of smokers over time to see what percentage of them would develop lung cancer.

In general, conducting prospective studies is more difficult compared to retrospective studies, since we might have to follow participants for a long time. However, they tend to produce more reliable data compared to retrospective studies (e.g., people might not remember their past very accurately). In the above example, our data might not be very reliable if we ask smokers how long they have been smoking.

In the above example, to make a reasonable conclusion about the relationship between smoking and lung cancer, the smoking habits of patients should be compared to that of another group who do not have lung cancer but are similar in all other

respects (age, gender, etc.). To this end, we can select a sample of patients along with a sample of people who do not have lung cancer and investigate the smoking habits of all participants (with or without lung cancer). We refer to the group of patients as the **case** group. The group of participants without lung cancer is called the **control** group. If there is a substantial difference between the two groups in terms of smoking (assuming that everything else is similar between the two groups), we may conclude that lung cancer is related to smoking. This type of study is called a **case-control** study.

In the above example, we assume that the individuals in the cases group are not related to those in the control group. In some case-control studies, however, we pair each individual in the control group with a related individual from the case group. For example, instead of randomly sampling a group of people without lung cancer, we can pair each lung cancer patient in the case group with a sibling who is not suffering from lung cancer. This way, we hope to make the two groups as comparable as possible (especially, with respect to hereditary factors). Note that in these situations, the usual assumption that samples are independent does not hold anymore, and the pairing of observations should be into account in our data analysis. We will discuss these situations in Chap. 8.

1.7.3 More on Experiments

We mentioned that **randomization** is a key concept in designing experiments. Another key concept is **replication**, which refers to the assignment of multiple subjects to each treatment. Replications allow us to observe the variability of treatment effects. For example, if we are interested in investigating the effect of aspirin on heart attack, we could randomly assign some subjects to the treatment group, who would take aspirin regularly, and some subjects to the control group, who would receive the *placebo*, which is similar to the actual drug but missing the main or active ingredients. Of course, the effect of treatment would not be the same for all subjects; not everyone taking aspirin will become immune from heart disease, and not everyone taking placebo would suffer from heart disease. In the study conducted by Sack et al., not everyone following the DASH diet would have a lower blood pressure than those in the control group. The observed variability in the response variable (e.g., disease status, blood pressure) contributes to our uncertainty regarding the treatment effect. In a proper statistical inference, the extent of our uncertainty should be expressed along with our conclusion regarding the effectiveness of the treatment. The simplest design using randomization and replication is called a **complete randomized design**.

In experiments, we want the subjects assigned to different treatments to be as comparable as possible, so the only difference between them would be the type of treatment they receive. In some experiments, however, the treatment groups might be different (due to chance) with respect some factors that are known to influence the results but are not the main focus of the study. We refer to them as **nuisance**

factors. For example, when studying the effect of aspirin on heart disease, age might be an important factor but is not of main interest. It is quite possible that when we randomly assign subjects to one of the two treatments, they might not be comparable with respect to age (e.g., participants in one group tend to be older). To avoid this issue, a common approach is to use **blocking**, which refers to the division of subjects into subgroups such that subjects within a block are considered to be similar in terms of the nuisance factors. The design of such experiments is called **randomized block design**. In these experiments, subjects are first divided into blocks, and then randomization (assigning subjects to different treatments randomly) is performed within each block. For the above example, we could first group the subjects into several age blocks (e.g., below 55, 55–65, and above 65) and then randomly assign the subjects within each block to one of the possible treatments (i.e., aspirin or placebo).

A common randomized block design is when each block is comprised of a pair of related subjects. For example, each block may include two siblings, who are randomly assigned to one of the two possible treatments (e.g., aspirin or placebo). Such design is referred to as **matched pairs design**. Occasionally, the pairs that create a block are the same subject, who receives both treatments one after each other. (The order of treatments are decided randomly.) For example, we can select a sample of people from the population, assign each person to one of the two possible diets (e.g., high sodium vs. low sodium) for one week, measure the blood pressure, then switch them to the other diet for another week, and measure their blood pressure again. This way, each person acts as his or her own control, so the two treatment groups are quite comparable. Such experiments are known as **crossover** experiments.

When assigning subjects to different treatments, the researchers may not tell the subjects their assignments to avoid prejudice and unintentional influence on the results. These experiments are called **single-blind** studies. The assignments of subjects may be hidden from the researchers (to avoid bias) as well as subjects. These experiments are **double-blind**.

1.7.4 Cross-Sectional, Longitudinal, and Time Series Data

In some studies, we collect data at some fixed time. We refer to such data, which represent a snapshot of our samples, as **cross-sectional** data. In some other studies, we follow the samples over time and repeatedly collect information and take measurements. The resulting data are called **longitudinal** data. These studies tend to be more complex but can provide a better understanding of patterns and relationships in the population. They are especially helpful for understanding how these patterns and relationships change over time. **Time series** data are also collected over a period of time. However, compared to longitudinal data, time series data are usually collected more frequently, but on smaller samples (e.g., one or two individuals).

1.8 Exercises

1. Read the paper entitled "A Critical Appraisal of 98.6°F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich" by Mackowiak et al. [19]. What is the scientific question that motivated this study? Comment on the type of study, its sampling design, and its findings. (The paper is available online at http://jama.ama-assn.org/cgi/reprint/268/12/1578.)

2. Read the paper entitled "Chocolate consumption in relation to blood pressure and risk of cardiovascular disease in German adults" by Buijsse et al. [4]. What is the objective of this study? Comment on the type of the study, its sampling design, and its findings. Could we use this study to conclude that chocolate consumption reduces the risk of cardiovascular disease? (This paper is available online at http://eurheartj.oxfordjournals.org/content/early/2010/03/18/eurheartj.ehq068.abstract.)

3. In another study, Taubert et al. [33] also studied the relationship between chocolate and blood pressure. Read their paper entitled "Effects of Low Habitual Cocoa Intake on Blood Pressure and Bioactive Nitric Oxide". (This paper is available online at http://jama.ama-assn.org/content/298/1/49.full.) Compare this study to the study [4]. Comment on the advantages and disadvantages of each study.

4. Read the paper entitled "A July Spike in Fatal Medication Errors: A Possible Effect of New Medical Residents" by Phillips et al. [26]. What is the scientific question? Comment on the type of study, the sample that was used in this study, and the conclusion. What kind of decision can be made based on the findings of this study? (The paper is available online at http://www.springerlink.com/content/n502614282p9266t.)

5. In the paper entitled "The Role of Estrogen in Schizophrenia", Seeman [30] reviewed three different studies related to the role of estrogen in Schizophrenia. Provide a summary for each of the three studies she reviewed and comment on the type of these studies, their samples, and their findings. (The paper is available online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1188751.)

6. Read the paper entitled "An Acute Effect of Cigarette Smoking on Platelet Function: A Possible Link Between Smoking and Arterial Thrombosis" by Levine [16]. What is the scientific question this study attempts to answer? Comment on the approach used in the study to answer the scientific question. (This paper is available online at http://circ.ahajournals.org/cgi/reprint/48/3/619.)

7. Kettunen et al. [14] study the effect of arthroscopy in patients with chronic patellofemoral pain syndrome. Discuss their approach, the sample they used, and their findings. What would be a reasonable decision based on the results of their study? (This paper is available online at http://www.biomedcentral.com/1741-7015/5/38.)

8. In an article published in the July 2010 issue of the journal of Pediatrics, Nafiu et al. [23] argue that measuring children's neck circumference could provide a

simple way to identify possible weight problems. Read the report by Reuters (http://www.reuters.com/article/idUSTRE6653R320100706) about this study and comment on its objective, its sampling design, and its findings. (The full article is available online at http://pediatrics.aappublications.org/cgi/content/abstract/peds.2010-0242v1.)

9. Read the article entitled "Caloric restriction improves memory in elderly humans" by Witte et al. [39]. This article was published in January 2009 in the Proceedings of the National Academy of Sciences (PNAS). Comment on the scientific question that motivated this study, the population of interest, and how the study was designed. What would be a reasonable decision based on the results of this study? (This paper is available online at http://www.pnas.org/content/106/4/1255.full.)

10. In an article entitled "Thought for Food: Imagined Consumption Reduces Actual Consumption", which was published in Science in December 2010, Morewedge et al. [21] study the effect of imaginary eating on the actual consumption of imagined food. (The article is available online at http://www.sciencemag.org/content/330/6010/1530.full.html.) Read this article and provide a summary of their study design.

Chapter 2
Data Exploration

2.1 Data Visualization and Summary Statistics

After clearly defining the scientific question we try to answer, selecting a set of representative members from the population of interest and collecting data (either through observational studies or randomized experiments), we usually begin our analysis with data exploration. This chapter focuses on data exploration for one variable at a time. (Data exploration techniques aimed at identifying possible relationship between two or more variables are discussed in the next chapter.) Our objective is to develop a high-level understanding of the data, learn about the possible values for each characteristic, and find out how a characteristic varies among individuals in our sample. In short, we want to learn about the *distribution* of variables. Recall that for a variable, the distribution shows the possible values, the chance of observing those values, and how often we expect to see them in a random sample from the population.

The data exploration methods allow us to reduce the amount of information so that we can focus on the key aspects of the data. We do this by using data visualization techniques and summary statistics. The visualization techniques and summary statistics we use for a variable depend on its type. Therefore, before we continue with data exploration methods, we briefly discuss different variable types. (More discussion is provided in Chap. 4.)

2.2 Variable Types

Let us revisit the `Pima.tr` data discussed in the previous chapter (Fig. 2.1). For each individual, there are eight measurements for eight different variables. In this book, variables will be represented by capital letters, such as X, Y, Z. Each observation in our sample has an index i, where $i = 1, 2, \ldots, n$, and n is the total sample size. Here, the term *observation* refers to an observed value of a variable, and the term *sample* refers to the collection of these observations. We denote by x_i the ith

Fig. 2.1 Viewing the
`Pima.tr` data in
R-Commander

	npreg	glu	bp	skin	bmi	ped	age	type
1	5	86	68	28	30.2	0.364	24	No
2	7	195	70	33	25.1	0.163	55	Yes
3	5	77	82	41	35.8	0.156	35	No
4	0	165	76	43	47.9	0.259	26	No
5	0	107	60	25	26.4	0.133	23	No
6	5	97	76	27	35.6	0.378	52	Yes
7	3	83	58	31	34.3	0.336	25	No
8	1	193	50	16	25.9	0.655	24	No
9	3	142	80	15	32.4	0.200	63	No
10	2	128	78	37	43.3	1.224	31	Yes
11	0	137	40	35	43.1	2.288	33	Yes
12	9	154	78	30	30.9	0.164	45	No
13	1	189	60	23	30.1	0.398	59	Yes
14	12	92	62	7	27.6	0.926	44	Yes
15	1	86	66	52	41.3	0.917	29	No
16	4	99	76	15	23.2	0.223	21	No
17	1	109	60	8	25.4	0.947	21	No
18	11	143	94	33	36.6	0.254	51	Yes

observed value of variable X. For example, if the variable `age` is denoted by X, then $x_5 = 23$ means that the 5th individual in our sample is 23 years old. (Try checking this by viewing the `Pima.tr` data set.)

Based on the values a variable can take, we can classify it into one of two groups: **numerical** variables or **categorical** variables. In `Pima.tr`, variables `npreg`, `age`, and `bmi` in the `Pima.tr` data set are numerical variables since they take numerical values, and the numbers they take have their usual meaning. For example, we say that the second individual in our sample is older than the first individual since $x_2 = 55$ is bigger than $x_1 = 24$. We can also subtract their ages to find their age difference: $55 - 24 = 31$. For numerical variables, we can talk about the distance between two values.

If the values of a numerical variable are *counts* (e.g., number of pregnancies, number of physician visits), we refer to the variable as a **count variable** to distinguish it from other types of numerical variables. Often, the statistical methods we choose for count variables are different from the method we choose for other numerical variables.

The `type` variable in `Pima.tr` is categorical since the set of values it can take consists of a finite number of categories; here, `Yes` (for diseased) and `No` (for nondiseased). In other words, a categorical variable assigns one of the possible categories to each individual in our sample.

It is common to use numerical codings for categorical variables. Let us denote the `type` variable Y. We can use $Y = 1$ for nondiabetic individuals (i.e., `type=No`), and $Y = 2$ for diabetic women (i.e., `type=Yes`). Note, however, that these numbers merely represent different categories (disease status) and do not have their usual meaning. For example, we cannot talk about the distance between two values of the `type` variable or say that the value of this variable for diabetic women is two times more than that of nondiabetic women. Indeed, the assignment of numbers to different categories in this case is quite arbitrary. For the `type` variable, we could have decided to represent diabetics by $Y = 1$ and nondiabetics by $Y = 2$.

Fig. 2.2 Viewing the
`birthwt` data in
R-Commander

	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
85	0	19	182	2	0	0	0	1	0	2523
86	0	33	155	3	0	0	0	0	3	2551
87	0	20	105	1	1	0	0	0	1	2557
88	0	21	108	1	1	0	0	1	2	2594
89	0	18	107	1	1	0	0	1	0	2600
91	0	21	124	3	0	0	0	0	0	2622
92	0	22	118	1	0	0	0	0	1	2637
93	0	17	103	3	0	0	0	0	1	2637
94	0	29	123	1	1	0	0	0	1	2663
95	0	26	113	1	1	0	0	0	0	2665
96	0	19	95	3	0	0	0	0	0	2722
97	0	19	150	3	0	0	0	0	1	2733
98	0	22	95	3	0	0	0	1	0	2751
99	0	30	107	3	0	1	0	1	2	2750
100	0	18	100	1	1	0	0	0	0	2769
101	0	18	100	1	1	0	0	0	0	2769
102	0	15	98	2	0	0	0	0	0	2778
103	0	25	118	1	1	0	0	0	3	2782

Categorical variables are either **nominal** or **ordinal**, depending on the extent of information the numerical coding provides. For nominal variables, the numbers are simply labels, which are chosen arbitrarily. Therefore, they do not provide any information. The `type` variable in `Pima.tr` is nominal. For ordinal variables, although the numbers do not have their usual meaning, they preserve a rank ordering. Therefore, they provide information about the ordering of categories. For example, we would use an ordinal variable to denote the severity of a disease as $Y = 1$ for low, $Y = 2$ for medium, and $Y = 3$ for high. Although these numerical values do not suggest that medium is two times more severe than low, we can say that medium is more severe than low.

Now let us consider another data set called `birthwt`, which is also available from the `MASS` package. This data set includes the birth weight (in grams) of 189 newborn babies along with some characteristics (e.g., age, smoking status) of their mothers. The data were collected at Baystate Medical Center, Springfield, MA, during 1986. To load this data set, click `Data → Data in packages → Read data set from an attached package`. Select `MASS` under `Package` and `birthwt` under `Data set`.

View the data set by clicking the `View data set` button (Fig. 2.2). The data set includes the following variables:

- `low`: indicator of birth weight less than 2.5 kg ($0 =$ normal birth weight, $1 =$ low birth weight).
- `age`: mother's age in years.
- `lwt`: mother's weight in pounds at last menstrual period.
- `race`: mother's race ($1 =$ white, $2 =$ African-American, $3 =$ other).
- `smoke`: smoking status during pregnancy ($0 =$ not smoking, $1 =$ smoking).
- `ptl`: number of previous premature labors.
- `ht`: history of hypertension ($0 =$ no, $1 =$ yes).
- `ui`: presence of uterine irritability ($0 =$ no, $1 =$ yes).
- `ftv`: number of physician visits during the first trimester.
- `bwt`: birth weight in grams.

Variables `age`, `lwt`, `ptl`, `ftv`, and `bwt` are numerical variables. Among these variables, `ptl` and `ftv` are count variables. The variables `low`, `race`, `smoke`, `ht`, and `ui` are all categorical. Note that all categorical variables are coded with numerical values. In these situations, R and R-Commander cannot automatically recognize them as categorical variables. In fact, they are considered as numerical variables by default. Therefore, we need to convert them to categorical variables. To do this, make sure `birthwt` is the active data set, then click on `Data` → `Manage variables in active data set` → `Convert numeric variables to factors`. (In R, categorical variables are usually stored as *factors*.) Under `Variables`, select `low`, `race`, `smoke`, `ht`, and `ui`. Under `Factor Levels`, check the `Use numbers` option (unless you would like to provide specific names for each category). Click OK and accept the overwrite option when prompted. The data set is now ready for exploration and analysis.

2.3 Exploring Categorical Variables

In this section, we discuss visualizing and summarizing categorical data. Consider the `type` variable in `Pima.tr` data set. A simple way for summarizing the data is to create a table that shows the number of times each category has been observed.

> The number of times a specific category is observed is called **frequency**. We denote the frequency for category c by n_c.

Table 2.1 shows that in this sample, the number of women not affected by diabetes (`type=No`) is $n_1 = 132$, and the number of diabetic (`type=Yes`) women is $n_2 = 68$. Here, 1 represents "No", and 2 represents "Yes" for the `type` variable. To obtain the frequencies for this variable, click `Statistics` → `Summaries` → `Frequency distributions` and select `type` as the `Variable`. The results are displayed in the *Output* window (Fig. 2.3).
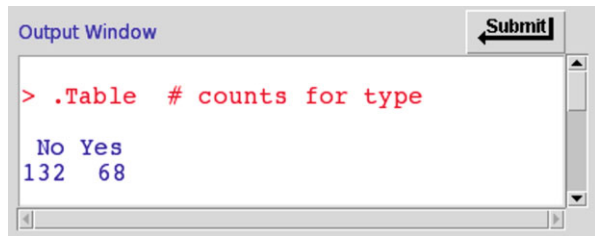
The sum of the frequencies for all categories is equal to the total sample size,

$$\sum_c n_c = n,$$

Table 2.1 Frequency table for the `type` variable in the `Pima.tr` data set

Type	Frequency
No	132
Yes	68
Total	200

Fig. 2.3 Using
R-Commander to obtain and
view the frequency table for
`type` from the `Pima.tr`
data set

```
Output Window                                    Submit

> .Table  # counts for type

 No Yes
132  68
```

where $\sum_c$ means the sum over all categories. For the `type` variable, we have

$$\sum_c n_c = n_1 + n_2 = 132 + 68 = 200.$$

2.3.1 Relative Frequency and Percentage

Follow the above steps to create the frequency table for the `race` variable in the
`birthwt` data set. For this variable, the frequencies are $n_1 = 96$, $n_2 = 26$, and
$n_3 = 67$ for "White", "African-American", and "Other" categories, respectively. The
sum of these frequencies is equal to the sample size $n = 189$.

Now suppose that we want to ensure that the racial make up of our sample is
similar to that of the whole US population. To do this, we use **relative frequencies**
or **percentages** as summary statistics.

> The relative frequency is the sample proportion for each possible category.
> It is obtained by dividing the frequencies n_c by the total number of observa-
> tions n:
>
> $$p_c = \frac{n_c}{n}. \tag{2.1}$$
>
> Relative frequencies are sometimes presented as percentages after multiplying
> proportions p_c by 100.

The relative frequencies and percentages for the `race` variable in `birthwt` are

$$p_1 = 96/189 = 0.508 = 50.8\%,$$
$$p_2 = 26/189 = 0.138 = 13.8\%,$$
$$p_3 = 67/189 = 0.354 = 35.4\%.$$

Therefore, 50.8% (almost half) of the women in the sample were white, 13.8% were
African-American, and the remaining 35.4% were from other races. We can now
compare these relative frequencies with their corresponding proportions in the US
population.

```
Output Window                                              ←Submit|

 1   2   3
96  26  67

> 100*.Table/sum(.Table)   # percentages for race

        1          2          3
50.79365  13.75661  35.44974
```

Fig. 2.4 Using R-Commander to obtain and view the frequencies and percentages of the `race` variable in the `birthwt` data set

In R-Commander, make sure `birthwt` is the active data set, then click `Statistics` → `Summaries` → `Frequency distributions`, and select `race` as the `Variable`. The frequencies and percentages are given in the *Output* window, as shown in Fig. 2.4. Note that R-Commander automatically multiplies the proportions by 100 to obtain the percentages.

For `race`, the category "1" (i.e., white women) has the highest frequency. In this case, we say that the **mode** of the variable `race` is "1".

> For a categorical variable, the mode of is the most common value, i.e., the value with the highest frequency.

For the `type` variable, if we use 1 for "No" (i.e., nondiabetic) and 2 for "Yes" (i.e., diabetic), the mode of the variable is 1.

Since the relative frequencies are proportions of the sample size, their sum is 1,

$$\sum_c p_c = 1,$$

where p_c is the relative frequency of category c. For the `race` variable, we have

$$\sum_c p_c = 0.508 + 0.138 + 0.354 = 1.$$

Similarly, the sum of the percentages for different categories is 100%. Table 2.2 shows the frequencies and relative frequencies of the three categories for `race`.

2.3.2 Bar Graph

For categorical variables, **bar graphs** are one of the simplest ways for visualizing the data. Using a bar graph, we can visualize the possible values (categories) a categorical variable can take, as well as the number of times each category has been

Table 2.2 Frequency table for the `race` variable in the `birthwt` data set

Race	Frequency	Relative frequency
White	96	0.508
African-American	26	0.138
Other	67	0.354
Total	189	1

Fig. 2.5 Using R-Commander to create and view a frequency bar graph for `type` in the `Pima.tr` data set. The heights of the bars sum to the sample size n. Overall, bar graphs show us how the observed values of a categorical variable in our sample are distributed
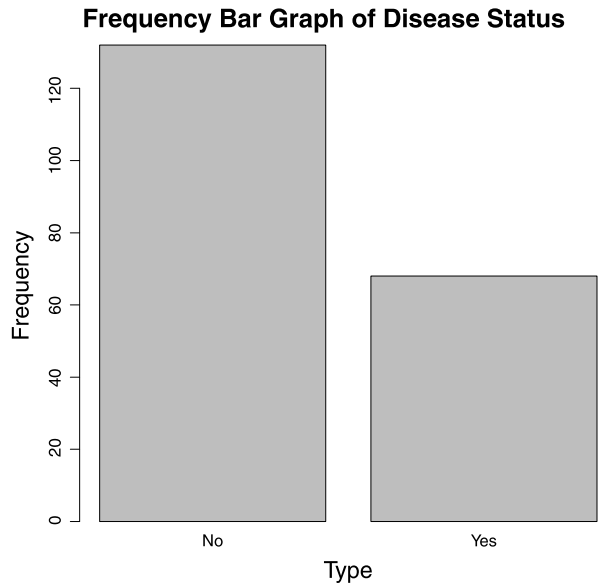
observed in our sample. The bar graph for variable `type` (Fig. 2.5) shows that the possible values are "No" (nondiseased) and "Yes" (diseased). The height of each bar in this graph shows the frequency of the corresponding category. Therefore, the bar heights (frequencies) add up to the total sample size (in this case, $n = 200$).

In R-Commander, make sure `Pima.tr` is the active data set. (If you have loaded `Pima.tr`, but it is not currently the active data set, click on the name of the active data set and select `Pima.tr` from the list of available data sets.) Then, create a bar graph for `type` by clicking `Graphs → Bar graph` and then selecting `type` as the `Variable`. (Notice how bar graphs can only be created for categorical variables.) On the resulting plot shown in Fig. 2.5, the horizontal axis represents the possible values of the variable, and the height of each bar represents the number of observations in that category. Indeed, a quick glance at the graph reveals that the number of nondiabetic women in our sample is almost two times more than the number of diabetic women. You can save this graph by clicking `Graphs → Save graph to file` and choosing either `as bitmap` or `as PDF/Postscrip/EPS` for the file format.

Fig. 2.6 Bar graph for
mother's race in the
`birthwt` data set, where *1*,
2, and *3* represent the
categories "white",
"African-American", and
"other", respectively

Frequency Bar Graph of Race
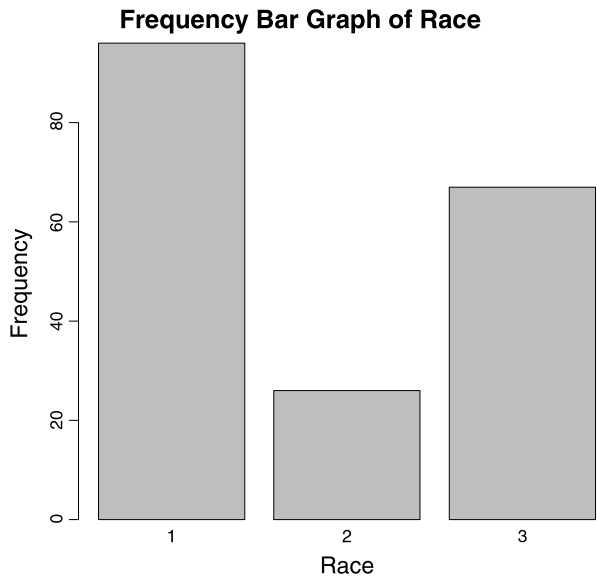
Fig. 2.7 Pie charts for the
`type` variable from
`Pima.tr` and the `race`
variable from `birthwt`,
where *1*, *2*, and *3* represent
the categories "white",
"African-American", and
"other", respectively
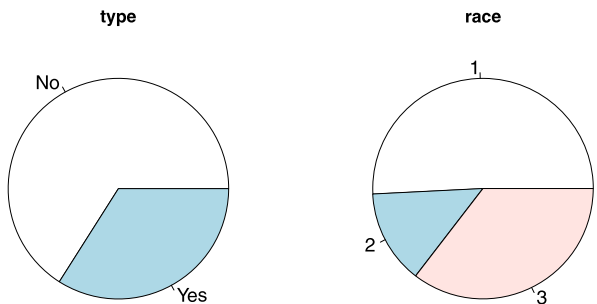
Follow the above steps to create the bar graph for the variable `race` in
`birthwt`. The resulting graph is shown in Fig. 2.6.

2.3.3 Pie Chart

We can use a `pie chart` to visualize the relative frequencies of different cate-
gories for a categorical variable. In a pie chart, the area of a circle is divided into
sectors, each representing one of the possible categories of the variable. The area of
each sector *c* is proportional to its frequency. To create pie charts in R-Commander,
click Graphs → Pie chart. Figure 2.7 shows the pie charts for the `type` vari-
able from `Pima.tr` and the `race` variable from `birthwt`.

Fig. 2.8 Three separate samples for variable X. Observations in Sample 1 are gathered around 2, whereas observations in Sample 2 and Sample 3 are gathered around 4. Observations in Sample 3 are more dispersed compared to those in Sample 1 and Sample 2
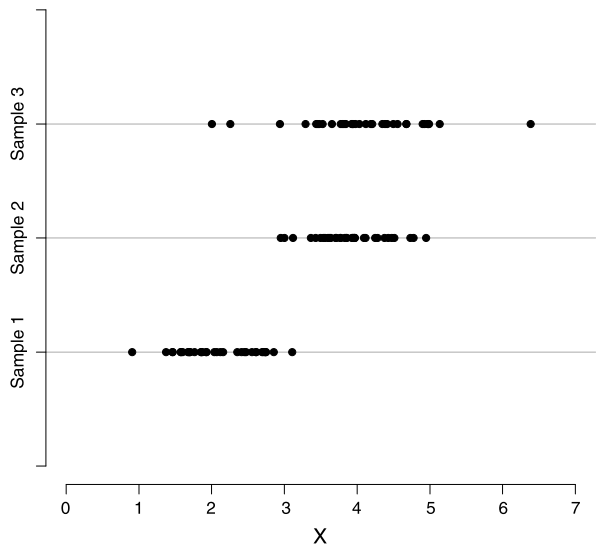
2.4 Exploring Numerical Variables

In this section, we discuss visualization and summarization of numerical data. As a running example, we consider a numerical variable, X, for which we have collected three sets (samples) of observations denoted as Sample 1, Sample 2, and Sample 3. (You can assume that each set of observations are collected from a distinct group in the population.) Figure 2.8 shows the **dot plots** for these three sets of observations. Here, each point represents one observation in the corresponding sample.

As before, we use data visualization techniques and summary statistics to learn about the distribution of variables. For numerical variables, we are especially interested in two key aspects of the distribution: its **location** and its **spread**. The location of a distribution refers to the *central tendency* of values, that is, the point around which most values are gathered. The spread of a distribution refers to the *dispersion* of possible values, that is, how scattered the values are around the location. In Fig. 2.8, we can see that the observed values in Sample 1 are gathered around $X = 2$; whereas, the observations in Sample 2 and Sample 3 are gathered around $X = 4$. Therefore, Sample 2 and Sample 4 have roughly the same location. On the other hand, Sample 1 and Sample 2 have roughly the same spread, which is smaller than the spread in Sample 3. The individual observations in Sample 3 tend to be further away from the location compared to those in Sample 1 and Sample 2. This might not be very clear from dot plots, where we show all the observed values. In what follows, we present more effective visualization techniques and summary statistics that reduce the amount of information in order to make it easier to learn about the distribution of numerical variables.
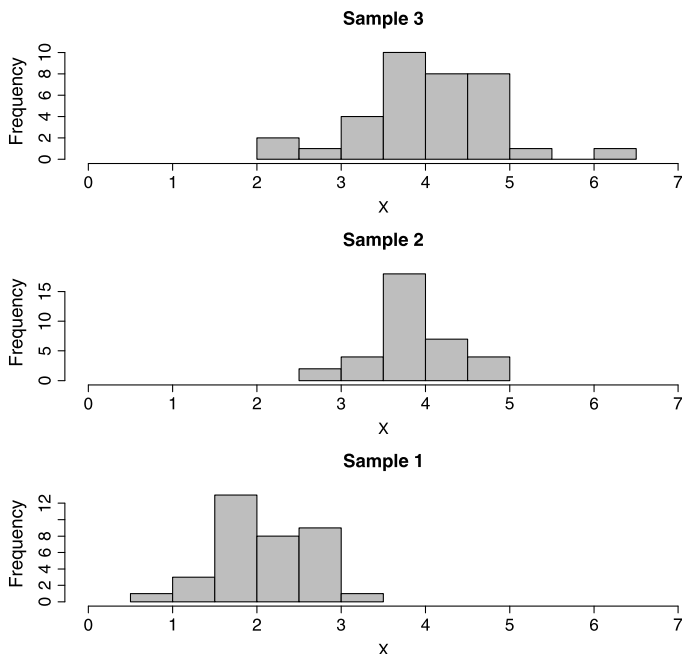
Fig. 2.9　Histograms for the three samples shown in Fig. 2.8

2.4.1 Histograms

Histograms are commonly used to visualize numerical variables. A histogram is similar to a bar graph after the values of the variable are grouped (binned) into a finite number of intervals (bins). For each interval, the bar height corresponds to the frequency (count) of observation in that interval. That is, we treat each interval as a category. Similar to bar graphs, the heights sum to sample size *n*. Figure 2.9 shows the histograms for Sample 1, Sample 2, and Sample 3. For Sample 1, observations are grouped into six intervals. Most observed values are around 2. Sample 2 and Sample 3 have roughly the same locations. However, the histogram for Sample 3 is more spread out compared to that of Sample 2.

As an example, we use the variable `bmi` in the `Pima.tr` data set and create its histogram. In R-Commander, click `Graphs` → `Histogram` and select `bmi` for the `Variable`. (Now we can only select from the numerical variables in our data set.) The resulting histogram is shown in Fig. 2.10. The *x*-axis represents `bmi`, where its observed values are divided into seven equal bins of width $w = 5$. The height of each bar shows the frequency (count) in the corresponding interval. Indeed, a quick glance of the plot suggests that the age interval (30, 35] has the highest frequency. The notation (30, 35] is the interval greater than 30 and less than or equal to 35. By default, each interval includes the right-hand point (here, 35) but not the left-hand point (here, 30). For the `bmi` variable, Fig. 2.11 shows that most observations are gathered around 32.5, and the observed values spread roughly from 15

Fig. 2.10 The frequency histogram for the numerical variable `bmi` in the `Pima.tr` data set. The height of the rectangles represent the frequency of the interval and sum to the total sample size n. Here, the values of the variable are divided into seven bins

Frequency Histogram of BMI
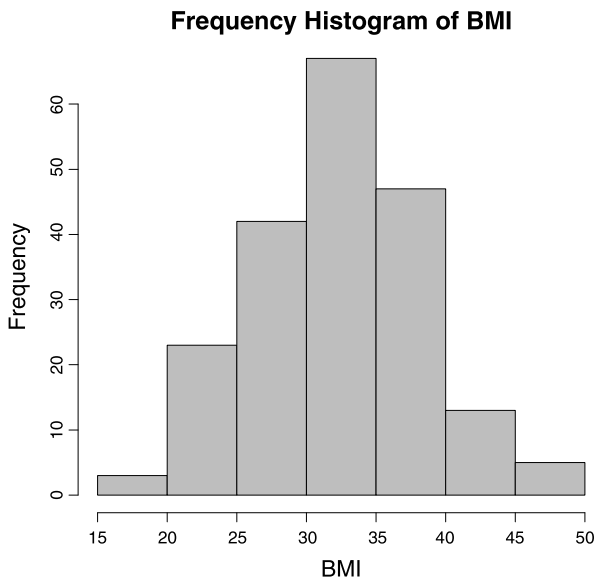
Fig. 2.11 The density histogram for `bmi` from the `Pima.tr` data set. Here, the scale on the y-axis is density (not frequency). Once again, the values of `bmi` are divided into seven bins of width $w = 5$

Density Histogram of BMI
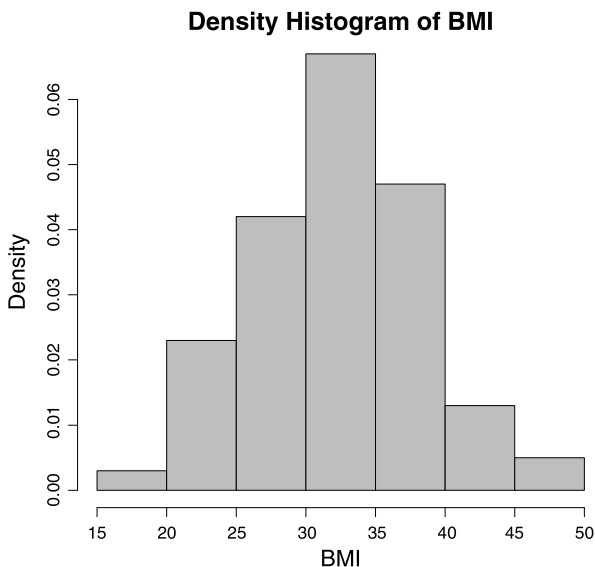
to 50. (Later, we use summary statistics to describe these features of data more precisely.) As before, you can save this graph by clicking `Graphs` → `Save graph to file` and choosing either `as bitmap` or `as PDF/Postscrip/EPS` for the file format.

In the above example, the bar height for each interval, c, is equal to its frequency, n_c. Alternatively, the bar height for each interval could be set to its relative frequency $p_c = n_c/n$, or the percentage $p_c \times 100$, of observations that fall into that

interval. For histograms, however, it is more common to use the **density** instead of
the relative frequency or percentage.

The density is the relative frequency for a unit interval. It is obtained by di-
viding the relative frequency by the interval width:

$$f_c = \frac{p_c}{w_c}. \tag{2.2}$$

Here, $p_c = n_c/n$ is the relative frequency with n_c as the frequency of interval
c and n as the total sample size. The width of interval c is denoted w_c.

Let us try calculating the density of the interval (30, 35], which is the fourth
interval. There are $n_4 = 67$ observations in this interval. Therefore, the relative fre-
quency is $p_4 = 67/200 = 0.335$. The interval width is $w_4 = 5$. The density for the
this interval is therefore

$$f_4 = 0.335/5 = 0.067.$$

To create the *density histogram* for `bmi` in R-Commander, click `Graphs` →
`Histogram`, select `bmi` as the `Variable`, and choose `Densities` for the
`Axis Scaling`. The resulting histogram (Fig. 2.11) is similar to that of Fig. 2.10.
However, the height of each bar in this histogram shows the density of the corre-
sponding interval (as opposed to its frequency).

For each interval c, the area of the corresponding bar in the density histogram is
calculated as follows (hight × width):

$$
\begin{aligned}
a_c &= f_c \times w_c \\
&= \frac{p_c}{w_c} \times w_c \\
&= p_c.
\end{aligned}
$$

Therefore, the area of each bar (rectangle) is the relative frequency for the corre-
sponding interval. Since the sum of relative frequencies is 1, the total area of bars in
a density histogram is 1.

Number of Bins   We typically use the same width, denoted as w, for all bins. When
creating a histogram, it is important to choose an appropriate value for w. This is
equivalent to choosing an appropriate number of bins. In R-Commander, by default,
the number of bins is selected automatically using Sturges' formula [32].

You can set the number of bins manually. In R-Commander, click `Graphs` →
`Histogram`, select `bmi` for the `Variable`, and set `Number of bins` to 3.
Compare the resulting histogram to Fig. 2.10.

Fig. 2.12 An example of a symmetric histogram
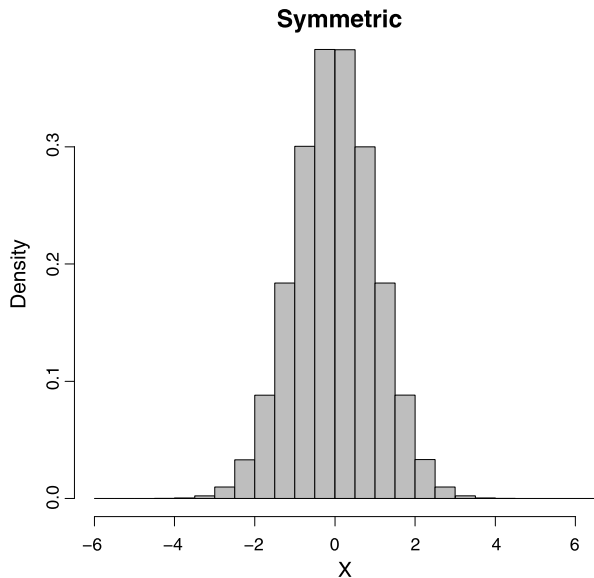
Shapes of Histograms   Besides the location and spread of a distribution, the shape of a histogram also shows us how the observed values spread around the location. Consider the histograms shown in Fig. 2.12. We say that this histogram is **symmetric** around its location (here, zero) since the densities are the same for any two intervals that are equally distant from the center. In reality, we rarely see perfectly symmetric histograms such as the one shown in Fig. 2.12. However, we usually consider a histogram as symmetric if the densities are almost the same for intervals that are equally distant from the location. For example, we can consider the histogram of bmi in Fig. 2.11 as symmetric.

In many situations, we find that a histogram is stretched to the left or right. We call such histograms **skewed**. More specifically, we call them **left-skewed** if they are stretched to the left, or **right-skewed** if they are stretched to the right. For instance, the histogram of Y in Fig. 2.13 is left-skewed. The majority of observations are around 102, but the decrease in densities is slower on the left of the location than on the right. This gives the histogram a long left (lower) tail. On the other hand, the histogram of variable Z in Fig 2.13 is **right-skewed**. The histogram is stretched to the right and has a long right (upper) tail. In the birthwt data set, the histogram of lwt (mother's weight in pounds at last menstrual period) is right-skewed (Fig. 2.14).

The above histograms, whether symmetric or skewed, have one thing in common: they all have one *peak* (or mode). The overall pattern (disregarding minor details) for these histograms can be described as rising to a single peak and then declining. We call such histograms (and their corresponding distributions) **unimodal**. Sometimes histograms have multiple modes. For example, the histogram of variable W in Fig. 2.15 is said to be **bimodal**, since it has two peaks. (Here, a smooth curve has been superimposed to show the overall pattern.)

Left–Skewed   **Right–Skewed**
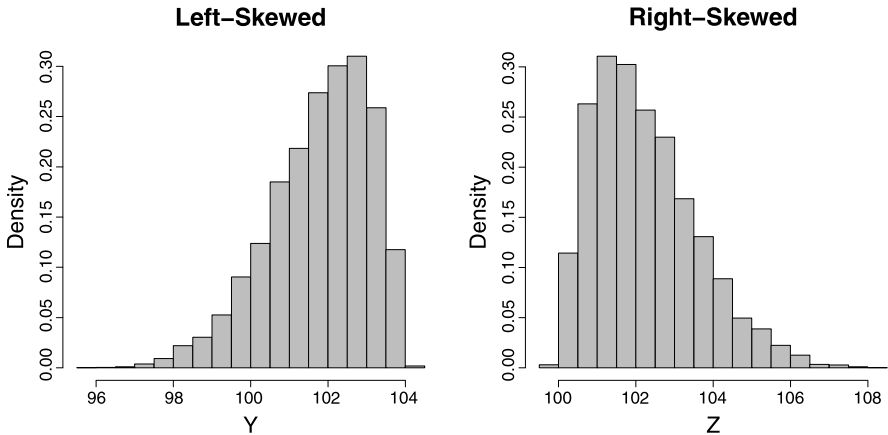
Fig. 2.13 *Left panel*: Histogram of variable *Y* whose histogram is left-skewed. *Right panel*: Histogram of variable *Z* whose histogram is right-skewed

Fig. 2.14 Histogram of variable lwt in the birthwt data set. The histogram is right-skewed
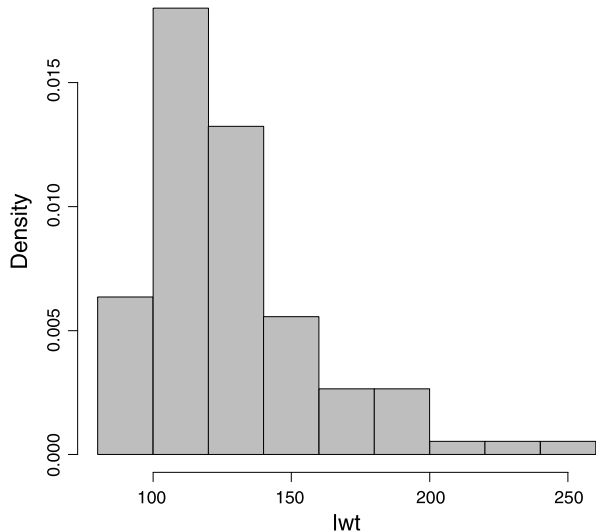
The bimodal histogram appears to be a combination of two unimodal histograms. Indeed, in many situations bimodal histograms (and multimodal histograms in general) indicate that the underlying population is not *homogeneous* and may include two (or more in case of multimodal histograms) subpopulations. For example, the variable *W* in Fig. 2.15 represents blood pressure, and the sample might have been obtained from a population comprised of two groups: a healthy group, whose blood pressure is normal (around 120), and a hypertensive group, who suffer from high blood pressure (around 150).

As another example, suppose that we want to study the protein consumption of European countries [9]. Download the Protein data set from http://lib.stat.cmu.edu/DASL/Datafiles/Protein.html. In R-Commander, import the Protein data set

Fig. 2.15 Histogram of a bimodal distribution. A *smooth curve* is superimposed so that the two peaks are more evident
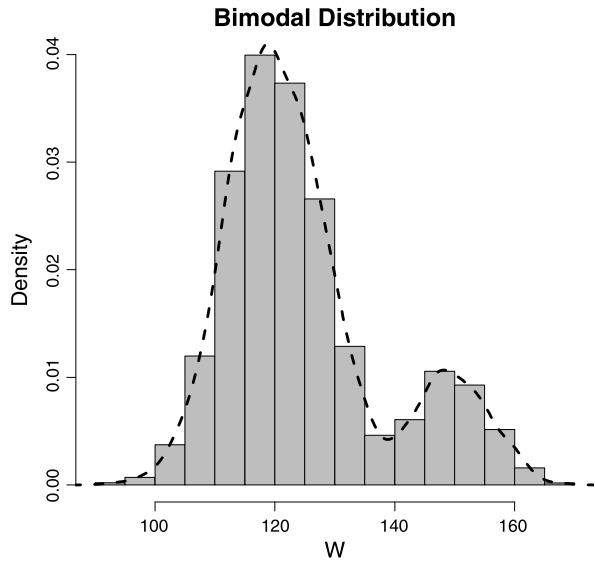
Fig. 2.16 Histogram of protein consumption in 25 European countries for white meat. The histogram is bimodal, which indicates that the sample might be comprised of two subgroups
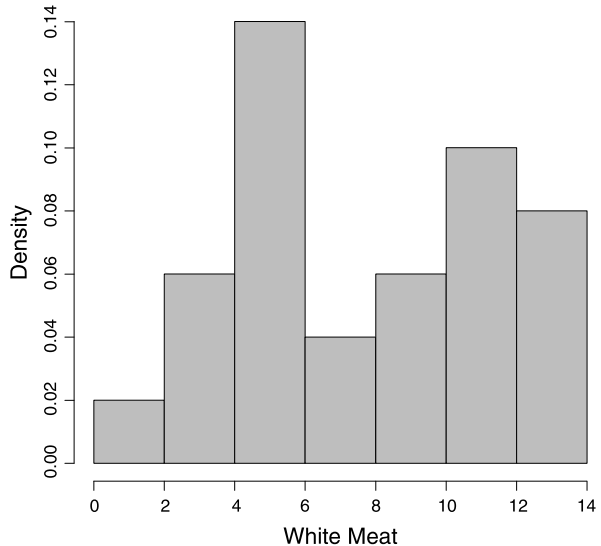
and view it. This data set was collected in 1973 and includes the consumption measurements of nine food groups: RedMeat, WhiteMeat, eggs, Milk, Fish, Cereals, Starch (starchy foods), nuts (pulses, nuts, and oil-seeds), and Fr.Veg (fruits and vegetables). Use the steps described above to plot the density histogram of WhiteMeat. Figure 2.16 shows that the resulting histogram is bimodal. It seems that European countries are divided into two subgroups with respect to the amount of protein consumption from white meat.

Fig. 2.17 Plotting the three
samples from Fig. 2.8 along
with their means (*short
vertical lines*)
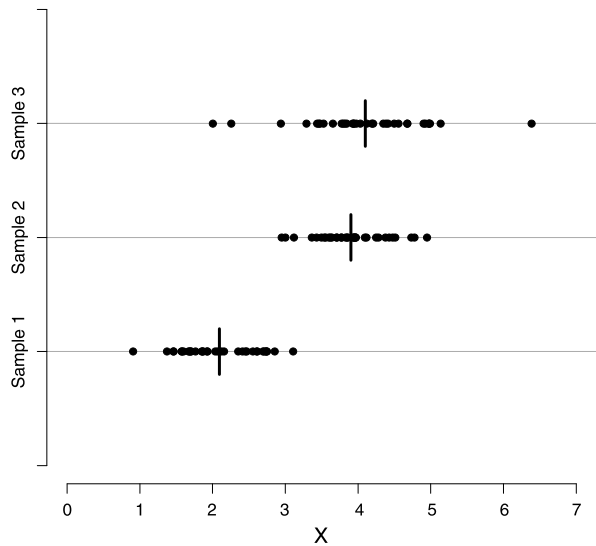
2.4.2 Mean and Median

Histograms are useful for visualizing numerical data and identifying their location
and spread. However, we typically use summary statistics for more precise specifica-
tion of the central tendency and dispersion of observed values. A common summary
statistic for location is the **sample mean**.

> The **sample mean** is simply the average of the observed values. For observed
> values $x_1, \ldots, x_n$, we denote the sample mean as $\bar{x}$ and calculate it by
>
> $$\bar{x} = \frac{\sum_i x_i}{n}, \tag{2.3}$$
>
> where x_i is the ith observed value of X, and n is the sample size.

For Sample 1, Sample 2, and Sample 3, the means are 2.1, 3.9, and 4.1, respec-
tively. The means are shown as short vertical lines in Fig. 2.17.

The sample mean for `bmi` in `Pima.tr` is 32.3. In Fig. 2.18, the mean is shown
by a solid line. In this case, the mean 32.3 appropriately represents the location
(center) of the distribution and the central tendency of the observed values.

While sample mean is a very useful summary statistic for location, it is sensitive
to very large or very small values, which might be outliers (unusual values). For
instance, suppose that we have measured the resting heart rate (in beats per minute)
for five people. The five measurements are {74, 80, 79, 85, 81}. We can calculate

Fig. 2.18 Histogram of `bmi` with the mean (*solid line*) and the median (*dashed line*) are shown as *vertical lines*. The mean and median are nearly equal since the histogram is symmetric
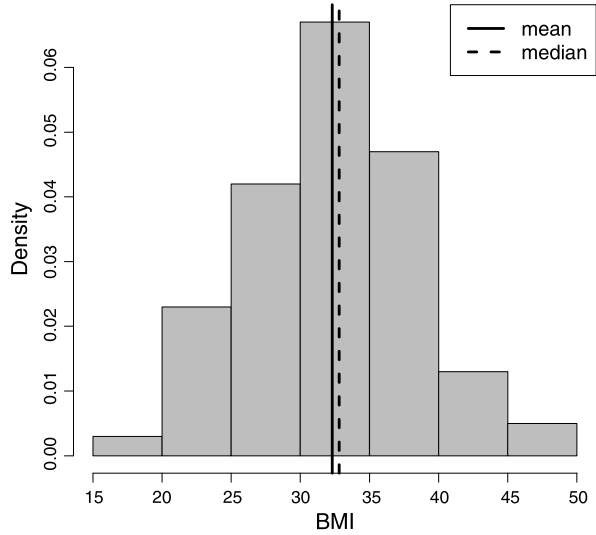
the sample mean as

$$x = \{74, 80, 79, 85, 81\}, \qquad \bar{x} = \frac{74 + 80 + 79 + 85 + 81}{5} = 79.8.$$

In this case, the sample mean is 79.8, which seems to be a good representative of the data.

Now suppose that the heart rate for the first individual is recorded as 47 instead of 74. Compared to other four people, this is a much smaller number, which is either due to a data recording mistake, or the first person is in fact a well-trained athlete with low resting heart rate. In this case, the sample mean is heavily affected by this observation, which is regarded as an outlier, and it is drastically reduced to 74.4:

$$x = \{47, 80, 79, 85, 81\}, \qquad \bar{x} = \frac{47 + 80 + 79 + 85 + 81}{5} = 74.4.$$

Now, the sample mean does not capture the central tendency of the observed data since four out of five measurements are much larger than $\bar{x} = 74.4$.

The **sample median** is an alternative measure of location, which is less sensitive to outliers. For observed values $x_1, \ldots, x_n$, the median is denoted $\tilde{x}$ and is calculated by first sorting the observed values (i.e., ordering them from the lowest to the highest value) and selecting the middle one. If the sample size n is odd, the median is the number at the middle of the sorted observations. If the sample size is even, the median is the average of the two middle numbers.

Fig. 2.19 Histogram of `lwt`
with the mean (*solid line*) and
the median (*dashed line*)
shown as *vertical lines*. The
mean is shifted to the right of
the median because the
histogram is skewed to the
right
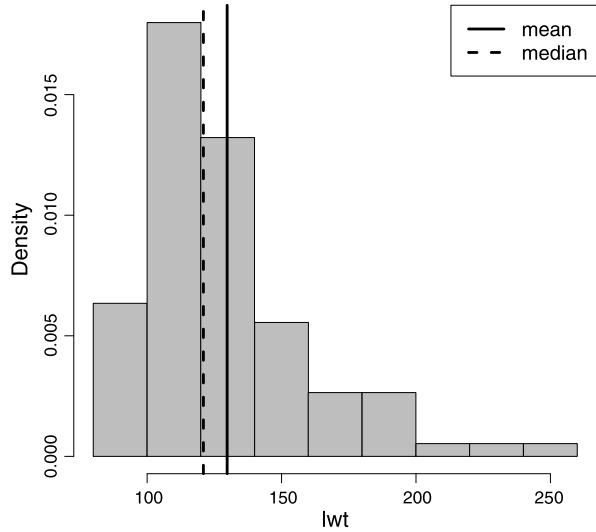
The sample medians for the above two scenarios are

$$x = \{74, 79, 80, 81, 85\}, \qquad \tilde{x} = 80;$$

$$x = \{47, 79, 80, 81, 85\}, \qquad \tilde{x} = 80.$$

In this example, the median remains equal to 80, which properly captures the central tendency of the observed values. In general, the median is not heavily influenced by outliers. We say that the median is more **robust** against outliers.

When there are no outliers and the histogram is almost symmetric, such as the histogram of bmi in Fig. 2.18, both the mean (solid line) and the median (dashed line) are close to each other, and both reasonably represent the location of data. However, when there are outliers, or when the histogram is skewed, such as the histogram of `lwt` in Fig. 2.19, the mean (solid line) moves toward the outliers or the direction of skewness in the histogram more than the median.

Occasionally, we might find situations in which neither the mean nor the median is a good representative of the central tendency. For example, Fig. 2.20 shows that the mean (solid line) and the median (dashed line) for the `WhiteMeat` variable do not capture the central tendency of the data. Most observed values in this case are clustered away from the mean and median. This is usually true for bimodal distributions.

2.4.3 Variance and Standard Deviation

While summary statistics such as mean and median provide insights into the central tendency of values for a variable, they are rarely enough to fully describe a distribution. We need other summary statistics that capture the dispersion of the distribution.

Fig. 2.20 Histogram of
`WhiteMeat` in the
`Protein` data set with the
mean (*solid line*) and the
median (*dashed line*) shown
as *vertical lines*. Neither
mean nor median is a good
measurement for central
tendency since the histogram
is bimodal
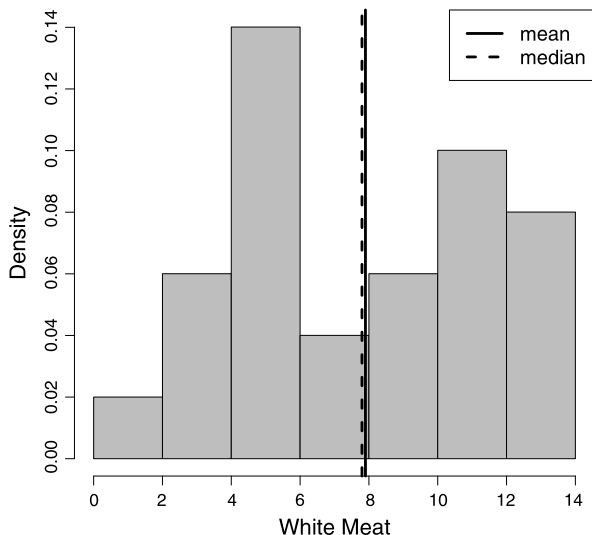
For example, consider Sample 2 and Sample 3 in Fig. 2.17. The two samples have
similar locations, but Sample 3 is more dispersed than Sample 2. The deviations
(differences) of observations from the center (e.g., mean) tend to be larger in Sample 3 compared to Sample 2.

As a further example, consider the following measurements of blood pressure (in
mmHg) for two patients:

$$\text{Patient A:} \quad x = \{95, 98, 96, 95, 96\}, \qquad \bar{x} = 96, \quad \tilde{x} = 96.$$
$$\text{Patient B:} \quad y = \{85, 106, 88, 105, 96\}, \quad \bar{y} = 96, \quad \tilde{y} = 96.$$

While the mean and median for both patients are 96, the readings are more dispersed
for Patient B. Suppose that we choose 96 as the representative value of systolic
blood pressure for both patients. For Patient A, there is a good chance that the next
reading of blood pressure would be close to 96, for example, in the [95, 97] range.
For Patient B, the chance of seeing a blood pressure value close to 96 (e.g., in the
[95, 97] range) would be relatively smaller. For a better description of a variable, we
need summary statistics that measure the dispersion (i.e., variability) of its observed
values.

Two common summary statistics for measuring dispersion are the **sample vari-
ance** and **sample standard deviation**. These two summary statistics are based on
the **deviation** of observed values from the mean as the center of the distribution.
For each observation, the deviation from the mean is calculated as $x_i - \bar{x}$. It is easy
to show that the sum of these deviations over all observed values is always zero.
(Note that $\bar{x} = n \sum x_i$.) Therefore, we cannot simply use the sum of the deviations
as a measure of dispersion. However, the sum would not be zero in general if we
ignore the sings of these deviations (i.e., focus on the distances from the mean). For
this, we can either take the absolute value of deviations, $|x_i - \bar{x}|$, or square them,

$(x_i - \bar{x})^2$. Either way, the sign of deviations becomes irrelevant. Taking the squares of the deviations is a more popular choice. We can then use the average of these squared deviations over all observations as a measure of dispersion:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}. \tag{2.4}$$

Instead of dividing by n, it is more common to divide by $n - 1$. (This increases the above dispersion measurement by a small amount.) The result is called the sample variance.

The sample variance is a common measure of dispersion based on the squared deviations. The variance, denoted s^2, is calculated as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}. \tag{2.5}$$

If we take the square root of the variance,

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}, \tag{2.6}$$

the result is called the sample standard deviation:

Table 2.3 shows the steps for calculating the sample variance and sample standard deviation of blood pressure readings for Patient A and Patient B in the above example. In comparison, the standard deviation for Patient A is much smaller than the standard deviation for Patient B. Thus, we can conclude that the observed blood pressure values are less dispersed for Patient A compared to Patient B.

Table 2.3 Calculating the sample variance and sample standard deviation for Patient A and Patient B in the blood pressure example

Patient A			Patient B		
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
95	−1	1	85	−11	121
98	2	4	106	10	100
96	0	0	88	−8	65
95	−1	1	105	9	81
96	0	0	96	0	0
Σ	0	6	Σ	0	366
$s^2 = 6/4 = 1.5$			$s^2 = 366/4 = 91.5$		
$s = \sqrt{1.5} = 1.22$			$s = \sqrt{91.5} = 9.56$		

Fig. 2.21 Obtaining the
five-number summary
(minimum, maximum, and
quartiles) along with the
mean and standard deviation
for `bwt` in R-Commander
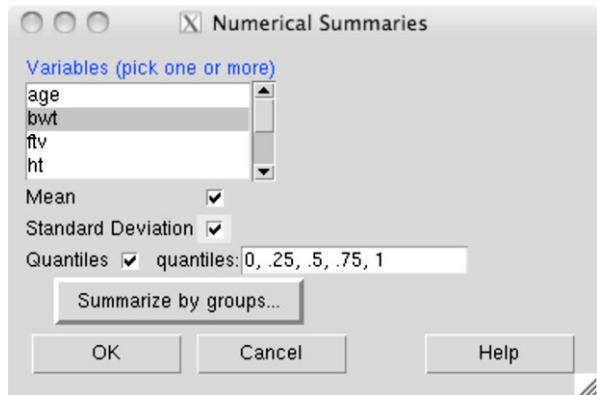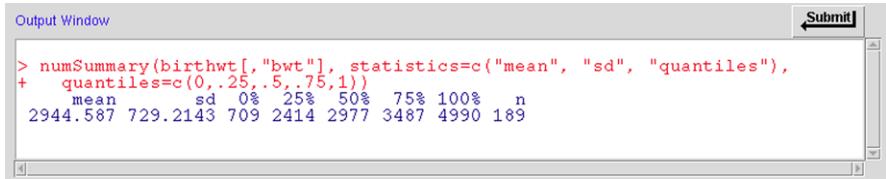
2.4.4 Quantiles

Informally, the sample median could be interpreted as the point that divides the
ordered values of the variable into two equal parts. More precisely, the median is
the point that is greater than or equal to at least half of the values and smaller than
or equal to at least half of the values. Therefore the median is called the 0.5 **quantile**,
which, as we discussed above, provides a measure of location. Similarly, the 0.25
quantile is the point that is greater than or equal to at least 25% of the values and
smaller than or equal to at least 75% of the values. In general, the q quantile is the
point that is greater than or equal to at least $100q\%$ of the values and smaller than or
equal to at least $100(1-q)\%$ of the values. Sometimes, we refer to the q quantile
as the $100q$th **percentile**. For example, the 0.25 quantile is the 25th percentile, and
the median is the 50th percentile.

We can divide the ordered values of a variable into four equal parts using 0.25,
0.5, and 0.75 quantiles. The corresponding points are denoted Q_1, Q_2, and Q_3,
respectively. Note that Q_2 is the 0.5 quantile and is therefore the same as the median.
Q_1 is the point that divides the lower half of the data (i.e., below the median) into
two equal parts. Q_3 is the point that divides the upper half of the data into two
equal parts. We refer to these three points as **quartiles**, of which Q_1 is called the
first quartile or the *lower quartile*, Q_2 (i.e., median) is called the *second* quartile,
and Q_3 is called the *third quartile* or *upper quartile*. The interval from Q_1 (0.25
quantile) to Q_3 (0.75 quantile) covers the middle 50% of the ordered data.

The **minimum** (min), which is the smallest value of the variable in our sample, is
in fact the 0 quantile. On the other hand, the **maximum** (max), which is the largest
value of the variable in our sample, is the 1 quantile. The minimum and maximum
along with quartiles (Q_1, Q_2, and Q_3) are known as **five-number summary**. These
are usually presented in the increasing order: min, first quartile, median, third quar-
tile, max. This way, the five-number summary provides 0, 0.25, 0.50, 0.75, and 1
quantiles.

We can use R-Commander to obtain the five-number summary along with mean
and standard deviation. Make sure `birthwt` is the active data set. Click `Statis-`
`tics` → `Summaries` → `Numerical summaries` (Fig. 2.21). Now select

```
Output Window                                                              Submit

> numSummary(birthwt[,"bwt"], statistics=c("mean", "sd", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
     mean        sd  0%  25%  50%  75% 100%    n
 2944.587 729.2143 709 2414 2977 3487 4990  189
```

Fig. 2.22 Summary statistics for `bwt` from the `birthwt` data set. Here, `sd` denotes standard deviation

`bwt`. (You can select multiple variables by holding down the "control" key.) Make sure `Mean`, `Standard Deviation`, and `Quantiles` are checked. The default for quantiles are the five-number summary. The resulting summary statistics are shown in Fig. 2.22.

> The five-number summary can be used to derive two measures of dispersion: the **range** and the **interquartile range**. The range is the difference between the maximum observed value and the minimum observed value. The interquartile range (IQR) is the difference between the third quartile (Q_3) and the first quartile (Q_1). Compared to the range, the IQR is less sensitive to outliers, which usually fall below Q_1 or above Q_3.

Using the results in Fig. 2.22, the range for `bwt` is $4990 - 709 = 4281$ grams, while the IQR is $3487 - 2414 = 1073$ grams. For this variable, 50% of the birth weight values fall within the [2414, 3487] interval. The birth weight for 25% of babies is above 3487 grams, and for 25% of babies is below 2414 grams.

2.4.5 Boxplots

To visualize the five-number summary, the range and the IQR, we often use a **box-plot** (a.k.a. **box and whisker** plot). Figure 2.23 shows the boxplot for `bwt` along with the plot of actual observed values. The thick line at the middle of the "box" shows the median $\tilde{x} = 2977$. The left side of the box shows the lower quartile $Q_1 = 2414$. Likewise, the right side of the box is the upper quartile $Q_3 = 3487$. Therefore, the box stretches from the lower quartile to the upper quartile and represents the middle 50% of the values of the ordered data. The length of the box is therefore the IQR, which in this case is equal to 1073. 25% of the observations are to the left of this box, and 25% are to the right of it.

The dashed lines extending from the box are known as the **whiskers**. The whisker on the right of the box extends to the largest observed value or $Q_3 + 1.5 \times$ IQR, whichever it reaches first. The whisker on the left extends to the lowest value or $Q_1 - 1.5 \times$ IQR, whichever it reaches first. Data points beyond the whiskers (i.e.,

Fig. 2.23 Horizontal boxplot along with the actual observed values of birth weight from the `birthwt` data set. The *gray box* shows the middle 50% of ordered observed values. The *thick line* in the middle of the box is the median (Q_2) of 2977
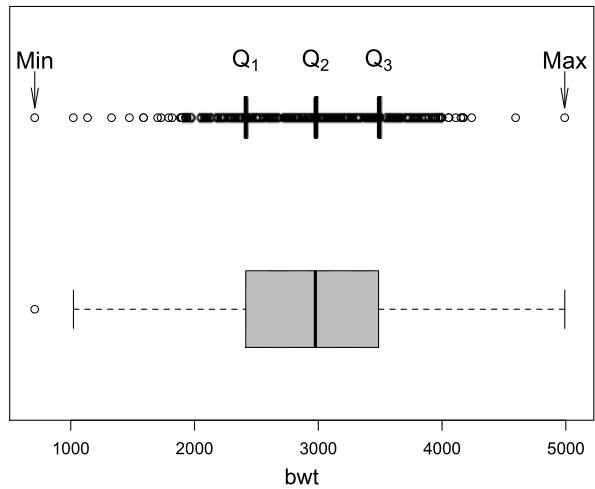
Fig. 2.24 Vertical boxplot for `bwt` using R-Commander
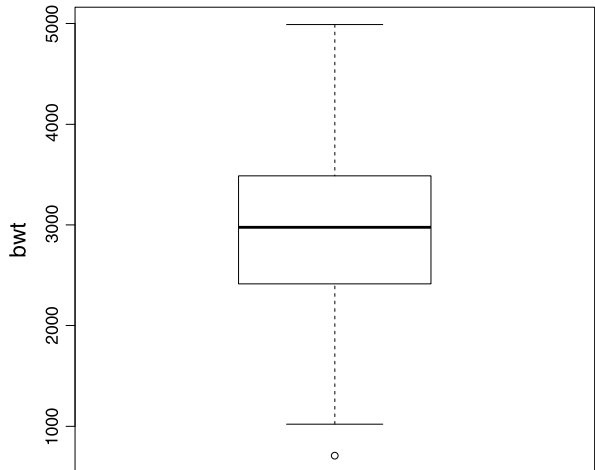
either less than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$) are shown as circles and considered as possible outliers. For `bwt`, the right whisker extends to the maximum value 4990 since it reaches to this value before $3487 + 1.5 \times 1073 = 5096.5$. The left whisker extends to $2414 - 1.5 \times 1073 = 804.5$ since it reaches this point before it reaches the minimum value 709. There is one observation to the left of this whisker, which is shown as a circle. This is, in fact, the minimum observed value, 709, which in this case is considered as a potential outlier.

Very often, boxplots are drawn vertically. This is the default option in R-Commander. To create a boxplot for `bwt` in R-Commander, make sure `birthwt` is the active dataset, click Graphs $\rightarrow$ Boxplot, and select `bwt`. The resulting boxplot is shown in Fig. 2.24. This is the same as the boxplot shown in Fig. 2.23 after 90° rotation.

Fig. 2.25 Vertical boxplot of `lwt`. This plot reveals that the variable `lwt` is right-skewed and there are several possible outliers, whose values beyond the whisker on the top of the box
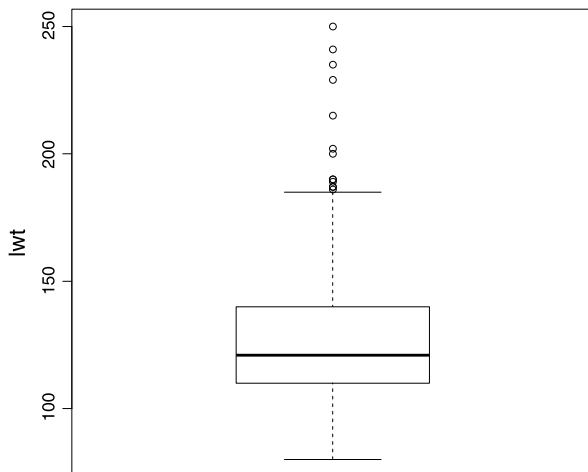
Now, consider the boxplot of `lwt` (Fig. 2.25), whose is distribution is right-skewed. The sample median ($\tilde{x} = 121$) is closer to the bottom ($Q_1 = 110$) than to the top ($Q_3 = 140$) of the box. This is an indication of skewed distribution. Moreover, the upper whisker extends substantially further than the lower whisker. There are several possible outliers, whose observed values fall beyond the whisker on the top of the box.

2.5 Data Preprocessing

Many of the data sets we have been using as examples have been collected in scientific studies. Typically, such data are not ready for immediate analysis. The most common issues are missing values and outliers. For example, the original data on women of Pima Indian Heritage (collected by US National Institute of Diabetes and Digestive and Kidney Diseases) included many observations with missing values. The data set we have been using so far (`Pima.tr`) was obtained after removing these observations. We refer to data in their original form (i.e., collected by researchers) as the **raw** data. Before using the original data for analysis, we should thoroughly check them for missing values and possible outliers. Data exploration techniques we discussed in this chapter can help us to identify data issues that need to be addressed before further analysis. Collectively, we refer to the process of preparing the raw data for analysis as **data preprocessing**. Here, we discuss some simple preprocessing steps.

2.5.1 Missing Data

For our first example, we look at the `Pima.tr2` data set, which includes the `Pima.tr` data set plus many other observations with missing values. The

Fig. 2.26 Viewing the
`Pima.tr2` data set in
R-Commander. Many
observations in this data set
have missing values (`NA`)

	npreg	glu	bp	skin	bmi	ped	age	type
193	1	128	48	45	40.5	0.613	24	Yes
194	2	112	68	22	34.1	0.315	26	No
195	1	140	74	26	24.1	0.828	23	No
196	2	141	58	34	25.4	0.699	24	No
197	7	129	68	49	38.5	0.439	43	Yes
198	0	106	70	37	39.4	0.605	22	No
199	1	118	58	36	33.3	0.261	23	No
200	8	155	62	26	34.0	0.543	46	Yes
201	2	134	70	NA	28.9	0.542	23	Yes
202	10	75	82	NA	33.3	0.263	38	No
203	0	146	70	NA	37.9	0.334	28	Yes
204	1	180	NA	NA	43.3	0.282	41	Yes
205	5	104	74	NA	28.8	0.153	48	No
206	9	164	78	NA	32.8	0.148	45	Yes
207	1	80	55	NA	19.1	0.258	21	No
208	4	171	72	NA	43.6	0.479	26	Yes

`Pima.tr2` is available in the `MASS` package. Follow the steps described in the previous chapter to load the `MASS` package and select `Pima.tr2` (which is located right after `Pima.tr` in the list) as the active data set. Figure 2.26 shows a part of this data set. Here, missing values are denoted `NA` (Not Available).

In general, it is up to the researcher to decide whether to remove the observations with missing values or impute (guess) the missing values in order to keep the observations. If we choose to remove all observations with missing values (this is how the `Pima.tr` data set was created based on `Pima.tr2`), we can do so by clicking `Data → Active data set → Remove cases with missing data`. Under `Name for new data set` enter `Pima.complete`. This creates a data set, which does not include any observation with missing values. (Notice that `Pima.complete` becomes the active data set.) Indeed, this data set is exactly the same as `Pima.tr`, which we have been using so far.
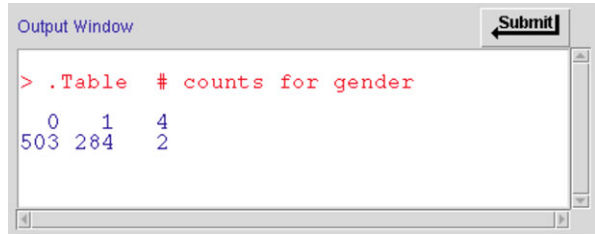
While simply removing observations with missing values is an easy approach for handling missing data, it is quite wasteful and inefficient. On the other hand, missing data imputation techniques, i.e., using statistical methods to fill-in missing values, tend to be complex. However, if done properly, they can improve our analysis. For an overview of statistical methods for analyzing data with missing values, refer to [18].

Sometimes we can temporarily ignore missing values if the variable whose values are missing is not the focus of our analysis at the moment. In the above example, if we are focusing on the `bp` (blood pressure) variable, we do not need to remove observations 201, 202, 203, 205, …. Of course, we still need to either remove or impute the observation 204 and any other observation whose blood pressure reading is missing. To remove individual observations, click `Data → Active data set → Remove row(s) from active data` and enter the row numbers (the leftmost number in the data set) for observations you want to remove.

2.5.2 Outliers

Dealing with missing values is not the only challenge of working with raw data. Sometimes, an observed value of a variable is suspicious since it does not follow the

Fig. 2.27 Frequency table for `gender` from the `AsthmaLOS` data set. The value of `gender` for two observations are entered as "4", while `gender` can only take 0 or 1


```
Output Window                                                    Submit

> .Table   # counts for gender

   0    1    4
 503  284    2
```

overall patterns presented by the rest of the data. We refer to such observations as outliers. Suppose, for example, that almost all BMI values in our sample are between 20 and 40. Observing a BMI value of 50 would be suspicious. Further investigation might reveal that in fact this is the correct value of BMI for an individual in our sample. In this case, this outlier is a legitimate value. However, a BMI value of 500 or –50 is clearly an erroneous observation, which is possibly due to a data entry mistake.

We could identify outliers using data exploration techniques. As an example, we use the `AsthmaLOS` data collected by [12] to study the length of stay in hospital for asthmatic children in the USA. Download the data set from the book website (http://extras.springer.com) and import it to R-Commander. The variables in this data sets are:

- `los`: length of stay in hospital (in days).
- `hospital.id`: hospital ID.
- `insurer`: the insurer, which is either 0 or 1.
- `age`: the age of the patient.
- `gender`: the gender of the patient; 1 for female, and 0 for male.
- `race`: the race of the patient; 1 for white, 2 for Hispanic, 3 for African-American, 4 for Asian/Pacific Islander, 5 for others.
- `bed.size`: the number of beds in the hospital; 1 means 1 to 99, 2 means 100 to 249, 3 means 250 to 400, 4 means 401 to 650.
- `owner.type`: the hospital owner; 1 for public, 2 for private.
- `complication`: if there were any treatment complication; 0 means there were no complications, 1 means there were some complications.

Before working with this data set, follow the steps discussed in Sect. 2.2 to convert the variables `hospital.id`, `insurer`, `gender`, `race`, `owner type`, and `complication` to factors (categorical). Next, obtain the frequency tables for `gender`. The resulting tables are shown in Fig. 2.27. Notice that while the `gender` variable can take only two values, 1 for female and 0 for male, the data include two observations whose values for `gender` is "4". These values are entered by mistake and should be either removed (as described above) or corrected if possible. If we know the correct values for these observations (e.g., by examining the medical records), we can edit the data and keep the observations. To edit a data set, click `Edit data set` button in front of its name on the menu bar. This opens the *R Data Editor* window, where you can find the erroneous values and correct them.

Fig. 2.28 The boxplot of
`los` with two extremely large
values
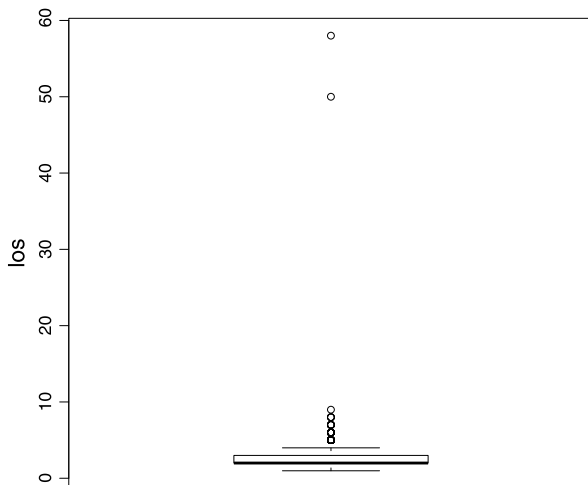
Now consider the variable `los` (length of stay) in the `AsthmaLOS` data set. Figure 2.28 shows the boxplot for this variable. As we can see, there are two children whose length of stay is extremely large (50 and 58). These values are not consistent with the rest of data. (All other values are less than 10.) However, if we find that they are legitimate and correctly recorded values, we should keep them in our data since they provide important information on the distribution of the variable (e.g., how extreme could be). Of course, such observations can drastically affect our results. For analyzing such data, we could use statistical methods that are more robust against outliers (e.g., median, IQR).

2.5.3 Data Transformation

Occasionally, we rely on data transformation techniques (i.e., applying a function to the variable) to reduce the influence of extreme values in our analysis. Two of the most commonly used transformation functions for this purpose are *logarithm* and *square root*. The logarithm function, $\log(x)$, is usually used to transform right-skewed variables with positive values. The square root function is usually used for right-skewed count variables. We use these transformations to reduce the skewness, i.e., to make it more symmetric, and reduce the influence of extreme values.

Consider the `lwt` variable in the `birthwt` data set. As shown in the left panel of Fig. 2.29, the variable is right-skewed. To use log-transformation, click `Data →` `Manage variables in active data set → Compute new vari-able`. Under `New variable name`, enter `log.lwt`, and under `Expression to compute`, enter `log(lwt)`. (If we want to use the square root transformation, we use `sqrt` instead of `log`.) This creates a new variable `log.lwt` whose values are the natural logarithm of `lwt`. Next, create the density histogram for this newly created variable. As shown in the right panel of Fig. 2.29, the resulting variable is less skewed compared to the original variable.
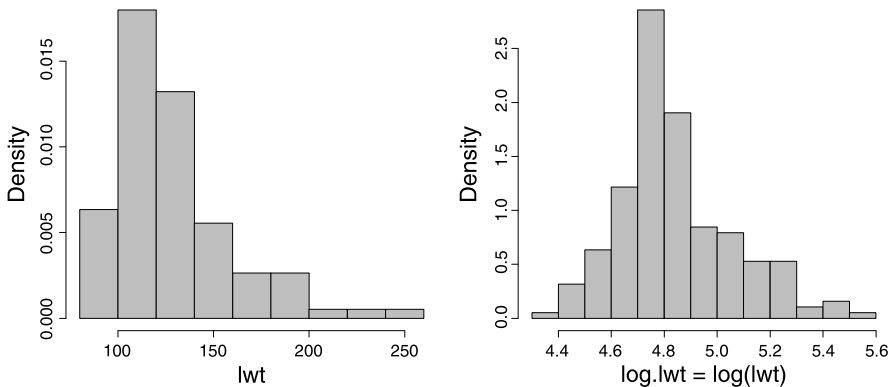
Fig. 2.29 *Left panel*: Histogram of variable `lwt` in the `birthwt` data set. *Right panel*: Histogram of variable `log(lwt)`, log-transformation of `lwt`

The transformation techniques discussed so far are used commonly in statistical analysis. You can of course use the above approach to transform a variable in many other ways. For example, suppose that you want to apply the square transformation to a variable X. (This is also a common transformation in regression analysis.) To do this, you can follow the above steps and simply enter `X^2` under `Expression to compute`. (Here, symbol "`^`" is used for exponentiation.)

2.5.4 Creating New Variable Based on Two or More Existing Variables

In the previous chapter, we discussed creating new variables based on existing ones as a common data preprocessing step. Here, we show how we can create a new variable based on two or more existing variables. Consider the `bodyfat` data set, which includes weight and height. Using these two variables, we can calculate BMI for each person in the sample using the equation

$$BMI = \frac{weight \times 703}{(height)^2},$$

where weight is in pounds, and height is in inches.

To create BMI, click `Data → Manage variables in active data set → Compute new variable`. Under `New variable name`, enter `BMI`, and under `Expression to compute`, we enter (Fig. 2.30)

$$(\text{weight} * 703)/(\text{height}^2)$$

This will create a new variable called `BMI`. You can now investigate the linear relationship between this variable and percent body fat by calculating their sample correlation coefficient. Pearson's correlation coefficient between `siri` and `BMI` is 0.72, which indicates a strong positive linear relationship as expected.

Fig. 2.30 Creating a new variable `BMI` based on weight and height for each person in the `bodyfat` data set
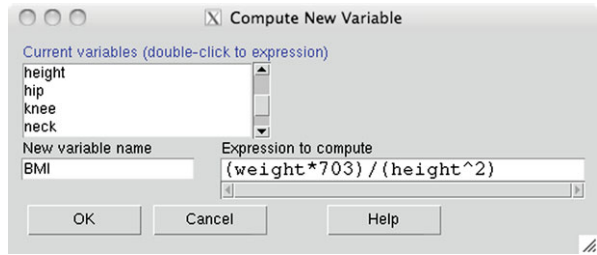
Table 2.4 Standard weight status based on BMI according to CDC

BMI	Weight Status
Below 18.5	Underweight
18.5–24.9	Normal
25.0–29.9	Overweight
30.0 and Above	Obese

2.5.5 Creating Categories for Numerical Variables

Another common preprocessing technique is to create categorical variables based on numerical variables. This could help us to see the patterns more clearly and identify relationships more easily. Recall that histograms are created by dividing the range of a numerical variable into intervals. Instead of using arbitrary intervals, we might prefer to group the values in a meaningful way. This way, we can create a categorical variable based on a numerical variable. For example, according to the Centers for Disease Control and Prevention (CDC), the standard weight status categories associated with BMI ranges for adults are as in Table 2.4.

In R-Commander, let us divide subjects based on their `bmi` (from the `Pima.tr`) into four groups: Underweight, Normal, Overweight, and Obese. Click `Data →` `Manage variables in active data set → Recode variables`. Select `bmi` as the `Variable to recode` and enter "weight.status" as the `New variable name` (Fig. 2.31). Then in the `Enter recode directives` box, type

```
0:18.5 = "Underweight"

18.5:24.9 = "Normal"

25.0:29.9 = "Overweight"

30.0:100 = "Obese"
```

Now view the `Pima.tr` data set. The newly created variable `weight.status` is added to the data set. This variable is categorical. More specifically, it is an ordinal variable. To specify the order of categories in R-Commander, click `Data` `→ Manage variables in active data set → Reorder factor`

Fig. 2.31 Recoding the
numerical variable `bmi` to be
categorical
(`weight.status`)
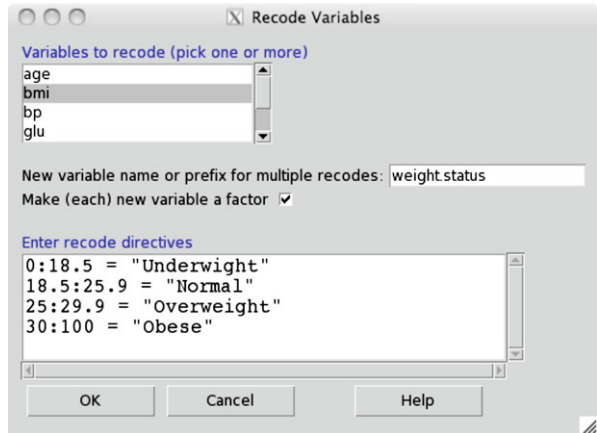
Fig. 2.32 Reordering the
categories for the variable
`weight.status` such that
"Underweight" is the first
category, "Normal" is the
second category,
"Overweight" is the third
category, and "Obese" is the
fourth category

`levels`. Then select `weight.status`. R-Commander will open a window to
reorder levels of the categorical variable. Change the order according to Fig. 2.32.
(Note that the default order is alphabetical.) Now you can create the barplot
for `weight.status` (Fig. 2.33). The graph of the `weight.status` variable
clearly indicates that the "Obese" category has the highest frequency.

2.6 Advanced

In this section, we discuss some data exploration and data transformation techniques
that are slightly more advanced. We also discuss some commonly used R functions
for data exploration.

2.6.1 Coefficient of Variation

Suppose that we want to compare the dispersion of `bwt` to that of `lwt` using their
standard deviations. Use R-Commander to obtain the means and standard deviations
for `bwt` and `lwt` in the `birthwt` data set. Based on the results shown in Fig. 2.34,

Fig. 2.33 The bar graph for `bmi` after converting the numerical variable to a categorical variable
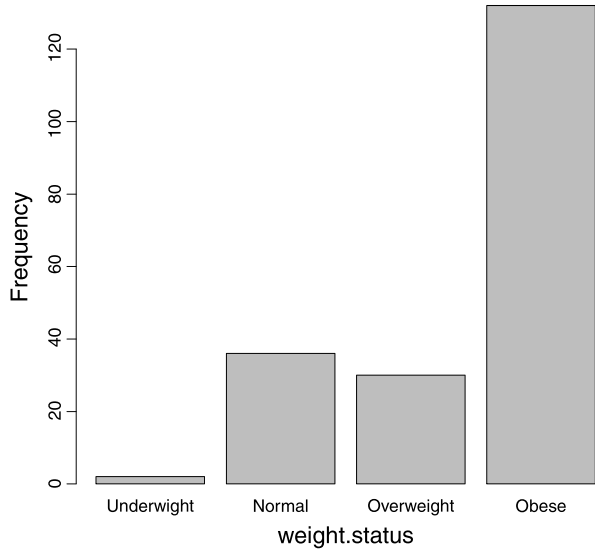
Fig. 2.34 Summary statistics for `bwt` and `lwt` from the `birthwt` data set


```
Output Window                                              Submit
> numSummary(birthwt[,c("bwt", "lwt")], statistics=c("mean", "sd"),
+   quantiles=c(0,.25,.5,.75,1))
          mean        sd    n
bwt 2944.5873 729.21430 189
lwt  129.8148  30.57938 189
```
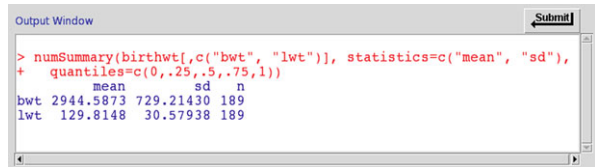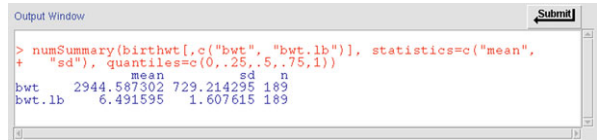
Fig. 2.35 Creating a new variable `bwt.lb` (birth weight in pounds) and obtaining its summary statistics
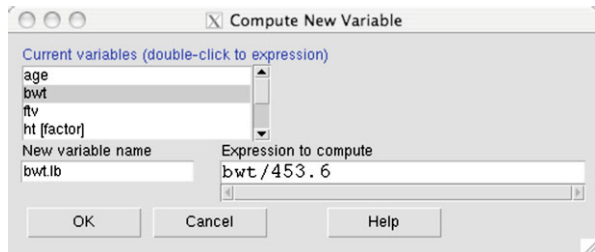
it seems that `bwt` is more dispersed than `lwt` since it has higher standard deviation compared to `lwt`. However, the two variables are not comparable; they have different units. Let us change the unit of `bwt` from grams to pounds. For this, we need to divide its values by 453.6. In R-Commander, click `Data → Manage variables in active data set → Compute new variable`. This opens a window (Fig. 2.35), where we create new variable for birth weight in pounds. Under `new variable name`, enter `bwt.lb`. Under `Expression to compute`, enter `bwt/453.6`. The newly created variable `bwt.lb`, whose values are birth weight in pound, will be added to the `birthwt` data set. (View the data set to make sure that this is done correctly.)

Fig. 2.36 Creating a new
variable `bwt.lb` (birth
weight in pounds) and
obtaining its summary
statistics

Now, use R-Commander to find the mean and standard deviation of `bwt` and
`bwt.lb`. The results are shown in Fig. 2.36. After changing the measurement unit
from grams to pounds, the standard deviation changes from 729.2 to 1.6. Now, this
is much smaller than the standard deviation of `lwt`, which is 30.6 (see Fig. 2.34).
This is of course expected since the values of `lwt` are much larger than the values
of `bwt.lb`. As a result, `lwt` has much larger sample mean and larger deviations
around the mean compared to `bwt.lb`.

The above results illustrate how difference in measurement units and large differ-
ences in sample means make it difficult to compare variables based on their standard
deviations. In many situations, we can avoid these issues by using another measure
of variation called the **coefficient of variation** instead of standard deviation.

> To quantify dispersion independently from units, we use the coefficient of
> variation, which is the standard deviation divided by the sample mean (as-
> suming that the mean is a positive number):
>
> $$CV = \frac{s}{\bar{x}}. \tag{2.7}$$

The coefficient of variation for `bwt` (birth weight in grams) is 729.2/
2944.6 = 0.25 and for `bwt.lb` (birth weight in pounds) is 1.6/6.5 = 0.25. There-
fore, the coefficient of variation is the same, even though `bwt` has a larger stan-
dard deviation compared to `bwt.lb`. Comparing this coefficient of variation to
30.6/129.8 = 0.24, which is the coefficient of variation for `lwt`, suggests that the
two variables have roughly the same dispersion in terms of CV. In general, the
coefficient of variation is used to compare variables in terms of their dispersion
when the means are substantially different (possibly as the result of having different
measurement units).

2.6.2 Scaling and Shifting Variables

To see why the coefficient of variation ($CV = s/\bar{x}$) is independent of measurement
units in the above example, we need to learn about how the mean and standard
deviation change when we change the scale of a variable. For example, we changed
the scaled of `bwt` by multiplying it by the constant 1/453.6 (i.e., dividing it by
453.6).

In general, when we multiply the observed values of a variable by a constant a, its mean, standard deviation, and variance are multiplied by a, $|a|$, and a^2, respectively. That is, if $y = ax$, then

$$\bar{y} = a\bar{x},$$
$$s_y = |a|s_x,$$
$$s_y^2 = a^2 s_x^2,$$

where $\bar{x}$, s_x, and s_x^2 are the sample mean, standard deviation, and variance of the original observations x, and $\bar{y}$, s_y, and s_y^2 are the sample mean, standard deviation, and variance of scaled observations y.

In the above example, the mean and standard deviation of bwt (denoted x) were $\bar{x} = 2944.6$ and $s_x = 729.2$, respectively (Fig. 2.22). To convert the measurement unit to pounds, we multiplied bwt by $a = 1/453.6$ to create a new variable bwt.lb (denoted y). The mean and standard deviation of bwt.lb are therefore as follows:

$$\bar{y} = \frac{1}{453.6} \times 2944.6 = 6.5,$$

$$s_y = |a|s_x = \frac{1}{453.6} \times 729.2 = 1.6,$$

which are the same values as what we obtained by using R-Commander (Fig. 2.36).

When the measurement units are changed by multiplying the observed values by a positive constant (e.g., multiplying by $1/453.6$ in the above example to convert grams to pounds), the coefficient of variation is not affected since both mean and standard deviation will be multiplied by that number. If $y = ax$ (where a is a positive constant), then

$$CV_y = \frac{s_y}{\bar{y}} = \frac{as_x}{a\bar{x}} = \frac{s_x}{\bar{x}} = CV_x.$$

What happens if instead of scaling the observed value, we shift them by a constant b (which can be negative): $y = x + b$? For example, suppose after researchers collected the birthwt data set, they realized that the weighting scale they used to measure birth weight was not calibrated properly, and they need to add 20 grams to the weight of each child, i.e., $y = x + 20$. Therefore, all the observed values for bwt will be shifted upwards by 20 points. Intuitively, this shifts the sample mean by 20 points. However, since the difference between observed values and the mean do not change, the standard deviation and variance remain unchanged. In general, if we shift the observed values by b, i.e., $y = x + b$, then

$$\bar{y} = \bar{x} + b,$$
$$s_y = s_x,$$
$$s_y^2 = s_x^2.$$

If we multiply the observed values by the constant a and then add the constant b to the result, i.e., $y = ax + b$, then

$$\bar{y} = a\bar{x} + b,$$

$$s_y = |a|s_x,$$

$$s_y^2 = a^2 s_x^2.$$

Therefore, when changing measurement units involved adding a constant (e.g., adding 273.15 to convert Celsius to Kelvin), the coefficient of variation will change. If $y = ax + b$ (assuming $a > 0$ and $b \neq 0$), then

$$CV_y = \frac{s_y}{\bar{y}} = \frac{as_x}{a\bar{x} + b} \neq \frac{s_x}{\bar{x}}.$$

2.6.3   Variable Standardization

Variable standardization is a common *linear* transformation, where we subtract the sample mean $\bar{x}$ from the observed values and divide the result by the sample standard deviation s, in order to shift the mean to zero and make the standard deviation 1:

$$y_i = \frac{x_i - \bar{x}}{s}.$$

Using such transformation is especially common in regression analysis (Chap. 11) and clustering (Sect. 12.1). Following the rules we discussed above, subtracting $\bar{x}$ from the observations shifts the sample mean to zero. This, however, does not change the standard deviation. Dividing by s, on the other hand, changes the sample standard deviation to 1. The mean is also divided by s. However, since the sample mean has become zero after subtracting $\bar{x}$, it remains zero. Therefore, variable standardization creates a new variable with mean 0 and standard deviation 1.

Suppose that we want to standardize `lwt` using R-Commander. For this, we can follow the steps for computing a new variable (Sect. 2.6.1), enter `std.lwt` under New variable name, and $(lwt - 129.8)/30.6$ under Expression to compute. This creates the standardized version of `lwt` called `std.lwt`. Now, find the mean and standard deviation of `std.lwt`. Alternatively, we can standardize a variable by clicking Data → Manage variables in active data set → Standardize variables. Select `lwt` under Variables. This will create a new variable called `Z.lwt`, which will be added to the data set. View the `birthwt` data set and find the mean and standard deviation of the newly created variable `Z.lwt`.

2.6.4 Data Exploration with R Programming

Writing your own R commands (as opposed to using R-Commander) gives you more control over the output and a deeper understanding of the material. In Appendix B, we provide a brief introduction to R programming. Here, we review the functions that are commonly used for data exploration. We start by loading the `Pima.tr` data set, which is available from the `MASS` package.

```
> library(MASS)
> data(Pima.tr)
```

The `library()` command loads the `MASS` package, and the `data()` command loads the `Pima.tr` data set. Note that the package should be loaded first before we can access its data sets.

Type `Pima.tr` to view the entire data set. If the data set is large, it is better to use the `head()` function, which shows only the first part (few rows) of the data set.

```
> head(Pima.tr)

  npreg glu bp skin  bmi   ped age type
1     5  86 68   28 30.2 0.364  24   No
2     7 195 70   33 25.1 0.163  55  Yes
3     5  77 82   41 35.8 0.156  35   No
4     0 165 76   43 47.9 0.259  26   No
5     0 107 60   25 26.4 0.133  23   No
6     5  97 76   27 35.6 0.378  52  Yes
```

When you obtain a data set from a package, you can use the `help()` function to view the description on the data available in the package.

```
> help(Pima.tr)
```

Bar Graphs and Frequencies   A common summary statistic for categorical variables is its frequencies, n_c. Use the `table()` function to obtain the frequencies for the categorical variable `type` from the `Pima.tr` data set.

```
> type.freq <- table(Pima.tr$type)
> type.freq

 No Yes
132  68
```

Note that the $ symbol is being used to access the `type` variable in the `Pima.tr` data set.

Now, use the `type.freq` table to create the bar graph. Bar graphs show us how observations categorical variables are distributed in the sample.

```
> barplot(type.freq, xlab = "Type", ylab = "Frequency",
+       main = "Frequency Bar Graph of Type")
```

The first parameter to the `barplot()` function is the frequency table. The options `xlab` and `ylab` label the x and y axes, respectively. Likewise, the `main` option puts a title on the plot.

Often it is more informative to report the relative frequencies. The relative frequency is the percentage or proportion in each category and is calculated by $p_c = n_c/n$ as in Eq. 2.1. Therefore, we need the frequencies n_c (stored in the `type.freq` table) and the total sample size n. Since the sum of the frequencies is the total sample size, $\sum_c n_c = n$, we can use the `sum()` function to add the entries in the frequency table:

```
> n <- sum(type.freq)
> n
```

```
[1] 200
```

Now create a table of relative frequencies by dividing the frequency table by the sample size:

```
> type.rel.freq <- type.freq/n
```

Use the `round()` function to limit the output to 2 decimal places:

```
> round(type.rel.freq, 2)
```

```
  No   Yes
0.66 0.34
```

We can also multiply the relative frequencies by 100 to get the percentages:

```
> round(type.rel.freq, 2) * 100
```

```
No Yes
66  34
```

Finally, you can create a relative frequency barplot with

```
> barplot(type.rel.freq, xlab = "Type",
+       ylab = "Relative Frequency",
+       main = "Relative Frequency Bar Graph of Type")
```

If the levels of a categorical variable in the data set is coded as numbers, we need to convert the type of variable to *factor* using the `factor()` function, so that R recognizes it as categorical. You can use the function `is.factor()` to examine whether a variable is a factor. For example, the `smoke` variable (smoking status) in `birthwt` is coded as 0 for mothers who did not smoke during their pregnancy and 1 for mothers who smoked during their pregnancy. R automatically considers this variable as numerical. To convert the variable to categorical, use the following code:

```
> data(birthwt)
> is.factor(birthwt$smoke)

[1] FALSE

> birthwt$smoke <- factor(birthwt$smoke)
> is.factor(birthwt$smoke)

[1] TRUE

> table(birthwt$smoke)

  0   1
115  74
```

Histograms   Histograms are commonly used to visualize numerical variables. To create a *frequency* histogram for `age`, use the `hist()` function with the `freq` option set to "TRUE" (which is the default):

```
> hist(Pima.tr$age, freq = TRUE,
+      xlab = "Age", ylab = "Frequency",
+      col = "grey", main = "Frequency Histogram of Age")
```

Then create a *density* histogram of `age` by setting the `freq` option to "FALSE":

```
> hist(Pima.tr$age, freq = FALSE,
+      xlab = "Age", ylab = "Density",
+      col = "grey", main = "Density Histogram of Age")
```

Summary Statistics   We can obtain the mean and median of numerical data with the `mean()` and `median()` functions. Find these statistics for numerical variables in `Pima.tr`:

```
> mean(Pima.tr$npreg)
```

```
[1] 3.57
```

```
> median(Pima.tr$bmi)
```

```
[1] 32.8
```

The `quantile()` function with the `probs` option returns the specified quantiles:

```
> quantile(Pima.tr$bmi, probs = c(0.1, 0.25, 0.5, 0.9))

   10%    25%    50%    90%
24.200 27.575 32.800 39.400
```

Here, the desired quantiles are specified as a vector using the combine `c()` function. The five-number summary along with the mean can simply be obtained with the `summary()` function:

```
> summary(Pima.tr$bmi)

 Min. 1st Qu. Median   Mean 3rd Qu.   Max.
18.20   27.58  32.80  32.31   36.50  47.90
```

We can present the five-number summary visually with a boxplot:

```
> boxplot(PIma.tr$bmi, ylab = "BMI")
```

While the default is to create vertical boxplots, we can also create horizontal boxplots by specifying the `horizontal` option to true:

```
> boxplot(Pima.tr$bmi, ylab = "BMI", horizontal = TRUE)
```

Find the interquartile range (IQR) with the `IQR()` function:

```
> IQR(Pima.tr$bmi)
```

```
[1] 8.925
```

The smallest and largest observations can be obtained with the `range()` function (the functions `min()` and `max()` could also be applied):

```
> minMax <- range(Pima.tr$bmi)
> minMax
```

```
[1] 18.2 47.9
```

Here, we created a vector object `minMax` with the minimum as the first element and the maximum as the second element. Obtain the range by subtracting the first element from the second:

```
> minMax[2] - minMax[1]

[1] 29.7
```

The variance and standard deviation are also easily calculated with `var()` and `sd()`:

```
> var(Pima.tr$bmi)

[1] 37.5795

> sd(Pima.tr$bmi)

[1] 6.130212
```

Creating Categories for Numerical Variables   The `hist()` function automatically divides the range of possible values into several intervals. Instead, as discussed above, we can create more meaningful intervals, which will be treated as categories. To create a categorical variable `weight.status` based on the `bmi` variable in `Pima.tr`, we can go through each observation one by one and assign each observation to one of the four categories: "Underweight", "Normal", "Overweight", and "Obese". To do this, we can use **loops** and **conditional** statements, which are discussed in Appendix B.

First, we start by creating an empty vector of size 200 within the `Pima.tr` data frame:

```
> Pima.tr$weight.status <- rep(NA, 200)
```

Next, we set the values of `weight.status` for all observations by using `if-else()` statements within a `for()` loop:

```
> for (i in 1:200) {
+     if (Pima.tr$bmi[i] < 18.5) {
+         Pima.tr$weight.status[i] <- "Underweight"
+     }
+     else if (Pima.tr$bmi[i] >= 18.5 &
+         Pima.tr$bmi[i] < 24.9) {
+         Pima.tr$weight.status[i] <- "Normal"
+     }
+     else if (Pima.tr$bmi[i] >= 24.9 &
+         Pima.tr$bmi[i] < 29.9) {
```

```
+           Pima.tr$weight.status[i] <- "Overweight"
+      }
+      else {
+           Pima.tr$weight.status[i] <- "Obese"
+      }
+ }
```

Here, the loop counter goes from 1 to 200. Use the `head()` function to view the result:

```
> head(Pima.tr)

  npreg glu bp skin  bmi   ped age type weight.status
1     5  86 68   28 30.2 0.364  24   No         Obese
2     7 195 70   33 25.1 0.163  55  Yes    Overweight
3     5  77 82   41 35.8 0.156  35   No         Obese
4     0 165 76   43 47.9 0.259  26   No         Obese
5     0 107 60   25 26.4 0.133  23   No    Overweight
6     5  97 76   27 35.6 0.378  52  Yes         Obese
```

Before we use the newly created variable `weight.status` in statistical analysis, we should convert its type to factor.

```
> Pima.tr$weight.status <- factor(Pima.tr$weight.status)
```

While the above code makes `weight.status` a factor variable, it does not take into account the ordering of levels. The levels are ordered alphabetically and can be examined using the `levels()` function:

```
> levels(Pima.tr$weight.status)

[1] "Normal"      "Obese"
[3] "Overweight"  "Underweight"
```

We can provide the right ordering when we use the `factor()` function to convert the variable:

```
> Pima.tr$weight.status <- factor(Pima.tr$weight.status,
+     levels = c("Underweight", "Normal",
+        "Overweight", "Obese"))
> levels(Pima.tr$weight.status)

[1] "Underweight" "Normal"
[3] "Overweight"  "Obese"
```

Handling Missing Data in R   To find missing values of a variable, we can use the `is.na()` function, which returns "TRUE" when the value is missing and "FALSE" otherwise. Consider the `Pima.tr2` data set from the MASS library (the `Pima.tr` data set is obtained from `Pima.tr2` by removing observations with missing values):

```
> data(Pima.tr2)
> is.na(Pima.tr2$bp)
```

To obtain the indices of observations whose values are missing, we can use the `which()` function along with the `is.na()` function. In general, `which()` can be used to find the indices of "TRUE" values for a logical vector:

```
> which(is.na(Pima.tr2$bp))
```

The `complete.cases()` function returns a logical vector indicating which cases (observations) in the data set are complete, i.e., have no missing values:

```
> complete.cases(Pima.tr2)
```

To remove cases with missing values, we can use the `na.omit()` function:

```
> Pima.complete <- na.omit(Pima.tr2)
```

Here, the newly created `Pima.complete` data set includes only the complete cases from `Pima.tr2`.

2.7 Exercises

1. Download the calcium data set from the Data and Story Library: http://lib.stat. cmu.edu/DASL/Datafiles/Calcium.html. The data were collected to investigate whether increasing calcium intake reduces blood pressure. 21 people participated in this experiment, where ten of them took a calcium supplement for 12 weeks, while the remaining 11 received a placebo. The blood pressure of each subject was measured before and after the 12-week period. Plot the histogram of the variables `Begin` and `End`. Compare the two histograms in terms of their central tendency and the form of their histogram.
2. Download the "Survival.txt" data set from the book website (http://extras. springer.com). This data set appeared in Haberman (1976) and was obtained from the UCI Machine Learning Repository. The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The variables are:

Table 2.5 Height (in inches) and weight (in pounds) for five newborn babies

Observation	Height	Weight
1	18	7.8
2	21	9.1
3	17	8.2
4	16	6.4
5	19	8.8

Fig. 2.37 Boxplot of variable X
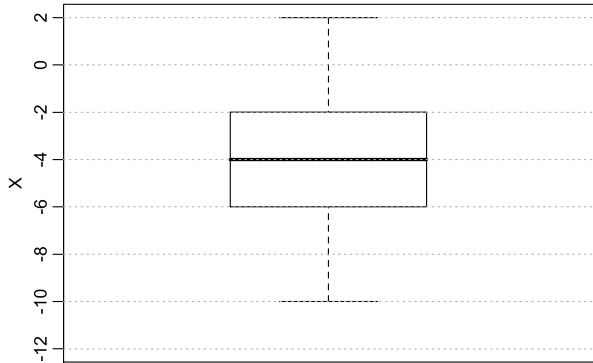
- `Age`: Age of patient at time of operation.
- `Nodes`: Number of positive axillary nodes detected.
- `Status`: Survival status.

  Plot the boxplot for `Age` and the bar graph for `Status`. Plot the histograms for `Nodes` and $\sqrt{\text{Nodes}}$. Which one is more skewed?
3. Show that the total area of rectangles in a density histogram is 1.
4. We have measured the height (in inches) and weight (in pounds) for five newborn babies. Manually calculate the mean and standard deviation of height and weight; show all the steps (Table 2.5).
5. Based on the boxplot in Fig. 2.37, write down the five-number data summary, range and IQR of variable X.
6. Download the "BodyTemperature.txt" from the book website (http://extras. springer.com), and find the five-number data summary for all numerical variables. For numerical variables, provide the histograms and boxplots. Comment on the central tendency and the form of the histograms. Are there any outliers in the data?
7. For the previous question, find the coefficient of variation for `Age` and `Temperature` variable. Show that the coefficient of variation remains the same if we change the units of `Age` to months (i.e., multiplying by 12). Change the body temperature scale to Celsius and recalculate the coefficient of variation. Comment on your findings.
8. The coefficient of variation for variable X is 2. If the sample mean of this variable is 3, what is the sample variance?

9. Download the "AsthmaLOS.txt" data from the book website (http://extras.
   springer.com). Read the description of variables provided in Sect. 2.5. Using R-
   Commander, identify data entry errors for `race` and `owner.type`. Remove
   the corresponding observations (i.e., rows) from the data set. Plot the histogram
   `age` and comment on its shape. For this variable, find the mean, variance, range,
   and IQR.

10. Upload the `Animals` data from the `MASS` package. This data set includes aver-
    age brain and body weights for 28 species of land animals. Plot the histograms
    of the two numerical variables. Next, use the log transformation for both vari-
    ables and plot the histograms again. Comment on the shapes of these new his-
    tograms.

Chapter 3
Exploring Relationships

3.1 Visualizing and Summarizing Relationships Between Variables

In the previous chapter, we focused on using graphs and summary statistics to explore the distribution of individual variables. This chapter is dedicated to using graphs and summary statistics to investigate relationships between two or more variables. Our objective is to develop a high-level understanding of the type and strength of relationships between variables. Note that at this point, we are not making formal conclusions regarding the existence of relationship or whether the relationship, if exists, is strong or not. We do that formally later in this book. Here, we explore the observed data to detect possible relationships and use summary statistics to measure the strength of relationships.

As before, the appropriate tools for exploring data depend on the types of variables. Therefore, this chapter is organized as follows. We start by discussing some techniques for exploring relationships between two numerical variables. Next, we look at some statistics that are commonly used to capture the relationship between two categorical variables. Mainly, we focus on cases where both categorical variables are binary (i.e., variables can take only two possible values). More general situations are discussed later in the book. Finally, we discuss some common techniques for exploring relationships between a categorical variable and a numerical variable.

3.2 Relationships Between Two Numerical Random Variables

We start our discussion of relationships between numerical variables by looking at a data set based on a study conduced by Dr. Fisher from Human Performance Research Center at Brigham Young University [25]. This observational study involved measuring percent body fat as the target variable, along with several explanatory variables such as age, weight, height, and abdomen circumference for

a sample of 252 men. The collected data set `bodyfat` is available online at
http://lib.stat.cmu.edu/datasets/bodyfat.You can also obtain this data set from the
`mfp` package in R. To install this package, enter the following command in *R Console* (the same way you installed `R-Commander`):

```
install.packages("mfp", dependencies=TRUE)
```

Once the package is installed, it can be loaded into R using the following command
(the same way you loaded the `Rcmdr` package):

```
library(mfp)
```

Now you can access `bodyfat` by clicking `Data` → `Data in packages`
→ `Read data set from an attached package` and selecting (double-clicking) `mfp` under `packages`. You can learn more about this data set by looking
at its accompanying help file. In R-Commander, click `Data` → `Active data
set` → `Help on active data set`.

   Suppose that we are interested in examining the relationship between percent
body fat and abdomen circumference among men. Load the `bodyfat` set from the
`mfp` package. Make sure `bodyfat` becomes the active data set and then view it. For
now, we are focusing on two variables, `siri` and `abdomen`. The `siri` variable
shows the percent body fat measurements derived based on body density using Siri's
equation (percent body fat $= 495/\text{density} - 450$). The `abdomen` variable shows the
abdomen circumference in centimeters.

   Both `siri` and `abdomen` are numerical variables. A simple way to visu-
alize the relationship between two numerical variables is with a **scatterplot**.
In R-Commander, click `Graphs` → `Scatterplot` and select `abdomen` for
the `x-variable` (to be represented by the horizontal axis) and `siri` for the
`y-variable` (to be represented by the vertical axis). Under `Options`, uncheck
`Marginal boxplots` and `Smooth line`.

   On the resulting scatterplot, shown in the left panel of Fig. 3.1, the *x*-axis repre-
sents possible values of abdomen circumference, and the *y*-axis represents possible
values of percent body fat. Each point on the graph represents one individual in the
sample. The plot suggests that the increase in percent body fat tends to coincide
with the increase in abdomen circumference. Therefore, the two variables seem to
be related with each other. In this case, the relationship is simply an association
and should not be regarded as causation since the data come from an observational
study.

   As the second example, we examine the relationship between the annual mor-
tality rate due to malignant melanoma for US states and the latitude of their geo-
graphical centers. The data, which are discussed in Fisher and van Belle (1993), are
collected from the population of white males in the US during 1950–1969. You can
obtain this data set, called `USmelanoma`, from the `HSAUR2` package. (Follow the
above steps to install and load the package.) The right panel in Fig. 3.1 shows the
scatterplot for mortality rate for different states and the latitude. (Each point rep-
resent a US state.) The two variables are clearly associated since the increase in
latitude tends to coincide with the decrease in mortality rate.
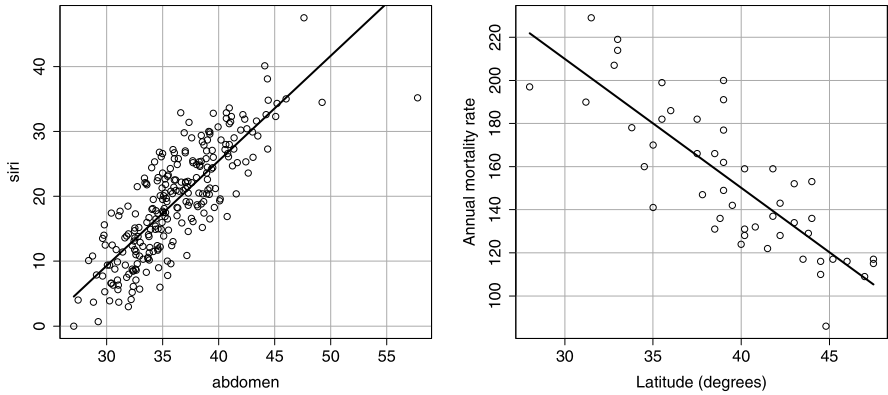
Fig. 3.1 *Left panel*: The scatterplot of percent body fat by abdomen circumference. There is a clear positive linear relationship between the two variables. *Right panel*: The scatterplot of annual mortality rate (per 100,000,000 population) and latitude in degrees. There is a clear negative linear relationship between the two variables

Using scatterplots, we could detect possible relationships between two numerical variables. If a relationship exists, we could also learn about its type. From the scatterplots shown in Fig. 3.1 we can see that changes in one variable coincides with substantial **systematic** changes (increase or decrease) in the other variable. Therefore, the two variables seem to be related. (We make a formal judgement regarding the existence of relationship later.) In these examples, the systemic changes are captured by the straight lines passing through the data points on the graphs. (Around the lines, there are also random changes in observed values without any clear patterns.) When two variables are related, and the overall relationship can be presented by a straight line, we say that the two variables have **linear relationship**. More specifically, we say that percent body fat and abdomen circumference have *positive linear relationship* since increase in one variable tends to coincide with increase in the other one. In contrast, we say that annual mortality rate due to malignant melanoma and latitude have *negative linear relationship* since increase in one variable tends to coincide with decrease in another one. Note that the directions (positive or negative) of these two linear relationships correspond to the slope of the straight lines presenting the overall patterns.

Compare the scatterplots in Fig. 3.1 with the scatterplot in the left panel of Fig. 3.2 for two variables X and Y. Here, changes in one variable, X, also coincide with systematic changes in the other variable, Y. Therefore, the two variables seem to be related. However, we could not use a straight line to capture the overall pattern properly. Instead, the dashed curve seems to provide a better representation for the overall pattern. In this case, as X increases, Y tends to increase first, then it starts decreasing systematically (i.e., ignoring random variations around the dashed
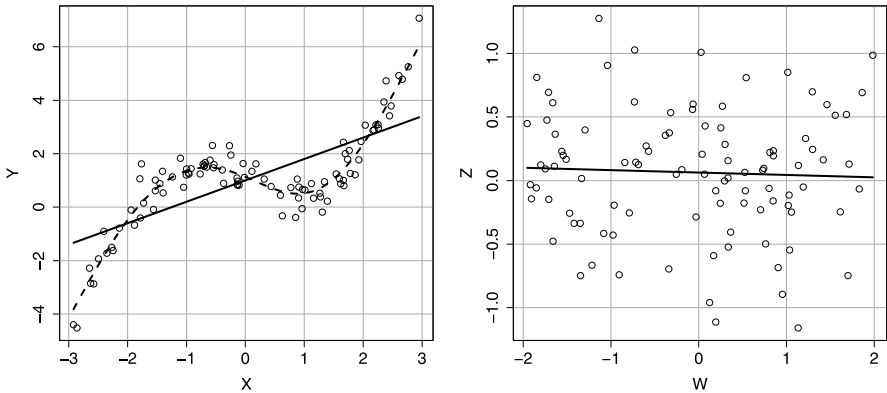
Fig. 3.2 *Left panel*: Scatterplot for two numerical variables with nonlinear relationship. *Right panel*: Scatterplot for two numerical variables that seem to be unrelated
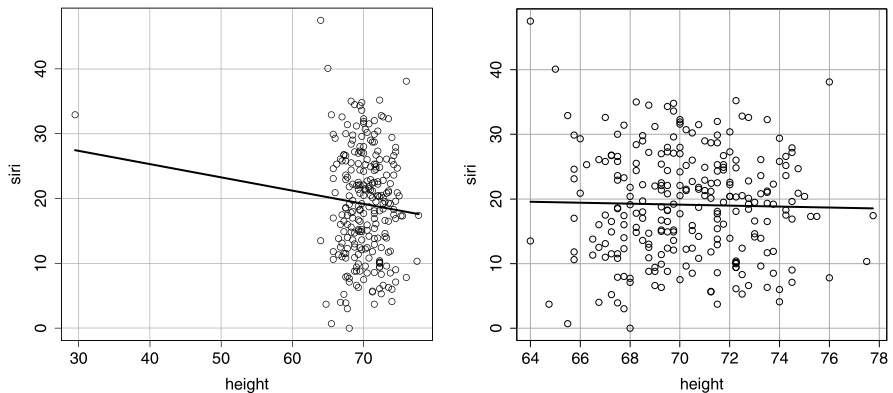
Fig. 3.3 *Left panel*: The scatterplot of percent body fat by height from the `bodyfat` data set. The isolated point at the left of the graph is an outlier, which has a drastic influence on the overall pattern. *Right panel*: The scatterplot of percent body fat by height after removing the outlier. The two variables seem to be unrelated

curve), and finally it tends to increase again. When two variables are related, but we cannot use a straight line to capture the overall relationship, we say that the two variables have **nonlinear relationship**.

Finally, compare the scatterplots in the right panel of Fig. 3.2, with the scatterplots in Fig. 3.1 and the one in the left panel of Fig. 3.2. In this case, changes in one variable, W, does not coincide with substantial systematic changes in the other one, Z. In this case, the overall pattern can be properly represented by a straight line that is almost horizontal. The two variables, W and Z, do not seem to be related.

Now let us examine the relationship between the `height` variable (height in inches) and `siri`. Follow the above steps, but this time select `height` for the

`x-variable`. Figure 3.3 (left panel) shows the resulting scatterplot. The isolated leftmost point is an outlier: a person whose height is around 30 inches, and his percent body fat is relatively hight. In this sample, everyone else's hight is above 60 inches. This is possibly a data entry mistake. Looking at the straight line that represents the overall pattern, there seems to be a negative linear relationship between the two variables. However, the overall pattern (represented by the straight line) is heavily influenced by the outlier. Here, for the illustration purpose, we assume that this is in fact a data entry mistake. Further, we assume that we cannot find the correct values for this subject. Therefore, we remove the outlier from the sample.

In practice, we should never remove an outlier just simply because it does not follow the overall pattern. Some outliers are due to rare events, which provide important information about the distribution of the corresponding variable. Even when we identify a data entry mistake, we should try to correct the mistake and keep the observation if possible.

The right panel in Fig. 3.3 shows the scatterplot after removing the outlier. (Removing cases from a data set was discussed in the previous chapter.) Now, the two variables seem to be unrelated since the straight line, which properly captures the overall pattern, is almost horizontal. (Its slope is almost zero.) In this case, there is no substantial systematic changes (increase or decrease) in percent body fat as hight increases.

Using R-Commander, we can also create pairwise scatterplots of multiple numerical variables. This is useful when we are investigating possible relationships among several variables. To illustrate this, we use the `Protein` data set discussed in the previous chapter. Unlike the above two examples, where there was a single target variable (percent body fat or mortality rate), in this example, we are interested in possible relationships between multiple sources of food. In R-Commander, make sure `Protein` is the active data set, then click `Graphs` → `Scatterplot matrix`, and select `Cereals`, `Eggs`, `RedMeat`, and `Fish` (as shown in Fig. 3.4). Uncheck `Smooth lines` and for the `On Diagonal`, select `Histograms`.

The resulting output, shown in Fig. 3.4, is a matrix of plots. The diagonal plots are histograms of the respective variable. The off-diagonals are the scatterplots of the ith row variable and jth column variable. The 1st row and column correspond to `Cereals`; the 2nd row and column correspond to `Eggs`; the 3rd row and column correspond to `RedMeat`, and the 4th row and column correspond to `Fish`. Consumption of eggs seems to be negatively related to consumption of cereals. That is, in European countries, high consumption of eggs tends to coincide with low consumption of cereals. On the other hand, consumption of eggs has a positive linear
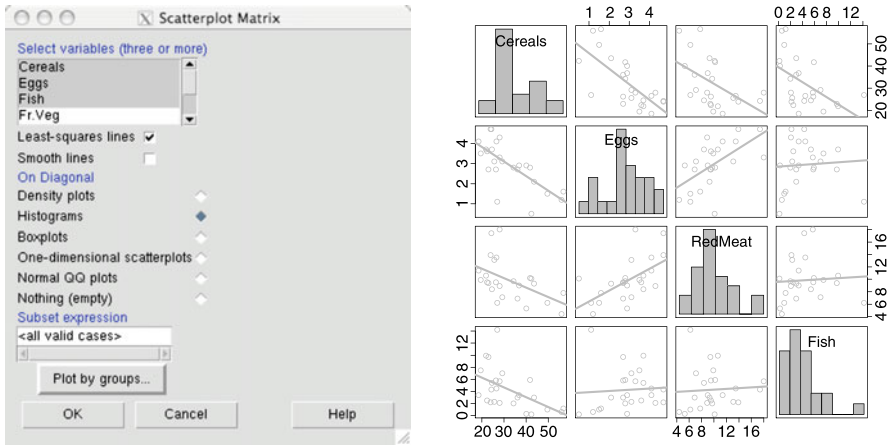
Fig. 3.4 Obtaining and viewing a scatterplot matrix in R-Commander. The diagonal elements are histograms, and the off-diagonals are scatterplots with a trend line

relationship with consumption of red meat. While consumption of fish seems to have a negative linear relationship with consumption of cereals, it does not seems to be linearly related to consumptions of eggs and red meat.

Correlation   As shown in Fig. 3.4, consumption of cereals has negative linear relationships with consumption of eggs, red meat, and fish. Which linear relationship seems to be stronger? To quantify the strength and direction of a *linear* relationship between two numerical variables, we can use **Pearson's correlation coefficient**, r, as a summary statistic. The values of r are always between -1 and $+1$. When r is close to zero, the linear relationship between the two variables is weak. As r moves away from zero either toward -1 or 1, it indicates that the linear relationship is relatively strong. The sign of r shows the direction (negative or positive) of the linear relationship. A positive correlation coefficient means that when one variable increases, the other variable tends to increase too. A negative correlation coefficient, on the other hand, indicates that when one variable increases, the other variable tends to decrease.

Consider a set of observed pairs of values, $(x_n, y_n), (x_n, y_n), \ldots, (x_n, y_n)$, for a sample of n observations. For these data, Pearson's correlation coefficient is calculated as follows:

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

(3.1)

Table 3.1 Height (in inches) and weight (in pounds) for five individuals

Index	Height	Weight
1	62	160
2	71	198
3	65	173
4	73	182
5	60	143
Mean	66.2	171.2
Standard deviation	5.6	21.0

Here, x_i is the observed value of the variable X for the ith observation, and y_i is the observed value of the variable Y for the same observation. For the two variable, $\bar{x}$ and $\bar{y}$ denote the sample means, and s_x and s_y denote the sample standard deviations. If the standard deviations are removed from the denominator, the statistic is called the **sample covariance**,

$$v_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}. \tag{3.2}$$

Therefore,

$$r_{xy} = \frac{v_{xy}}{s_x s_y}. \tag{3.3}$$

For example, suppose that we have measured the height in inches and weight in pounds for five people. We denote height as X and weight as Y. In the following pairs of observations, the first element is the hight of an individual, and the second element is his or her weight:

$$(62, 160), (71, 198), (65, 173), (73, 182), (60, 143).$$

We typically present such data in a tabular format (see Table 3.1). Now we can calculate r as in Table 3.2. Therefore, the sample correlation coefficient between height and weight is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{421.8}{4 \times 5.6 \times 21.0} = 0.89.$$

Here, the numerator 421.8 is obtained by adding up the last column of Table 3.2. Based on our data, height and weight seem to have a strong positive correlation.

We can use R-Commander to calculate the sample correlation coefficient. To calculate r for percent body fat and abdomen circumference, make sure `bodyfat` is

Table 3.2 Calculating Pearson's correlation coefficient for height and weight	Index	x	$x - \bar{x}$	y	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
	1	62	−4.2	160	−11.2	47.04
	2	71	4.8	198	26.8	128.64
	3	65	−1.2	173	1.8	−2.16
	4	73	6.8	182	10.8	73.44
	5	60	−6.2	143	−28.2	174.84

```
Output Window                                                    Submit
> cor(bodyfat[,c("abdomen","siri")], use="complete.obs")
          abdomen      siri
abdomen 1.0000000 0.8134323
siri    0.8134323 1.0000000
```

Fig. 3.5 Obtaining and viewing the correlation between percent body fat and abdomen circumference in R-Commander

```
Output Window                                                    Submit
> cor(Protein[,c("Cereals","Eggs","Fish","RedMeat")], use="complete.ob
           Cereals        Eggs        Fish     RedMeat
Cereals  1.0000000 -0.71243682 -0.52423080 -0.49987746
Eggs    -0.7124368  1.00000000  0.06557136  0.58560895
Fish    -0.5242308  0.06557136  1.00000000  0.06095745
RedMeat -0.4998775  0.58560895  0.06095745  1.00000000
```

Fig. 3.6 Correlation matrix for most of the numerical variables in the `Protein` data set

the active data set, then click `Statistics → Summaries → Correlation matrix`. Select both `abdomen` and `siri`. (You need to hold the *control* key.) The output is in the form of a symmetric matrix called the *correlation matrix*, where the value in row i and column j is the correlation coefficient between the ith and jth variables. As shown in Fig. 3.5, the correlation coefficient for `abdomen` and `siri` is $r = 0.81$, indicating a strong positive linear relationship. The diagonal elements in the correlation matrix are always 1, since a variable is perfectly and positively correlated with itself. Likewise, correlation matrices are symmetric since the order of the variables does not matter: $r_{xy} = r_{yx}$.

We can obtain the correlation matrix for multiple variables following the same steps as described above. Figure 3.6 shows the correlation matrix for `Cereals`, `Eggs`, `Fish`, and `RedMeat`. There is a strong negative linear relationship ($r = -0.71$) between `Cereals` and `Eggs`. The sample correlation coefficient between `Cereals` and `Fish` is also negative ($r = -0.52$). However, the negative linear relationship between `Cereals` and `Fish` is not as strong as the negative linear relationship between `Cereals` and `Eggs`. The correlation coefficient between `Fish` and `RedMeat` is close to zero ($r = 0.06$), which indicates that the linear relationship between these two variables is weak.

Table 3.3 Contingency table of heart attack status by the type of treatment (aspirin versus placebo)

	Heart attack	No heart attack	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

All the studies discussed in this section are observational studies. Therefore, if there is a relationship between two variables, it should not be considered as *causation*. All we can say in such cases is that the two variables are *associated* with each other.

3.3 Relationships Between Categorical Variables

In the previous section, we focused on the relationship between two numerical variables. Here, we discuss techniques for exploring relationships between categorical variables. For example, consider the five-year study to investigate whether regular aspirin intake reduces the risk of cardiovascular disease. The results of this study were published as "Findings from the aspirin component of the ongoing Physicians' health study" in *New England Journal of Medicine* in 1988 [36]. In this randomized experiment, 22071 physicians were randomly divided into two groups: 11037 physicians took an aspirin every other day, while 11034 physicians took a placebo. The investigators then recorded the number of people who suffered a heart attack within the five-year follow-up period.

For each individual in the study, a binary categorical variable indicates whether that person was assigned to the aspirin group or the placebo group. Another binary categorical variable indicates whether the person had a heart attack during the follow-up period. We usually use **contingency tables** to summarize such data. Contingency tables help us to investigate possible relationships between categorical variables. Here, we mainly focus on contingency tables for two categorical variables, each with two possible values (i.e., binary variables). More general forms of contingency tables are discussed in Chap. 10. For the above example, Table 3.3 summarizes the observed data for investigating the relationship between taking aspirin (a binary variable indicating whether a person has been taking aspirin or not) and heart attack (a binary variable indicating whether the person has suffered from heart attack).

Each cell shows the frequency of one possible combination of disease status (heart attack or no heart attack) and experiment group (placebo or aspirin). For example, according to this table, 189 people took placebo and suffered from heart attack. Using these frequencies, we can calculate the **sample proportion** of people who suffered from heart attack in each experiment group separately. There were 11034 people in the placebo group, of which 189 had heart attack. The proportion of people suffered from a heart attack in the placebo group is therefore

$p_1 = 189/11034 = 0.0171$. On the other hand, 104 people out of 11037 in the aspirin group had heart attack. Therefore, the proportion of people suffered from heart attack in the aspirin group is $p_2 = 104/11037 = 0.0094$. This is lower than the corresponding sample proportion in the placebo group. We say that the **risk** (here, the sample proportion is used to measure risk) of heart attack is smaller for those who took aspirin.

Recall that the sample proportion is a commonly used summary statistic for exploring the distribution of categorical variables. For the above example, substantial difference in the sample proportions of heart attack between the two experiment groups indicates that the distribution of heart attack changes from one group to another. This is interpreted as a possible relationship between the two binary variables, one indicating the experiment group, and the other one indicating the disease status. Since the study is designed as a randomized experiment, we can regard the relationship, if exists, as causation. That is, we can conclude that taking aspirin affects the risk of heart attack. Here, of course, we are simply exploring possible relationship between two categorical variables. Later, we formally evaluate whether such relationship exists through hypothesis testing methods.

As mentioned above, substantial difference between the sample proportion of heart attack between the two experiment groups could lead us to believe that the experiment group and disease status are related. Therefore, one way of measuring the strength of the relationship is to calculate the **difference of proportions**, $p_2 - p_1$. Here, the difference of proportions is $p_2 - p_1 = -0.0077$. The proportion of people suffered from heart attack reduces by 0.0077 in the aspirin group compared to the placebo group. We can present this difference as a percentage using the sample proportion (risk) in the placebo group as the baseline:

$$\frac{p_2 - p_1}{p_1} \times 100\% = \frac{-0.0077}{0.0171} \times 100\% = -45\%.$$

Note that the negative sign here indicates the decrease in the risk of heart attack from first group (placebo) to the second group (aspirin). In this example, the risk of heart attack reduces by 45% in the aspirin group compared to the placebo group.

Another common summary statistic for comparing sample proportions is the **relative proportion** p_2/p_1. Since the sample proportions in this case are related to the risk of heart attack, we refer to the relative proportion as the **relative risk**. Here, the relative risk of suffering from heart attack is $p_2/p_1 = 0.0094/0.0171 = 0.55$. This means that the risk of a heart attack in the aspirin group is 0.55 times of the risk in the placebo group.

If the two sample proportions are equal, the relative proportion (risk) is equal to 1, which is interpreted as no relationship between the two categorical variables. Values of the relative proportion away from 1 (either below 1 or above 1) indicate that the relationship is strong.

Instead of comparing sample proportions, p, to measure the strength of relationship between two binary categorical variables, it is more common to compare the

sample odds,

$$o = \frac{p}{1 - p},$$ (3.4)

where p is the sample proportion for the event of interest (e.g., heart attack). The odds of a heart attack in the placebo group, o_1, and in the aspirin group, o_2, are

$$o_1 = \frac{0.0171}{(1 - 0.0171)} = 0.0174,$$

$$o_2 = \frac{0.0094}{(1 - 0.0094)} = 0.0095.$$

We usually compare the sample odds using the **sample odds ratio**

$$OR_{21} = \frac{o_2}{o_1}.$$ (3.5)

The index "21" shows that we are dividing the odds in the second group (here, the aspirin group) by the odds in the first group (here, the placebo group). An odds ratio equal to 1 means that the odds are equal in both groups and is interpreted as no relationship between the two categorical variables. Values of the odds ratio away from 1 (either greater than or less than 1) indicate that the relationship is strong. Note that the odds ratio cannot be negative. Therefore, its smallest possible value is zero.

The odds ratio in the above example is

$$OR_{21} = \frac{o_2}{o_1} = \frac{0.0095}{0.0174} = 0.54.$$

The odds of a heart attack for those taking aspirin regularly is 0.54 times of the odds of heart attack for the placebo group. In other words, taking aspirin regularly seems to reduce the odds of heart attack.

In general, instead of OR_{21}, we could use OR_{12}, i.e., dividing the odds in the first group by the odds in the second group, for comparing the odds. For the above example, however, the interpretation of the results based on OR_{21} is more meaningful since we can talk about the effect of taking aspirin.

As another example, we investigate the relationship between the variable `low`, indicating whether the baby's birth weight was less than 2.5 kg, and the variable `smoke`, indicating the mother's smoking status during pregnancy, using the `birthwt` data set. In R-Commander, load the `birthwt` data set and make sure the variables `low` and `smoke` are converted to factors (categorical) variables. (R-Commander automatically considers these variables as numerical since they have numerical codings.) Now, create a 2 × 2 (two rows and two columns) contingency table. Click `Statistics` → `Contingency tables` → `Two-way`

Fig. 3.7 Creating the
contingency table for smoke
and low
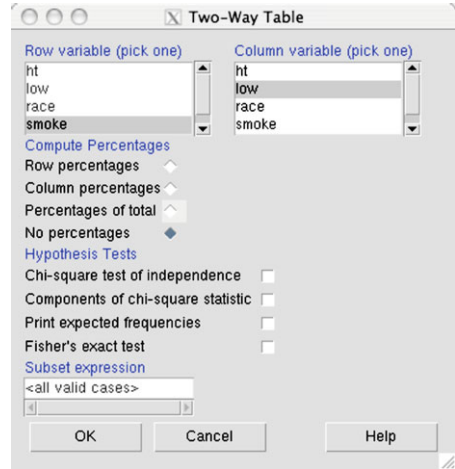
Fig. 3.8 Contingency table
for smoke and low

table. Select smoke for the Row variable and low for the Column variable, as in Fig. 3.7. For now, uncheck Chi-square test of independence under Hypothesis Tests.

The output is the 2×2 contingency table shown in Fig. 3.8. The proportion of low weight babies (low=1) among nonsmoking (during pregnancy) mothers is $p_1 = 29/(86 + 29) = 0.25$, whereas the proportion among smoking mothers is $p_2 = 30/(44 + 30) = 0.41$. Therefore, the proportion of low weight babies is $p_2 - p_1 = 0.16$ higher for smoking mothers. There is a 64% increase in the risk of having low-weight babies for mothers who smoke during pregnancy compared to those who do not smoke.

$$\frac{0.41 - 0.25}{0.25} \times 100 = 64.$$

The relative risk of having a low-weight baby is $p_2/p_1 = 0.41/0.25 = 1.64$, which means that the risk of having a low-weight baby is 1.64 times higher among smoking mothers compared to nonsmoking mothers. Furthermore, the odds of hav-

ing a low-weight baby for the two groups are

$$\text{nonsmoking:} \quad o_1 = \frac{0.25}{(1 - 0.25)} = 0.33,$$

$$\text{smoking:} \quad o_2 = \frac{0.41}{(1 - 0.41)} = 0.69.$$

Therefore, the odds ratio is

$$OR_{21} = \frac{o_2}{o_1} = \frac{0.69}{0.33} = 2.1,$$

which means the odds of having low-weight baby is 2.1 times higher when mothers smoke during pregnancy.

3.4 Relationships Between Numerical and Categorical Variables

Very often, we are interested in the relationship between a categorical variable and a numerical random variable. When the sample size is small, we can visualize the relationship by simply creating dot plots of the numerical variable for different levels of the categorical variable. As an example, we use the `cabbages` data set available from the `MASS` package. Figure 3.9 shows the dot plots of ascorbic acid (one form of vitamin C) content (numerical) by cultivar (categorical). The categorical variable has two possible categories: c39 and c52.

Figure 3.9 shows that the distribution of vitamin C content is different between the two cultivars. More specifically, the central tendency for the observed values in the c39 group is around 50, whereas the central tendency for the c59 group is around 65.

> In general, we say that two variables are related if the distribution of one of them changes as the other one varies.

In the above example, the two variables, vitamin C content and cultivar, seem to be related. We can use R-Commander to create a dot plot (a.k.a. strip chart) similar to the one presented in Fig. 3.9. Of course, R-Commander uses the horizontal axis for the categorical variable and the vertical axis for the numerical variables. Make sure the data set `cabbages` from `MASS` is the active data set, then click `Graphs` → `Strip chart`. For the `Factors`, choose `Cult` (cultivar of the cabbage), and for the `Response Variable`, choose `VitC` (vitamin C content). The resulting plot is shown in Fig. 3.10. Here, multiple observations with the same value of the numerical variable are stacked toward the right. Overall, vitamin C content tends to be higher in the c52 group compared to the c39 group.

Fig. 3.9 *Dot plots* of vitamin C content (numerical) by cultivar (categorical) for the cabbages data set from the MASS package
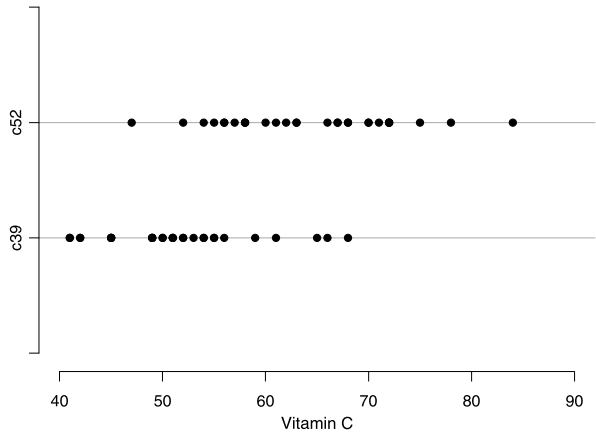
Fig. 3.10 Strip chart (*dot plot*) for vitamin C content (VitC) by cultivar (Cult) from the canbbages data set using R-Commander
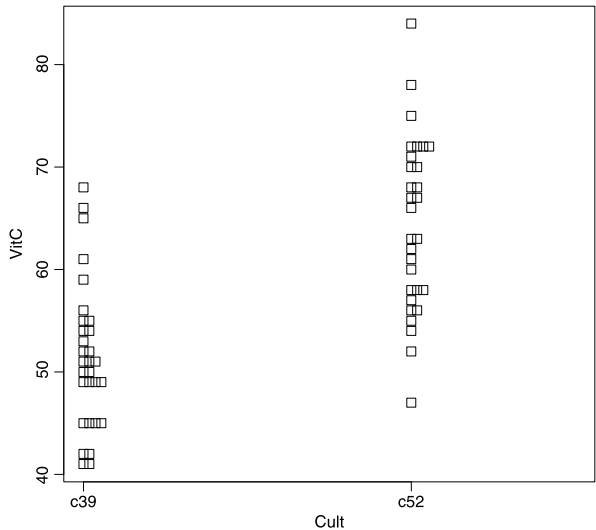
A more common way of visualizing the relationship between a numerical variable and a categorical variable is to create boxplots, as opposed to dot plots, of the numerical variable for different values of the categorical variable. This is especially useful when the sample size is large. By focusing on some key aspects of the distributions, namely the five-number summaries, boxplots make the patterns easier to detect. In R-Commander, click Graphs → Boxplot; select VitC as the Variable. Then click on Plot by groups button and in the resulting window, select Cult as the Groups variable. The resulting plot is shown in Fig. 3.11, which suggests that vitamin C content tends to be higher in the c52 group compared to the c39 group. This is indicative of a possible relationship between these two variables.

We can measure changes in the distribution of the numerical variable by obtaining its summary statistics for different levels of the categorical variable. In

Fig. 3.11 Boxplot of
vitamin C content for
different cultivars
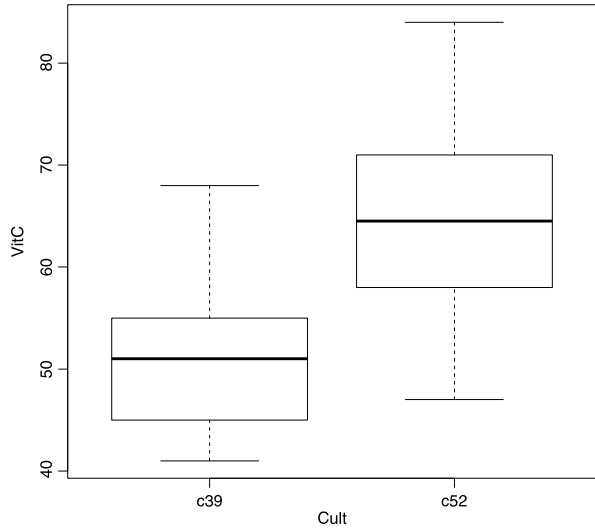
Fig. 3.12 Summary statistics
of vitamin C content (`VitC`)
by cultivar (`Cult`) from the
`cabbages` data set
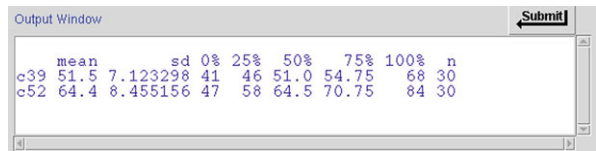
R-Commander, click `Statistics` → `Summaries` → `Numerical sum-`
`maries` and select `VitC` as the `Variable`. Then click on the `Summarize by`
`groups` button and choose `Cult` as the `Groups variable`. This way, the sum-
mary statistics will be calculated for the two groups (c39 and c59) separately. The
results are shown in Fig. 3.12. As we can see, the sample mean and the sample
median are substantially different between the two groups.

As mentioned above, we consider two variables as related when the distribu-
tion of one variable changes for different values of the other variable. Distribution
change could refer to the change of location, spread, or in general, form of a distribu-
tion. However, it is more common to focus on the change of location. Especially, it is
common to use the **difference of means** when examining the relationship between
a numerical variable and a categorical variable. In the above example, the difference
of means of vitamin C content is $64.4 - 51.5 = 12.9$ (see Fig. 3.12) between the
two cultivars. Later, we will use this measure to formally evaluate our hypothesis
regarding the relationship between cultivar of the cabbage and its vitamin C content.

When the categorical variable has multiple levels (categories), it is easier to com-
pare the means across different levels using the **plot of means**. For example, in the
previous chapter, we created a categorical variable called `weight.status` based
on BMI values in the `Pima.tr` data set. This variable had four categories: "Un-
derweight", "Normal", "Overweight", and "Obese". Here, we would like to inves-
tigate how blood pressure `bp` changes with `weight.status`, which is an *ordi-*

Fig. 3.13 Plotting the means
of `bp` for different weight
groups (which are defined
based on BMI). The
relationship between these
two variables is nonlinear
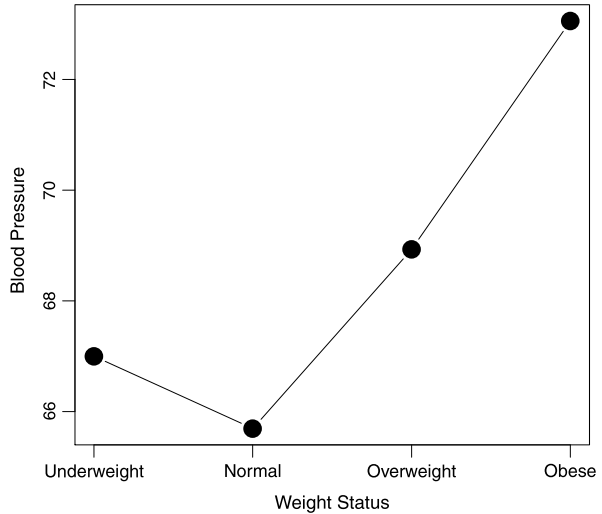
nal variable. In R-Commander, click `Graphs → Plot of means` and select
`weight.status` as the `Factors` and `bp` as the `Response Variable`. For
now, choose `no error bars`. The resulting graph (Fig. 3.13) shows the plot of
the mean blood pressure for each group. This plot shows that compared to the Nor-
mal group, the average blood pressure increases for both Underweight and Over-
weight group. The Obese group has the highest blood pressure average. Also, note
that as we move toward higher levels of weight group, average blood pressure first
decreases and then increases. The issue of high blood pressure among underweight
people is a well-studied phenomenon (see, for example, [22]).

3.5 Advanced

In this section, we discuss some useful R functions for examining the relationship
between two variables.

Two Numerical Variables   We start by installing and loading the `mfp` package,
which includes the `bodyfat` data set:

```
> install.packages("mfp", dependencies = TRUE)
> library(mfp)
> data(bodyfat)
```

An easy way to visualize the relationship between two numerical variables is
to use scatterplots. In R, you can use the `plot()` function for this purpose. For
example, the following code creates the scatterplot of percent body fat (`siri`) by
abdomen circumference (`abdomen`):

```
> plot(bodyfat$abdomen, bodyfat$siri,
+      xlab = "Abdomen", ylab = "Percent Body Fat")
```

The first parameter to the `plot()` function is the variable to be represented by the x-axis, and the second parameter is variable to be represented by the y-axis.

Next, we use Pearson's correlation coefficient to measure the strength and direction of the linear relationship between the two numerical variables. For this, we use the `cor()` function:

```
> cor(bodyfat$abdomen, bodyfat$siri)

[1] 0.8134323
```

The resulting correlation coefficient of $r = 0.81$ along with the scatterplot suggests that there is evidence of a strong positive linear relationship between percent body fat and abdomen circumference. Likewise, the `cor` function can be used to obtain the correlation matrix for multiple variables:

```
> cor.matrix <- cor(bodyfat[, c("siri", "weight",
+      "height", "abdomen")])
> round(cor.matrix, 2)

          siri weight height abdomen
siri      1.00   0.61  -0.09    0.81
weight    0.61   1.00   0.31    0.89
height   -0.09   0.31   1.00    0.09
abdomen   0.81   0.89   0.09    1.00
```

Here, we are using the combine function `c()` to specify that we want the correlation matrix for the columns labeled "siri", "weight", "height", and "abdomen". Then, the `round()` function is used to round the output to 2 decimal places.

Two Categorical Variables   To examine possible relationship between two categorical variables, we usually use contingency tables. From contingency tables we can obtain the proportions, relative proportions (risk), odds, and odds ratio. For instance, try creating the contingency table for `smoke` by `low` from the `birthwt` data set with the `table()` function:

```
> library(MASS)
> data(birthwt)
> table(birthwt$smoke, birthwt$low)

    0  1
  0 86 29
  1 44 30
```

The first parameter to the `table()` is the row variable, and the second parameter is the column variable.

Relationship Between a Numerical Variable and Categorical Variable   To visualize the relationship between a numerical variable, we can simply create a dot plot using the `plot()` function. The following code plots the dot plot of birthweight (`bwt`) by the smoking status (`smoke`) of mothers:

```
> plot(birthwt$smoke, birthwt$bwt)
```

The `plot()` function, however, shows the levels of the categorical variable as integers.

Using dot plots is usually recommended for data sets with a small sample size. In general, it is better to use boxplots to visualize the relationship between a numerical variable and categorical variable. For instance, create a boxplot of `bwt` for each level of `smoke`:

```
> boxplot(bwt ~ smoke, ylab = "Birthweight",
+      data = birthwt, xlab = "Smoking Status",
+      main = "Birthweight by Smoking Status")
```

The first parameter is a formula, using the $\sim$ symbol to plot `bwt` (the response variable) by `smoke` (the explanatory variable). Note that boxplot shows the actual levels of the categorical variable as opposed to their corresponding numerical values.

The summary statistics for `bwt` can be calculated for each level of `smoke`. Using the `which()` function, we can find the indices of smoking mothers (`smoke=1`) in the `birthwt` data set:

```
> smoke.ind <- which(birthwt$smoke == 1)
```

Now, obtain the summary statistics of this group:

```
> summary(birthwt$bwt[smoke.ind])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    709    2370    2776    2772    3246    4238

> sd(birthwt$bwt[smoke.ind])

[1] 659.6349
```

A more convenient way to obtain summary statistics by group is to use the `by` function.

```
> by(birthwt$bwt, birthwt$smoke, summary)

birthwt$smoke: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1021    2509    3100    3056    3622    4990
----------------------------------
```

```
birthwt$smoke: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    709    2370    2776    2772    3246    4238
```

The general form of the `by` function is `by(data, group, function)`. The first parameter of this function specifies the numerical variable (here, `bwt`), the second parameter specifies the indicator variable to identify the groups (here, `smoke`), and the last parameter (here, `summary`) specifies the function we want to apply to different groups. The following code returns the standard deviation of birthweight for different levels of `ht` (hypertension history):

```
> by(birthwt$bwt, birthwt$ht, sd)

birthwt$ht: 0
[1] 709.4418
------------------------------------
birthwt$ht: 1
[1] 917.3617
```

3.6 Exercises

1. Using the measurements (see Table 3.4) of height (in inches) and weights (in pounds) for 5 newborn babies, calculate the sample covariance and sample Pearson's correlation coefficient between height and weight; show all the steps.
2. Using the "BodyTemperature.txt" data set, create the scatterplot for body temperature by heart rate. Describe the pattern and comment on possible relationship between the two variables. Find the correlation coefficient between body temperature and heart rate. Finally, create boxplots of body temperature for men and women separately. Which one tends to be higher? Which one has higher dispersion?
3. In an article published in the July 2010 issue of the journal Pediatrics, Dr. Nafiu and colleagues argue that children's neck circumference instead of BMI should be used as a simple proxy for percent body fat. For the `bodyfat` data set, use graphs and summary statistics to investigate the relationship between BMI

Table 3.4 Height (in inches) and weight (in pounds) for five newborn babies

Observation	Height	Weight
1	18	7.8
2	21	9.1
3	17	8.2
4	16	6.4
5	19	8.8

Table 3.5 Frequencies of people with heart disease for different levels of snoring based on a sample of 2484 subjects

Snoring Severity	Heart Disease	Total
Never	24	1379
Occasionally	35	638
Nearly every night	21	213
Every night	30	254

and percent body fat (`siri`), and the relationship between neck circumference (`neck`) and percent body fat among adult men. Which one seems to have a stronger relationship with percent body fat?

4. Consider the data (Table 3.5) based on an epidemiological survey of 2484 people to investigate snoring as a risk factor for heart attack. The data set is collected by Norton and Dunn [24] and is discussed in Categorical Data Analysis by Agresti [1]. The first column shows "snoring severity" as reported by the spouses of subjects. The second column shows the number of people suffer from heart disease for each level of snoring severity, and the third column shows the total number of people for each snoring severity level. Create two groups based on snoring severity: Group 1 are those who never snore, and Group 2 are those who snore. Write down the contingency table and calculate the proportion of people with heart disease for each group. Then, find difference of proportions, relative risk, and odds ratio for heart disease in order to compare the two groups.

5. Follow the steps described at the beginning of this chapter to load the `GBSG` from the `mfp` package. Make sure `GBSG` is the active data set, then click `Data` → `Active data set` → `Help on active data set` to see the description of random variables. Using this data set, we want to investigate the effect of hormonal treatment on recurrence free survival of breast cancer patients. Here, `htreat` is a binary categorical variable, which indicates whether the subject has received hormonal therapy (`htreat = 1`) or not (`htreat = 0`). The `cens` variable is also binary indicating whether the patient had at least one recurrence of the disease or died. For patients who had at least one recurrence and/or did not survive, `cens=1` and `rfst` shows their actual survival time (in days). For patients who survived recurrence free during the study, `cenc=0`, create a new random variable called `rfs` (recurrence free survival) such that `rfs="No"` if the patient had at least one recurrence or died (i.e., `cenc=1`) and `rfs="Yes"` otherwise. Obtain the frequency table for the `rfs` variable. Next, create a 2×2 contingency table for `htreat` (hormonal treatment) and `rfs` (recurrence free survival). Find the relative risk, odds, and odds ratio for the two treatment groups.

6. For the `birthwt` data set, investigate the relationship between the history of hypertension (`ht`) and the risk of having low birthweight babies.

7. Layman et al. (1986) [15] investigated the effect of iontophoretic treatment with the nerve conduction-inhibiting chemical vincristine on elderly patients complaining of post-herpetic neuralgia. There were eighteen patients in the study.

The patients were interviewed six weeks after the initial treatment and were asked if the pain had been reduced. The data set, "neural.txt", for this study was obtained from [9] and is available online from the book website (http://extras.springer.com). There are five variables in this data:

- `Pain`: A binary variable indicating whether the pain was eased or not.
- `Treatment`: A binary variable indicating whether the patient underwent treatment.
- `Age`: The age of the patient in completed years.
- `Gender`: The gender of the patient: M (male) or F (female).
- `Duration`: The pretreatment duration of symptoms (in months).

Use contingency tables to examine the relationship between `Pain` and `Treatment`.

8. Use boxplots to investigate the relationship between `type` and three numerical variables, `bmi`, `bp`, `glu`, from the `Pima.tr` data set. Comment on your findings.

9. Using a plot of means, show how mean birthweight (`bwt`) changes among different races (`race`) in the `birthwt` data set. Find the sample mean and sample standard deviation of `bwt` for each race separately.

10. In R-Commander, load the `chickwts` data set from the `datasets` package. (Click Data → Data in packages → Read data set from an attached package.) The `chickwts` data set was collected based on an experiment to measure the effectiveness of various feed supplements (`feed`) on the growth rate (`weight`) of chickens. Find the five-number summary statistics for each feed type separately. Use boxplots and a plot of means to visualize the difference between feed types. Which feed type results in the lowest growth rate on average?

Chapter 4
Probability

4.1 Probability as a Measure of Uncertainty

In the previous chapters, we used plots and summary statistics to learn about the distribution of variables and to investigate their relationships. In the birthweight example, from a sample of 189 newborn babies, we found that average birthweight is 2944 grams. Also, we found that the risk of having a low-weight baby is 1.64 times higher among smoking mothers compared to nonsmoking mothers. In scientific studies, we would like to generalize our findings from a sample of observations to the whole population (here, all newborn babies). For example, we would like to comment on the average birthweight for all newborn babies. More generally, we would like to comment on the distribution of birthweight (e.g., its location, its spread, and its form) in this population. Also, we want to know whether our findings about the relationship between smoking and birthweight can be generalized to the whole population.

As discussed earlier, we always remain uncertain about the true distributions and relationships in the population since we almost never have access to all of its members. Furthermore, our findings based on the observed sample can change if different samples from the population were obtained. Therefore, when we generalize our findings from a sample to the whole population, we should explicitly specify the extent of our uncertainty.

The focus of this chapter is the use probability as a measure of uncertainty. In what follows, we use coin tossing, die rolling, and genetic variation as running examples. Since the latter involves some terminologies that might not be familiar to all readers, we provide a brief review of some common terms in statistical genetics in the next section.

4.2 Some Commonly Used Genetic Terms

A *gene* is a segment of double-stranded DNA, which itself is made of a sequence of four different nucleotides: adenine (A), guanine (G), thymine (T), or cytosine (C).

A DNA segment can be specified as a sequence of these four letters; for example, {TAGCAAT}. Genetic variation is caused by changes in the DNA sequence of a gene. Single Nucleotide Polymorphisms (SNPs) are the most common type of genetic variation. SNPs (pronounced "snips") occur when a single nucleotide is replaced by another one. An example of a SNP would be replacing "G" in the sequence {TA*G*CAAT} by "T" to create {TA*T*CAAT}.

The alternate forms of a gene ares called *alleles*. In the above example, the alleles could be denoted as T and G. Alleles are responsible for variation in *phenotypes*. Phenotypes, in general, are observable traits, such as eye color, disease status, and blood pressure, due to genetic factors and/or environmental factors (e.g., diet, smoking, sun exposure). Throughout this book, we only consider genes that are bi-allelic (two possible alleles). We denote the genes with bold face letters (e.g., **A**) and the two different alleles as capital and small letters (e.g., A and a).

Genetic materials are stored on *chromosomes*. Human somatic cells have two copies of each chromosome (one inherited from each parent); hence, they are called *diploid*. Each pair of similar chromosomes are called *homologous* chromosomes. The *genotype* (i.e., genetic makeup) of an individual for the bi-allelic gene **A** can take one of the three possible forms: *AA*, *aa*, or *Aa*. The first two genotypes, *AA* and *aa*, are called *homozygous*, which means the same version of the allele was inherited from both parents. That is, both homologous chromosomes have the same allele. The last genotype, *Aa*, is called *heterozygous*, which means different alleles were inherited.

The presence of a specific allele does not always result in its corresponding trait (a characteristic such as eye color). Some alleles are *recessive*, producing their trait only when both homologous chromosomes carry that specific variant. On the other hand, some alleles are *dominant*, producing their traits when they appear on at least one of the homologous chromosomes. For example, suppose that the allele a for gene **A** is responsible for a specific disease. Furthermore, assume that a is a recessive allele. Then, only a person with genotype *aa* will be affected by the disease. Individuals with genotype *AA* or *Aa* will not have the disease. For example, Cystic Fibrosis (CF) is an inherited chronic disease affecting the lungs and digestive system. The gene causing CF is recessive, which means a person can carry the gene without having the disease. That is, if we denote the allele causing CF as a and the normal allele as A, only people with *aa* genotype have CF. People with *Aa* genotype are carriers.

4.3 The Sample Space

We begin our discussion of probability with the concept of **randomness**. A phenomenon is called *random* if its outcome (value) cannot be determined with certainty before it occurs. For example, when a coin is tossed, the outcome is either heads H or tails T, but unknown before the coin is tossed. Die rolling is also a random phenomenon, whose outcome is an integer from 1 to 6, unknown before the die is rolled. Likewise, for a bi-allelic gene **A**, the possible alleles are A and a, and the

Fig. 4.1 Tree diagram of possible outcomes for the combination of genotype and smoking status. Genotypes (*AA*, *Aa*, and *aa*) are represented by the first set of branches and smoking status (*Y* for smokers and *N* for nonsmokers) is represented by the second set of branches

S={AA-Y, AA-N, Aa-Y, Aa-N, aa-Y, aa-N}
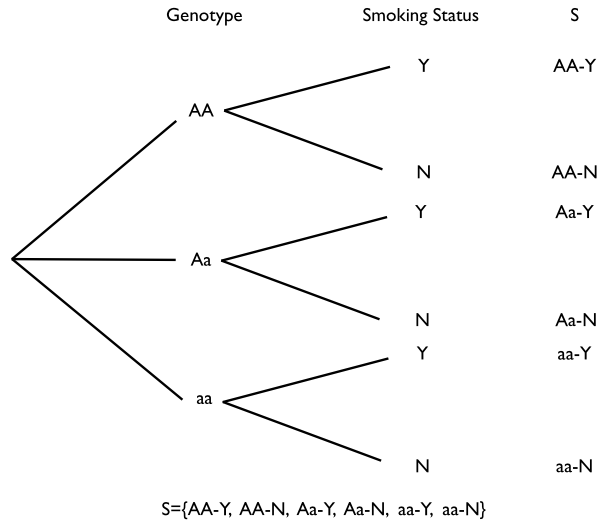
possible corresponding genotypes are *AA*, *Aa*, and *aa*. The collection of all possible outcomes is denoted *S* and is called the **sample space**. The sample spaces for the above random phenomena are:

$$\text{Coin tossing:} \quad S = \{H, T\},$$
$$\text{Die rolling:} \quad S = \{1, 2, 3, 4, 5, 6\},$$
$$\text{Bi-allelic gene:} \quad S = \{A, a\},$$
$$\text{Genotype:} \quad S = \{AA, Aa, aa\}.$$

The sample space might include an infinite number of possible outcomes. For example, the value of blood pressure is random since it cannot be determined with certainty before measuring it. The corresponding sample space for blood pressure values is (theoretically) the set of positive real numbers, which is infinite. In this chapter, we focus on random phenomena with finite number of possible outcomes.

For a complex random phenomenon that is a combination of two or more other random phenomena, it might be easier to view the sample space with **tree diagrams**. For example, suppose that we suspect that gene **A** is related to a specific disease, but genetic variation alone does not determine the disease status. Rather, it affects the risk of the disease. Further, we suspect that smoking (an environmental factor) is also related to the disease. In this case, the random phenomenon we are interested in is the combination of genotype and smoking status ("Y" for smoking and "N" for not smoking). All possible combinations (i.e., sample space) are identified using the tree diagram in Fig. 4.1. The tree starts at left from its root. The first set of branches (originating from the root) represents possible genotypes (*AA*, *Aa*, and *aa*), and the second set represents the smoking status. Following each branch from root to tip, we obtain the sample space $S = \{AA - Y, AA - N, Aa - Y, Aa - N, aa - Y, aa - N\}$. For example, $Aa - Y$ represents the outcome of having heterozygous genotype and smoking.

4.4 Probability Measure

To each possible outcome in the sample space, we assign a probability P, which is a number between 0 and 1, and it represents how certain we are about the occurrence of the corresponding outcome. As the probability of an outcome increases, we become more certain that it will occur. The total probability of all outcomes in the sample space is always 1.

For an outcome o, we denote the probability as $P(o)$, where $0 \le P(o) \le 1$. For the above examples, we have:

Coin tossing: $P(H) + P(T) = 1,$
Die rolling: $P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1,$
Bi-allelic gene: $P(A) + P(a) = 1,$
Genotype: $P(AA) + P(Aa) + P(aa) = 1.$

Probability is not defined for outcomes outside of the sample space. For example, the probability of rolling a 7 is not defined for a normal die.

Because the total probability of all outcomes in the sample space is always 1, if the outcomes are equally probable, the probability of each outcome is $1/n_S$, where n_S is the number of possible outcomes, i.e., the size of sample space.

If we believe that the coin is balanced (fair, symmetric) so that heads and tails are equally probable, then $P(H) = P(T) = 1/2$. When rolling a balanced die (i.e., we believe all the 6 possible numbers are equally probable), $P(1) = P(2) = \cdots = P(6) = 1/6$. Likewise, if we believe the two alleles are equally probable, we have $P(A) = P(a) = 1/2$. We do not however assume equal probabilities for the genotypes. To find the probability of each genotype, we first need to define **events**.

An **event** is a subset of the sample space S. A possible event for die rolling is $E = \{1, 3, 5\}$. This is the event of rolling an odd number. For the genotype example, $E = \{AA, aa\}$ is the event that a person is homozygous.

For the coin tossing example, $E_1 = \{H\}$, $E_2 = \{T\}$, $E_3 = \{H, T\}$, and $E_4 = \{\}$ are all the possible events. Note that these include the sample space $E_3 = \{H, T\} = S$ and the empty set $E_4 = \{\}$, which we denote $\emptyset$. In general, the sample space S and the empty set are possible events for any random phenomenon.

An event occurs when any outcome within that event occurs. For instance, if a person's genotype is AA, the homozygous event, $E = \{AA, aa\}$, has occurred. For die rolling example, define $E_1 = \{1, 2, 3\}$ (the outcome is less than 4) and $E_2 = \{1, 3, 5\}$ (the outcome is an odd number). If we roll the die and the outcome is 1, then both E_1 and E_2 have occurred. (The outcome is less than 4 and is an odd number.)

Fig. 4.2 Tree diagram of possible genotypes for a bi-allelic SNP. The first set of branches represents possible alleles for chromosome 1 (Ch_1) and the second set of branches represents possible alleles for chromosome 2 (Ch_2). Since the labels on homologous chromosomes are arbitrary, we can write the sample space as $S = \{AA, Aa, aa\}$
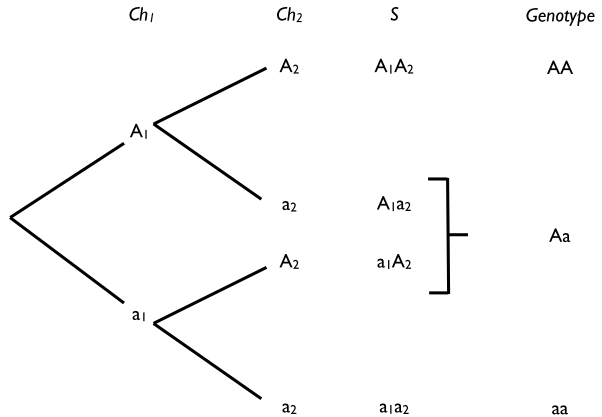
We denote the probability of event E as $P(E)$, where $0 \leq P(E) \leq 1$. The probability of an event is the sum of the probabilities for all individual outcomes included in that event. For instance, when rolling a symmetric die, the probability of event $E = \{1, 3, 5\}$ is $P(E) = 1/6 + 1/6 + 1/6 = 1/2$. The probability of the sample space is one, $P(S) = 1$, since it includes all the possible outcomes, and the total probability of all outcomes is 1. On the other hand, the probability of the empty set is $P(\phi) = 0$ since it does not include any of the possible outcomes.

Now let us apply these principles to determine the genotype probabilities for the bi-allelic gene **A**. So far, we have treated the genotype as a random phenomenon with three possible outcomes. Alternatively, we can treat the genotype as the combination of alleles on homologous chromosomes (one inherited from each parent). This way, the allele type on each chromosome is regarded as a random phenomenon by itself, and the genotype is regarded as the combination of two random phenomena. In the tree diagram in Fig. 4.2, the first set of branches represents the possible alleles for chromosome 1 (Ch_1). Likewise, the second set represents the possible alleles for chromosome 2 (Ch_2). By following the branches from the root to the tip, we can obtain the possible genotypes. Now, if we assume that all outcomes in the sample space are equally probable, then $P(A_1 A_2) = P(A_1 a_2) = P(a_1 A_2) = P(a_1 a_2) = 1/4$.

The labels for homologous chromosomes are arbitrary; we do not distinguish between genotypes $A_1 a_2$ and $a_1 A_2$. Therefore, we can create three new events: $AA = \{A_1 A_2\}$, $Aa = \{A_1 a_2, a_1 A_2\}$, and $aa = \{a_1 a_2\}$. The probabilities for these events are then $P(AA) = 1/4$, $P(Aa) = 1/4 + 1/4 = 1/2$, and $P(aa) = 1/4$. We can now treat AA, Aa, and aa as possible outcomes for the genotype events. The probabilities for these three outcomes are $1/4$, $1/2$, and $1/4$, respectively.

4.5 Complement, Union, and Intersection

In this section, we discuss some common operations on random events along with some general rules of probability. For this, we use two running examples. First,

we consider the die rolling example presented in the form of a Venn diagram in Fig. 4.3. All the possible outcomes are contained inside the sample space S, which is represented by the rectangle. We define two events. The event M (shown as a triangle) occurs when the outcome is less than 4. The event N (shown as an oval) occurs when the outcome is an odd number. In this example, $P(M) = \frac{1}{2}$ and $P(N) = \frac{1}{2}$.

For the second example, we consider a bi-allelic gene **A** with two alleles A and a. We assume that allele a is recessive and causes a specific disease. Then only people with the genotype aa have the disease. We can define four events as follows:

$$
\begin{array}{lll}
\text{The heterozygous event:} & HM = \{AA, aa\}, \\
\text{The heterozygous event:} & HT = \{Aa\}, \\
\text{The no-disease event:} & ND = \{AA, Aa\}, \\
\text{The disease event:} & D = \{aa\}.
\end{array}
$$

A schematic representation of these events is provided in Fig. 4.4. The shaded area shows the disease event (D); whereas the unshaded area shows the no-disease event (ND). The area with shaded border lines shows the homozygous event (HM). The remaining part of the sample space, which includes the outcome Aa only, corresponds to the heterozygous event. Assume that the probabilities for different genotypes are $P(AA) = 0.49$, $P(Aa) = 0.42$, and $P(aa) = 0.09$. (The sum of probabili-

Fig. 4.3 A schematic representation for the die rolling example. M is the event that the outcome is a number less than 4, and N is the event that the outcome is an odd number
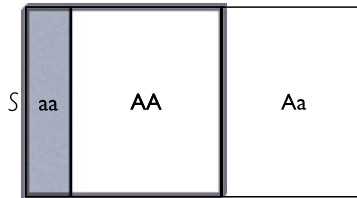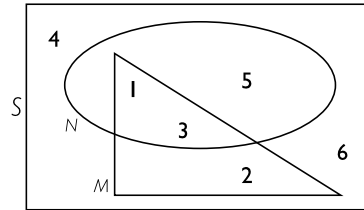
Fig. 4.4 A schematic representation for a bi-allelic gene with a recessive allele a that causes a specific disease. The *shaded area* shows the disease event (D). The *unshaded area* shows the no-disease event (ND). The *area with shaded border lines* shows the homozygous event (HM). The *remaining part* of the sample space, which includes the outcome Aa only, corresponds to the heterozygous event

ties is 1.) Then,

$$P(HM) = 0.49 + 0.09 = 0.58,$$

$$P(HT) = 0.42,$$

$$P(ND) = 0.49 + 0.42 = 0.91,$$

$$P(D) = 0.09.$$

4.5.1 Complement

For any event E, we define its **complement**, E^c, as the set of all outcomes that are in the sample space S but not in E. Schematically, the complement is the set of outcomes outside the region defined for the event but inside the sample space. In the die rolling example, the complement of M is $M^c = \{4, 5, 6\}$ (i.e., the outcome is greater than or equal to 4). This is the set of all outcomes inside the rectangle but outside of the triangle. The complement of N is $N^c = \{2, 4, 6\}$ (i.e., the outcome is an even number). This is the set of all outcomes inside the rectangle but outside of the oval.

For the gene-disease example, the complement of the homozygous event $HM = \{AA, aa\}$ is the heterozygous event $\{Aa\}$; we show this as $HM^c = HT$. Likewise, the complement of the disease event, $D = \{aa\}$, is the no-disease event, $ND = \{AA, Aa\}$; we show this as $D^c = ND$.

The complement of an event is an event by itself so we can talk about its probability.

> The probability of the complement event is 1 minus the probability of the event:
>
> $$P(E^c) = 1 - P(E). \tag{4.1}$$

For the event that the outcome is an odd number, we have

$$P(N^c) = 1 - P(N) = 1 - \frac{1}{2} = \frac{1}{2},$$

which is equal to the probability that the outcome is an even number. In the gene-disease example, the probability of the complement of the homozygous event is $P(HM^c) = 1 - P(HM) = 1 - 0.58 = 0.42$. This is, of course, equal to the probability of the heterozygous event $P(HT) = 0.42$. Likewise, the probability of the complement of the disease event is $P(D^c) = 1 - P(D) = 1 - 0.09 = 0.91$ and equal to the probability of the no-disease event, $P(ND) = 0.91$.

The **odds** of an event shows how much more certain we are that the event occurs than we are that it does not occur. For event E, we calculate the odds as follows:

$$\frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}.$$

For the gene-disease example, the odds for ND (i.e., not having the disease) are

$$\frac{P(ND)}{P(ND^c)} = \frac{P(ND)}{1 - P(ND)} = \frac{0.91}{1 - 0.91} = 10.11.$$

Therefore, it is almost 10 times more likely that a person is not affected by the disease than it is for having the disease. In this case, we say that the odds for not having the disease are 10 to 1.

4.5.2 Union

For two events E_1 and E_2 in a sample space S, we define their **union** $E_1 \cup E_2$ as the set of all outcomes that are at least in one of the events. The union $E_1 \cup E_2$ is an event by itself, and it occurs when *either E_1 or E_2* (or both) occurs. Schematically, the union $E_1 \cup E_2$ includes outcomes that are inside the regions defined for either E_1 or E_2 (or both). This description can be generalized to the union of more than two events.

The union of M and N in the above example is

$$M \cup N = \{1, 2, 3, 5\}.$$

The union of the two events in this case includes outcomes that are either less than 4 or odd or both. In Fig. 4.3, these are outcomes that are either inside the triangle or oval or both. The union of the heterozygous event, HT, and the disease event, D, is $\{Aa\} \cup \{aa\} = \{Aa, aa\}$.

Since the union of two events is an event by itself, we can talk about its probability. When possible, we can identify the outcomes in the union of the two events and find the probability by adding the probabilities of those outcomes. For the die rolling example,

$$P(M \cup N) = P(\{1, 2, 3, 5\}) = \frac{4}{6} = \frac{2}{3}.$$

Note that in general this is not equal to the sum of the probabilities of the two events: $P(M \cup N) \neq \frac{1}{2} + \frac{1}{2}$. Only under a specific condition (discussed below), we can write

the probability of the union of two events as the sum of their probabilities. For the union of the heterozygous event, HT, and the disease event, D,

$$P(HT \cup D) = P(\{Aa, aa\}) = 0.42 + 0.09 = 0.51.$$

In this special case, the probability of the union of the two events is equal to the sum of their individual probabilities.

4.5.3 Intersection

For two events E_1 and E_2 in a sample space S, we define their **intersection** $E_1 \cap E_2$ as the set of outcomes that are in both events. The intersection $E_1 \cap E_2$ is an event by itself, and it occurs when both E_1 *and* E_2 occur. Schematically, the intersection $E_1 \cap E_2$ includes outcomes that are inside the regions defined for both E_1 and E_2. This description can be generalized to the intersection of more than two events.

The intersection of M and N in the above example is

$$M \cap N = \{1, 3\}.$$

In this case, the intersection of the two events includes outcomes that are less than 4 *and* odd. In Fig. 4.3, these are outcomes that are in both the triangle and oval. The intersection of the heterozygous event and the no-disease event is $HM \cap ND = \{AA\}$.

The intersection of two events is an event by itself, so we can talk about its probability:

$$P(M \cap N) = P(\{1, 3\}) = \frac{2}{6} = \frac{1}{3},$$
$$P(HM \cap ND) = P(AA) = 0.49.$$

Now consider the intersection of the heterozygous event and the disease event. There is no common element between HT and D. Therefore, the intersection is the empty set $HT \cap D = \{\}$, and its probability $P(HT \cap D) = P(\emptyset) = 0$.

4.5.4 Joint vs. Marginal Probability

We refer to the probability of the intersection of two events, $P(E_1 \cap E_2)$, as their **joint probability**. Occasionally, we show this probability as $P(E_1 E_2)$. In contrast, we refer to probabilities $P(E_1)$ and $P(E_2)$ as the **marginal probabilities** of events E_1 and E_2.

For any two events E_1 and E_2, we have

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2). \tag{4.2}$$

That is, the probability of the union $P(E_1 \cup E_2)$ is the sum of their marginal probabilities minus their joint probability.

For the above die rolling example, we have

$$P(M \cup N) = \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = \frac{2}{3},$$

which is what we found before. Intuitively, adding their individual probabilities $P(M) = P(\{1, 2, 3\})$ and $P(N) = P(\{1, 3, 5\})$ counts their common elements $\{1, 3\}$ twice. We correct this by subtracting $P(\{1, 3\})$, which is $P(M \cap N)$.

The union of the heterozygous event and the no-disease event is

$$P(HM \cup ND) = P(HM) + P(ND) - P(HM \cap ND)$$
$$= 0.58 + 0.91 - 0.49 = 1.$$

This was of course expected since the union of HM and ND is the entire sample space: $HM \cup ND = \{AA, Aa, aa\} = S$.

4.6 Disjoint Events

Two events are called **disjoint** or **mutually exclusive** if they never occur together: if we know that one of them has occurred, we can conclude that the other event has not. Disjoint events have no elements (outcomes) in common, and their intersection is the empty set. For the above die rolling example, M and N are not disjoint. The outcomes 1 and 3 are shared by the two events. Therefore, both events happen simultaneously when either 1 or 3 occurs. On the other hand, $M = \{1, 2, 3\}$ and $Q = \{5, 6\}$ are two disjoint events. When rolling a die, the outcome cannot be less than 4 and greater than 4 at the same time. In the gene-disease example, the heterozygous event and the disease event are disjoint; they cannot occur at the same time. If a person is heterozygous, we know that he does not have the disease. If a person has the disease, we know that he cannot be heterozygous. The intersection of these two events is the empty set, $HT \cap D = \{\}$; hence, $P(HT \cap D) = P(\emptyset) = 0$.

For two disjoint events E_1 and E_2, the probability of their intersection (i.e., their joint probability) is zero:

$$P(E_1 \cap E_2) = P(\emptyset) = 0.$$

Therefore, according to Eq. 4.2, the probability of the union of two disjoint events is simply the sum of the their marginal probabilities:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2). \tag{4.3}$$

In general, if we have multiple disjoint events, $E_1, E_2, \ldots, E_n$, then the probability of their union is the sum of their marginal probabilities:

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n). \tag{4.4}$$

Accordingly, the probability of the union of the heterozygous and disease events is $P(HT \cup D) = 0.42 + 0.09 = 0.51$. Likewise, when we roll a die, the events $\{1, 2\}$, $\{4\}$, and $\{5, 6\}$ are disjoint. The occurrence of one event prevents the occurrence of the others. Therefore, the probability of their union is

$$P\big(\{1, 2\} \cup \{4\} \cup \{5, 6\}\big) = 1/3 + 1/6 + 1/3 = 5/6.$$

Now consider the three events $\{1, 2, 3\}$, $\{4\}$, and $\{5, 6\}$. These events are disjoint, and their union is the sample space S.

When two or more events are disjoint and their union is the sample space S, we say that the events form a **partition** of the sample space.

Two complementary events E and E^c always form a partition of the sample space since they are disjoint and their union is the sample space.

4.7 Conditional Probabilities

In this section, we discuss possible changes in the probability of one event based on our knowledge regarding the occurrence of another event.

The **conditional probability**, denoted $P(E_1|E_2)$, is the probability of event E_1 given that another event E_2 has occurred.

In contrast, the marginal probability $P(E_1)$ is the unconditional probability of E_1 regardless of the occurrences of other events. Consider the die rolling example. Recall that $P(M) = 1/2$. Now suppose that we are told that N has occurred, that is, the outcome is in fact an odd number. Then, the set of possible outcomes reduces to $S^* = N = \{1, 3, 5\}$. This new sample space is shaded in Fig. 4.5. Since the three
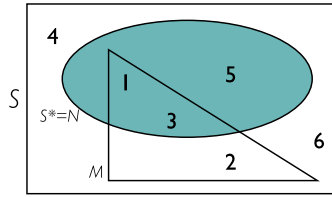
Fig. 4.5 A schematic representation of the die rolling example. Here, M (*triangle*) is the event that the outcome is less than 4, and N (*oval*) is the event that the outcome is an odd number. Assuming that M has occurred (the outcome is an odd number) reduces the number of possible outcomes (a new sample space) to N

possible outcomes, 1, 2, and 3, are still equally probable, the probability of each of them is now $1/3$. Within this smaller space, the event M occurs if the outcome is either 1 or 3. (The outcome of 2 is no longer a possibility.) These are two out of three equally probable outcomes. Therefore, the probability of M given that N has occurred (i.e., the conditional probability of M given N) is $P(M|N) = 2/3$. In this case, knowing that the outcome is an odd number increased the probability of E_1 from $1/2$ to $2/3$.

> The conditional probability of event E_1 given event E_2 can be calculated as follows: (assuming $P(E_2) \neq 0$)
>
> $$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}. \tag{4.5}$$
>
> This is the joint probability of the two events divided by the marginal probability of the event on which we are conditioning.

In the die rolling example, the intersection of the two events is $M \cap N = \{1, 3\}$ with probability $P(E_1 \cap E_2) = 2/6 = 1/3$. Therefore, the conditional probability of an outcome less than 4, given that the outcome is an odd number, is

$$P(M|N) = \frac{P(M \cap N)}{P(M)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

Now consider the gene-disease example. Suppose we know that a person is homozygous and are interested in the probability that this person has the disease, $P(D|HM)$. The probability of the intersection of D and HM is $P(D \cap HM) = P(\{aa\}) = 0.09$. Using Eq. 4.5, the conditional probability of having the disease knowing that the genotype is homozygous can be obtained as follows:

$$P(D|HM) = \frac{P(D \cap HM)}{P(HM)} = \frac{0.09}{0.58} = 0.16.$$

In this case, the probability of the disease has increased from $P(D) = 0.09$ (the unconditional probability) to $P(D|HM) = 0.16$ (the conditional probability).

Now let us find the conditional probability of not having the disease knowing that the person has a homozygous genotype: $P(ND|HM)$. The joint probability of ND and HM is $P(ND \cap HM) = P(\{AA\}) = 0.49$. The conditional probability is therefore

$$P(ND|HM) = \frac{P(ND \cap HM)}{P(HM)} = \frac{0.49}{0.58} = 0.84.$$

The information that the person is homozygous decreases the probability of no-disease from its 0.91 to 0.84.

Note that the two events ND and D are complementary, and the conditional probability of ND given HM is 1 minus the conditional probability of D given HM,

$$P(ND|HM) = 1 - P(D|HM) = 1 - 0.16 = 0.84.$$

In general, all the probability rules we discussed so far apply to conditional probabilities. Conditioning on an event only reduces the sample space (e.g., from the large rectangle to the shaded oval in Fig. 4.5). Within this shrunken sample space, all probability rules are valid. For example,

$$P\big(E_1^c\big|E_2\big) = 1 - P(E_1|E_2),$$
$$P(E_1 \cup E_2|E_3) = P(E_1|E_3) + P(E_2|E_3) - P(E_1 \cap E_2|E_3).$$

4.8 The Law of Total Probability

By rearranging Eq. 4.5 (i.e., moving $P(E_2)$ to the other side), we obtain the following useful equation:

$$P(E_1 \cap E_2) = P(E_1|E_2)P(E_2). \tag{4.6}$$

Therefore, the probability that both E_1 and E_2 occur, i.e., their joint probability, is the product of the conditional probability of E_1 given E_2 and the marginal probability of E_2. We will use this rule in the following sections.

Now suppose that a set of K events $B_1, B_2, \ldots, B_K$ forms a partition of the sample space. (See Fig. 4.6, where $K = 6$.) In other words, the events are disjoint (mutually exclusive), and their union is the entire sample space. For any event A (shown as a shaded area in Fig. 4.6) in the sample space, we can use Eq. 4.6 to write the probability of $A \cap B_k$ as follows:

$$P(A \cap B_k) = P(A|B_k)P(B_k),$$

where B_k is one of the partitioning events. In Fig. 4.6, the event $(A \cap B_k)$ is the intersection of the shaded area and B_k; this corresponds to the shaded area that
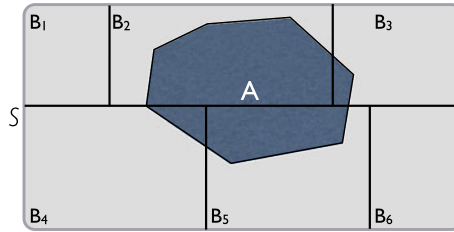
Fig. 4.6 A Venn diagram illustrating a set of partitioning events, $B_1, B_2, \ldots, B_6$. We can use the law of total probability to find the probability of any other event A, shown as a *shaded area*. Here, the probability of A given the conditional probabilities is $P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_6)P(A|B_6)$

falls inside B_k. The events $(A \cap B_1), \ldots, (A \cap B_K)$ themselves are disjoint. Using Eq. 4.4, we can write the union of these events as

$$P\big((A \cap B_1) \cup \cdots \cup (A \cap B_K)\big) = P(A \cap B_1) + \cdots + P(A \cap B_K)$$
$$= P(A|B_1)P(B_1) + \cdots + P(A|B_K)P(B_K).$$

From Fig. 4.6 it is clear that the union of $(A \cap B_1), \ldots, (A \cap B_K)$ is equal to the whole shaded area A. Therefore, the marginal probability of A can be calculated as follows:

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_K)P(B_K).$$

The above rule is known as the **law of total probability**, which can be written as

$$P(A) = \sum_{k=1}^{K} P(A|B_k)P(B_k), \tag{4.7}$$

where $B_1, B_2, \ldots, B_K$ form a partition of the sample space, and A is an event in the sample space.

For die rolling example, consider the three events $B_1 = \{1, 2\}$, $B_2 = \{3, 4\}$, and $B_3 = \{5, 6\}$, whose probabilities are $P(B_1) = P(B_2) = P(B_3) = 1/3$. These events form a partition of the sample space. The conditional probabilities of M (outcome less than four) given either of these three events are

$$P(M|B_1) = 1, \qquad P(M|B_2) = 1/2, \qquad P(M|B_3) = 0.$$

If we know that the event $B_1 = \{1, 2\}$ has occurred, we know for sure that the outcome is less than 4. Given $B_2 = \{3, 4\}$, the possible outcomes are now 3 and 4. One

of two possible outcomes corresponds to the event M, that is, the conditional probability of M given B_2 is 1/2. If we know that the event $B_3 = \{5, 6\}$ has occurred, then the probability that the number is less than 4 is zero: $P(M|B_3) = 0$. Using the law of total probability, we have

$$P(M) = P(M|B_1)P(B_1) + P(M|B_2)P(B_2) + P(M|B_3)P(B_3)$$
$$= 1 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} = \frac{1}{2},$$

which is the same as the probability we found directly based on the outcomes included in M.

4.9 Independent Events

Two events E_1 and E_2 are **independent** if our knowledge of the occurrence of one event does not change the probability of occurrence of the other event. That is, if E_1 and E_2 are independent, then the conditional probability of E_1 given E_2 (i.e., probability of E_1 knowing E_2 has occurred) is the same as the unconditional (or marginal) probability of E_1 (i.e., probability of E_1 regardless of E_2). Therefore, for two independent events,

$$P(E_1|E_2) = P(E_1).$$

Likewise,

$$P(E_2|E_1) = P(E_2).$$

For example, suppose that we toss two dice simultaneously. Knowing that the outcome of one of them is less than 4 does not change the probability that the outcome of the other one is an odd number. In this case, we say that the two events, "less than 4" for one die and "odd number" for the other one, are independent.

For our running example, where we are rolling *one* die only, the two events M and N are not independent. In this case, as we showed above, knowing that the outcome is an odd number, i.e., event N has occurred, increases the probability of M from 1/2 to 2/3. For the gene-disease example, we also showed that our knowledge that the genotype is homozygous increased the probability of the disease from $P(D) = 0.09$ (the unconditional probability) to $P(D|HM) = 0.16$ (the conditional probability). Therefore, the two events are dependent. This is of course consistent with our assumption that the disease is caused by gene **A**.

Equation 4.6 provides a general rule for the probability of the intersection of two events. However, if E_1 and E_2 are independent, then $P(E_1|E_2) = P(E_1)$. Substituting $P(E_1)$ for $P(E_1|E_2)$ into Eq. 4.6, we obtain the following rule for the probability that two independent events occur at the same time.

When two events E_1 and E_2 are independent, the probability that E_1 and E_2 occur simultaneously, i.e., their joint probability, is the product of their marginal probabilities:

$$P(E_1 \cap E_2) = P(E_1) \times P(E_2). \qquad (4.8)$$

In general, if events $E_1, E_2, \ldots, E_n$ are independent, then

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1) \times P(E_2) \times \cdots \times P(E_n).$$

For example, if we toss two fair coins simultaneously, then the probability of observing heads on both coins is $P(H_1 \cap H_2) = 1/2 \times 1/2 = 1/4$.

Using the above rule along with Eq. 4.2, we obtain the probability of the union of two independent events as follows:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1) \times P(E_2).$$

For the above coin tossing example, the probability that at least one of the two coins is heads is

$$P(H_1 \cup H_2) = 1/2 + 1/2 - 1/4 = 3/4.$$

4.10 Bayes' Theorem

In some situations, we know the conditional probability of E_1 given E_2, but we are interested in the conditional probability of E_2 given E_1. For example, suppose that the probability of having lung cancer is $P(C) = 0.001$ and that the probability of being a smoker is $P(SM) = 0.25$. Further, suppose we know that if a person has lung cancer, the probability of being a smoker increases to $P(SM|C) = 0.40$. We are, however, interested in the probability of developing lung cancer if a person is a smoker, $P(C|SM)$. Using Eq. 4.5, this conditional probability is

$$P(C|SM) = \frac{P(SM \cap C)}{P(SM)}.$$

From Eq. 4.6 the probability of being a smoker and having lung cancer at the same time is

$$P(SM \cap C) = P(SM|C)P(C).$$

Since $P(C)$ and $P(SM|C)$ are known, we can calculate the conditional probability of developing lung cancer for smokers:

$$P(C|SM) = \frac{P(SM|C)P(C)}{P(SM)} = \frac{0.4 \times 0.001}{0.25} = 0.0016.$$

Therefore, the probability of lung cancer for smokers increases from 0.001 to 0.0016. That is, the probability becomes 60% higher than the overall probability of lung cancer.

In general, for two events E_1 and E_2, the following equation shows the relationship between $P(E_2|E_1)$ and $P(E_1|E_2)$:

$$P(E_2|E_1) = \frac{P(E_1|E_2)P(E_2)}{P(E_1)}. \tag{4.9}$$

This formula is known as **Bayes' theorem** or **Bayes' rule**.

Now suppose that a set of K events $B_1, B_2, \ldots, B_K$ forms a partition of the sample space. We can write the Bayes' theorem for each of the partitioning events as follows:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)}.$$

Here, B_i is one of the partitioning events, and A is an event in the sample space. Using the law of total probability (Eq. 4.7), we have

$$P(A) = \sum_{k=1}^{K} P(A|B_k)P(B_k).$$

Therefore, we can write the general form of Bayes' theorem as

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{k=1}^{K} P(A|B_k)P(B_k)}. \tag{4.10}$$

In what follows, we use the above general form of Bayes' theorem for analyzing the results of medical tests.

4.10.1 Application of Bayes' Theorem in Medical Diagnosis

As an example, we use the "sweat test" to diagnose Cystic Fibrosis (CF). It is well known that patients with CF have a high concentration of chloride in their sweat. The sweat test is a simple procedure to detect CF by measuring the concentration of salt in a person's sweat. A high level of salt above a certain cutoff indicates CF. The possible outcomes from this medical test are shown schematically in Fig. 4.7. The vertical line represents the boundary between the disease event D (i.e., people with CF) and its complementary event H (i.e., people without CF). The shaded area represents the positive test results T^+, while the unshaded area represents the negative test results T^-.
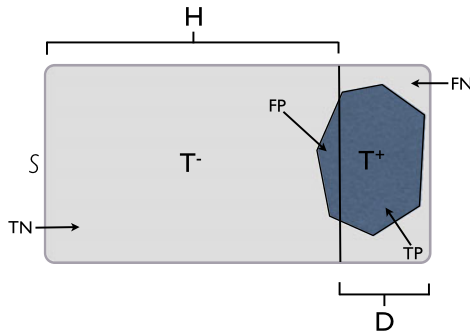
Fig. 4.7 A Venn diagram illustrating a typical medical diagnosis test. Here, the following abbreviations are used S for sample space, H for healthy, D for diseased, T^- for negative test result, T^+ for positive test result. The *shaded area to the right of vertical line* is the true positive TP, the *shaded area to the left of the vertical line* is the false negative FN, the *unshaded area to the left of the vertical line* is the true negative TN, and the *unshaded area to the right of the vertical line* is the false negative FN for the test

The sweat test successfully diagnoses many of the CF patients as positive. These cases are called **true positives**, TP, and represented by the shaded area (T^+) within the disease D region (right of the vertical line) in Fig. 4.7. The conditional probability of a positive diagnosis for CF patient, $P(T^+|D)$, is called the **sensitivity** of the test. The sweat test also successfully diagnoses most healthy people as negative for CF. These cases are called **true negatives**, TN, and represented by the unshaded area within H (left of the vertical line). The conditional probability of a negative result for a healthy person, $P(T^-|H)$, is called the **specificity** of the test.

Of course, there is always a chance of misdiagnosis. Cases where healthy people are diagnosed as positive are called **false positives**, FP. These cases are represented by the part of the shaded area that falls within the H region (left of the vertical line). The conditional probability of a positive result for a healthy person is $P(T^+|H)$. Likewise, some CF patients are misdiagnosed as negative. These cases are called **false negatives**, FN, and represented by the unshaded area (T^-) within the disease region D (right of the vertical line). The conditional probability of a negative result for a CF patient is $P(T^-|D)$.

The probability of the CF disease for a child whose parents are both carriers is $P(D) = 0.25$. Note that the gene causing CF is recessive. Therefore, if we denote the allele causing CF as a and the normal allele as A, only people with aa genotype have CF. People with Aa genotype are carriers. If both parents are carriers, the chance of transmitting a is 0.5 for each parent. Assuming that chromosomes from two parents are transmitted independently, there is the probability $P(D) = 0.5 \times 0.5 = 0.25$ that the child becomes affected (i.e., aa genotype). Then, the probability of being healthy is $P(H) = 1 - 0.25 = 0.75$.

Suppose that we use the sweat test and the child tests positive for the disease. Of course, this does not mean that he has CF for sure. Medical tests usually have nonzero false positive and false negative probabilities. Assume that the probability

of false positive for the sweat test is $P(T^+|H) = 0.04$ and the probability of false negative is $P(T^-|D) = 0.07$. As discussed in Sect. 4.7, we can use all probability rules for conditional probabilities. Because T^+ and T^- are complementary events, we have

$$P(T^-|H) = 1 - P(T^+|H) = 1 - 0.04 = 0.96,$$
$$P(T^+|D) = 1 - P(T^-|D) = 1 - 0.07 = 0.93.$$

Now we can calculate the updated probability of the disease knowing that the outcome of the test is positive. Notice that the two events D and H form a partition of the sample space (Fig. 4.7). Using the general form of Bayes' theorem, the conditional probability of the disease given a positive test result is

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|H)P(H)}$$
$$= \frac{0.93 \times 0.25}{0.93 \times 0.25 + 0.04 \times 0.75} = 0.89.$$

Therefore, the positive test result increases the probability of having the disease from $P(D) = 0.25$ to $P(D|T^+) = 0.89$.

4.10.2 Bayesian Statistics

In the CF diagnosis example discussed in this section, we assigned the probability of 0.25 to the disease event before seeing any new empirical data (i.e., test results). This probability is called the **prior probability**. In this case, the prior probability of disease was $P(D) = 0.25$. After obtaining new evidence, namely positive test results, we updated the probability of the disease from $P(D)$ to $P(D|T^+)$. We call this updated probability the **posterior probability**. In this case, the posterior probability of the disease was $P(D|T^+) = 0.89$. Therefore, based on the test result, we become more certain that the child is affected by the disease.

The above concept is the basis of **Bayesian Statistics**, which provides a framework to combine prior probability with new empirical data in order to perform statistical inference based on posterior probabilities.

4.11 Interpretation of Probability as the Relative Frequency

The random phenomena we have been discussing so far can be observed repeatedly. A coin can be tossed or a die can be rolled many times. Likewise, we can observe the genotypes of many people. These repeated experiments or observations are called **trials**. For such random phenomena, the probability of an event can be interpreted

Fig. 4.8 Simulation study of
the relative frequency of *AA*
genotype for different sample
size values. As *n* increases,
the sample relative frequency
n_{AA}/n approaches 1/4
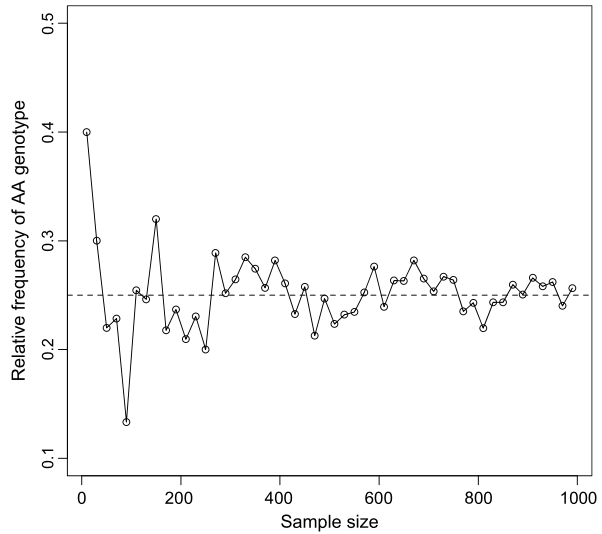
in terms of the relative frequency. The above view of probability is the basis of
Frequentist Statistics.

As an example, suppose that the probability of genotype *AA* is $P(AA) = 1/4$.
This probability could be interpreted as 1 out of 4 people in the population have
genotype *AA*. Suppose that we take a simple random sample of size *n* from the
population. If the genotype *AA* is observed n_{AA} times in the sample, the relative
frequency of *AA* in the sample is n_{AA}/n. If our probability assumption is true (i.e.,
$P(AA) = 1/4$), this sample relative frequency would be approximately 1/4. In this
case, as our sample size *n* increases, the sample relative frequency becomes closer
to the probability of 1/4; that is, it reaches the probability $P(AA) = 1/4$.

For illustration, we simulate (using computer programs) sampling people from
the population. The plot in Fig. 4.8 shows how the sample relative frequency of *AA*
genotype approaches the probability $P(AA) = 1/4$ as the sample size increases.

Note that the above interpretation of probability requires two important assump-
tions. First, we assume that the probability of events does not change from one trial
to another. For example, the probability of *AA* must remain 1/4. If the population
changes as we are sampling people (e.g., genotype *AA* becomes more prevalent),
then the sample relative frequency will not converge to 1/4. We also assume that
the outcome of one trial does not affect the outcome of another trial.

4.12 Advanced

In this section, we discuss the application of tree diagrams for obtaining joint prob-
abilities and making decisions under uncertainty.

Ch_1	Ch_2	S	Probability	Genotype Probability
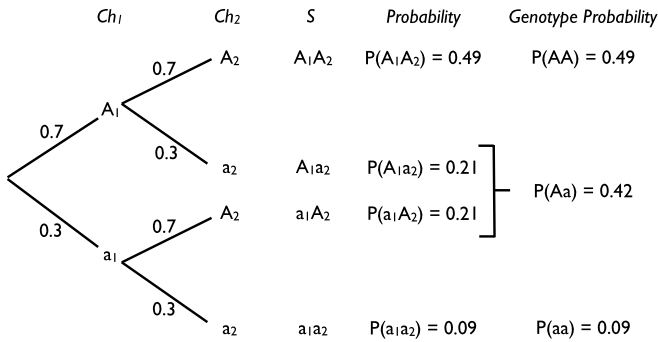
Fig. 4.9 Using a tree diagram to find the probability of genotypes assuming the alleles on homologous chromosomes are independent. The first set of branches represents possible alleles for one chromosome (Ch_1), and the second set represents possible alleles for the other chromosome (Ch_2)

4.12.1 Using Tree Diagrams to Obtain Joint Probabilities

Previously, we used tree diagrams to find the sample space for the combination of two random phenomena. Here, we extend the application of tree diagrams for calculating their joint probabilities. As an example, assume that the alleles on the homologous chromosomes are independent (i.e., the allele inherited from the mother has no influence on the allele inherited from the father). Also assume that for a bi-allelic gene **A**, the allele probabilities are $P(A) = 0.7$ and $P(a) = 0.3$. Then to find the genotype probabilities, we can use the tree diagram shown in Fig. 4.9. Note that this tree is similar to tree presented in Fig. 4.2, but this time we have put the probability of each possible outcome on its corresponding branch. The first set of branches represents possible alleles for one chromosome (Ch_1), and the second set represents possible alleles for the other chromosome (Ch_2). Since these events are independent, knowing the allele on the first chromosome has no influence on the probability of the allele on the second chromosome.

As before, the sample space is obtained by following a branch from root to tip: $S = \{A_1A_2, A_1a_2, a_1A_2, a_1a_2\}$. Since these events are independent, their joint probabilities are obtained by multiplying their marginal probabilities: $P(A_1A_2) = 0.7 \times 0.7 = 0.49$ (Eq. 4.8). Likewise, the probability of having a on the first chromosome and allele A on the second chromosome is $P(a_1A_2) = 0.3 \times 0.7 = 0.21$. Following similar approach, we can find the probability of each possible combination of two chromosomes. These probabilities are given in the column after the sample space in Fig. 4.9.

The labeling of the chromosomes is arbitrary. Therefore, we can drop the indices for A_1A_2 and a_1a_2 and write them as genotypes AA and aa, respectively. The genotype Aa can be considered as an event that includes two outcomes, A_1a_2 and a_1A_2. Therefore, $P(Aa) = 0.21 + 0.21 = 0.42$. This probability is shown in the last column in Fig. 4.9.

We can generalize the above example. Assume that the probability of observing the A allele is $P(A) = p$ and the probability of observing the a allele is $P(a) = q$.

Fig. 4.10 Tree diagram to
find probabilities when events
are dependent. Note that the
probabilities of the second set
of branches can change given
the outcomes on the first set
of branches

Genotype	Disease	S	Probability
	D	BB-D	0.064
BB			
0.64			
	H	BB-H	0.576
	D	Bb-D	0.032
0.32	Bb		
	H	Bb-H	0.288
	D	bb-D	0.008
0.04	bb		
	H	bb-H	0.032
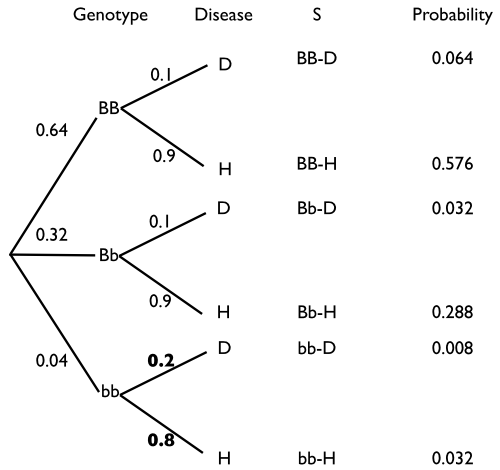
Then the genotype probabilities are

$$
\begin{aligned}
\text{Homozygous } AA: \quad & P(A_1 A_2) = p \times p = p^2, \\
\text{Heterozygous } Aa: \quad & P(A_1 a_2 \cup a_1 A_2) = p \times q + q \times p = 2pq, \\
\text{Homozygous } aa: \quad & P(a_1 a_2) = q \times q = q^2.
\end{aligned}
$$

Suppose, for example, that the allele probabilities for gene **B** are $P(B) = 0.8$ and
$P(b) = 0.2$ and that the alleles on homologous chromosomes are independent (i.e.,
they are transmitted from parents independently). Then the genotype probabilities
are $P(BB) = 0.8^2 = 0.64$, $P(bb) = 0.2^2 = 0.04$, and $P(Bb) = 2 \times 0.8 \times 0.2 = 0.32$.

This concept can be used to predict the genotype probabilities of children given
the allele probabilities of their parents in a population. However, this approach
requires that the population follows a very strict principle known as the *Hardy–
Weinberg* law (which assumes random mating, no selection, no mutation, no genetic
drift, no migration, and an infinite population size). A population adhering to this
law is said to be in Hardy–Weinberg equilibrium.

Now we discuss the use of tree diagrams to find probabilities when the outcomes
are not independent. Suppose that gene **B** in above example is related to a specific
disease, but it is not the only factor to determine the disease status. In particular, the
probability of having the disease is 0.2 for the *bb* genotype, whereas this probability
is 0.1 for the other two genotypes, *BB* and *Bb*. Therefore, the probability of the
disease depends on the genotype.

In Fig. 4.10, the first set of branches represents the genotype, and the second set
represents the disease status. The probabilities on the first set of branches are for
different genotypes: $P(BB) = 0.64$, $P(Bb) = 0.32$, and $P(bb) = 0.04$. The proba-
bilities on the second set of branches are conditional probabilities for the disease
status given the genotype: $P(D|BB) = 0.1$, $P(D|Bb) = 0.1$, and $P(D|bb) = 0.2$.
Since the healthy (H) and disease (D) events are complementary, the remaining
conditional probabilities are $P(H|BB) = 1 - 0.1 = 0.9$, $P(H|Bb) = 1 - 0.1 = 0.9$,

and $P(H|bb) = 1 - 0.2 = 0.8$. Therefore, unlike the tree for independent events (Fig. 4.9), the probabilities on the second set of branches depend on the outcomes on the first set of branches.

As before, we follow the branches from the root to tip and obtain the sample space:

$$S = \{BB - D, BB - H, Bb - D, Bb - H, bb - D, bb - H\}.$$

To find their probabilities, which are in fact the joint probabilities of genotype and disease status, we multiply the probabilities on the corresponding branches according to Eq. 4.6. For example, the probability of $Bb - D$ is the product of the conditional probability $P(D|Bb)$ and the marginal probability $P(Bb)$:

$$P(Bb - D) = P(Bb)P(D|Bb) = 0.32 \times 0.1 = 0.032.$$

4.12.2 Making Decisions under Uncertainty

Probability helps to quantify our uncertainty with respect to possible outcomes of a random phenomenon. However, probability alone is not enough to make decisions. For example, suppose that we have a choice between 1) winning \$10 with probability 0.9 and 2) winning \$1,000 with probability 0.8. While there is a higher probability of winning with the first option, we would be inclined to choose the second option after considering the potential (expected) gain for each possible outcome. As another example, suppose that medical tests show that a person might have cancer with probability 0.3. Although there is a probability of 0.7 for not having cancer, it is not reasonable to decide not to take any action just because the probability of the disease is lower compared to the alternative.

In the above examples, there is an explicit (in the first example) or implicit (in the second example) **utility function**, through which we attempt to quantify our gain or joy if a specific outcome occurs (win the money, recover from the disease). For the gambling example, the utility function assigns 10 to the winning event and 0 to the losing event (which has 0.1 probability) for the first option. Alternatively, we could use a **loss function** that assigns a value to the amount of loss due to any specific outcome (losing money, becoming more ill).

When making decisions, our goal is to maximize the **expected utility** (using a utility function) or minimize the **expected loss** (using a loss function). For simple cases, where the set of possible outcomes is finite (e.g., gambling), we can find the expected utility (or loss) for each option by multiplying the probability of each possible consequence (outcome) for that option by its corresponding utility (or loss) value and then summing over all possible consequences. For the first option in the gambling example, there are two possible consequences: winning with probability of 0.9 and utility of 10, or losing with probability of 0.1 and utility of 0. Our expected utility for this option is

$$EU_1 = 0.9 \times 10 + 0.1 \times 0 = \$9.$$

Decision	Outcome	Loss		Expected Loss
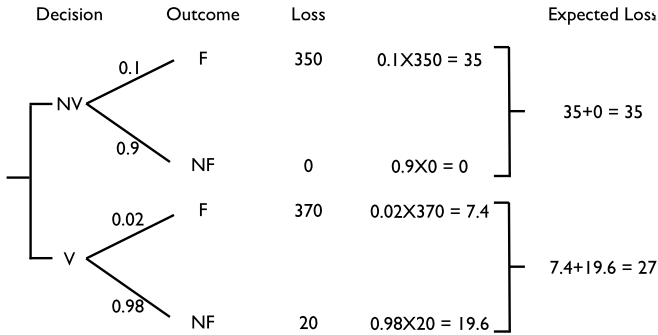
Fig. 4.11 Decision tree illustrating the expected loss for the flu example. The tree shows alternative actions, all possible outcomes with their corresponding probabilities and losses

For the second option, we either win probability of 0.8 and utility of 1000, or we lose with probability of 0.2 and utility of 0. Our expected utility in this case is

$$EU_2 = 0.8 \times 1000 + 0.2 \times 0 = \$800.$$

The expected utility is much higher for the second option. Therefore, we choose the second option.

As another example, consider whether to vaccinate against the seasonal flu. Suppose if we catch the flu, we spend $50 on medication and lose $300 in income from missing work. (There is also some amount of discomfort, which is not considered for simplicity.) Further suppose that the probability of catching the flu F without vaccination, NV, is 0.1, while the probability of the flu with vaccination V is 0.02. The cost of the vaccine is $20. Should we vaccinate?

To answer this question, we use the **decision tree** in Fig. 4.11. The first branches represent our decision: vaccination or no vaccination. The second set of branches represent the potential consequences of our decision. The probabilities of possible consequences are given on the branches. The "loss" column in Fig. 4.11 shows the loss due to the corresponding outcome when occurs. If we do not vaccinate and catch the flu, we lose $350 in the cost of medication and missed work. However, if we do not vaccinate and do not catch the flu, there would be no loss, $0. On the other hand, if we vaccinate and still catch the flu, our loss would be $20 + $350 = $370, since we have to pay for the vaccine and medication, and we miss work. Lastly, if we vaccinate and do not catch the flu, our loss is the $20 paid for the vaccination.

The last columns show our expected loss. If we decide not to vaccinate, our expected loss is $0.1 \times \$350 + 0.9 \times \$0 = \$35$. However, if we decide to vaccinate, our expected loss is $0.02 \times \$370 + 0.98 \times \$20 = \$7.4 + \$19.6 = \$27$. Therefore, we should vaccinate since its expected loss ($27) is less than the expected loss as the result of not vaccinating ($35).

4.13 Exercises

1. Consider two events E_1 and E_2, where $P(E_1) = 0.3$ and $P(E_2) = 0.5$. Calculate the following probabilities:
   (a) $P(E_1 \cup E_2)$ if the events are disjoint. In this case, are these two events partitioning the sample space?
   (b) $P(E_3)$, where $E_3 = (E_1 \cup E_2)^c$, and E_1 and E_2 are disjoint.
   (c) $P(E_1 \cap E_2)$ if the events are independent.
   (d) $P(E_1 \cup E_2)$ if the events are independent.
   (e) $P(E_2|E_1)$ if $P(E_1|E_2) = 0.35$. In this case, are these two events independent?

2. In a population that is in Hardy–Weinberg equilibrium, $P(a) = 0.1$ and $P(A) = 0.9$. Find the probability of each possible genotype.

3. Assume that the probability of having the disease is 0.4 and that the disease is not genetic (i.e., it is independent from the genotype of individuals). Also assume that the gene **A** has two alleles A and a such that $P(A) = 0.3$ and $P(a) = 0.7$. If the population is in Hardy–Weinberg equilibrium, write down the sample space for the combination of the disease status (D for diseased and H for healthy) and different genotypes along with the probability of each possible combination.

4. For the above question, find the probabilities for all possible combinations of genotypes and the disease status assuming that the disease is related to the gene **A** such that $P(D|aa) = 0.5$ and $P(D|Aa) = P(D|AA) = 0.3$.

5. Suppose that a pregnant woman is going to give birth to a girl or a boy with equal probabilities. However, if the baby is a boy, the probability that he has black (Bk) hair is 0.7, whereas this probability is 0.4 if the baby is a girl. Alternatively, the baby could have blond (Bd) hair. Using a tree diagram, find the sample space and the corresponding probabilities for all possible combinations of gender and hair color for the baby.

6. Suppose that the probability of being affected by H1N1 flu is 0.02. We found that among people who are affected by H1N1, the probability that a person washes her hands regularly is 0.3. If the probability of washing hands regularly in general (regardless of whether the person has the H1N1 flu or not) is 0.6. What is the probability of getting the H1N1 flu if a person washes her hands regularly?

7. A person has received the result of his medical test and realized that his diagnosis was positive (affected by the disease). However, the lab report stated that this kind of test has false positive probability of 0.06 (i.e., diagnosing a healthy person, H, as affected, D) and that the probability of false negative is 0.038 (i.e., diagnosing an affected person as healthy). Therefore, while this news was devastating, there is a chance that he was misdiagnosed. After some research, he found out that the probability of this disease in the population is $P(D) = 0.02$. Find the probability that he is actually affected by the disease given the positive lab result.

Chapter 5
Random Variables and Probability Distributions

5.1 Random Variables

In the previous chapter, we discussed random events and their probabilities. We used the possible genotypes of a bi-allelic gene **A** as an example. We defined its sample space, $S = \{AA, Aa, aa\}$, and various events, such as the homozygous event $HM = \{AA, aa\}$. We then discussed such concepts as the complement, union, intersection, conditional probability, and independence.

The focus of this chapter is random variables and their probability distributions. Formally, a **random variable** X assigns a numerical value to each possible outcome (and event) of a random phenomenon. For instance, we can define X based on possible genotypes of a bi-allelic gene **A** as follows:

$$X = \begin{cases} 0 & \text{for genotype } AA, \\ 1 & \text{for genotype } Aa, \\ 2 & \text{for genotype } aa. \end{cases}$$

In this case, the random variable assigns 0 to the outcome AA, 1 to the outcome Aa, and 2 to the outcome aa. The way we specify random variables based on a specific random phenomenon is not unique. For the above example, we can define another random variable Y as follows:

$$Y = \begin{cases} 0 & \text{for genotypes } AA \text{ and } aa, \\ 1 & \text{for genotype } Aa. \end{cases}$$

In this case, Y assigns 0 to the homozygous event and assigns 1 to the heterozygous event. When the underlying outcomes are numerical, the values the random variable assigns to each outcome can be the same as the outcome itself. For the die rolling example, we can define a random variable Z to be equal to $1, 2, \ldots, 6$ for outcomes $1, 2, \ldots, 6$, respectively. Alternatively, we can define a random variable W and set W to 1 when the outcome is an odd number and to 2 when the outcome is an even number.

The set of values that a random variable can assume is called its **range**. For the above examples, the range of X and Z is $\{0, 1, 2\}$, and the range of Z is $\{1, 2, \ldots, 6\}$.

After we define a random variable, we can find the probabilities for its possible values based on the probabilities for its underlying random phenomenon. This way, instead of talking about the probabilities for different outcomes and events, we can talk about the probability of different values for a random variable. Assume that the probabilities for different genotypes are $P(AA) = 0.49$, $P(Aa) = 0.42$, and $P(aa) = 0.09$. Then, instead of saying $P(AA) = 0.49$, i.e., the genotype is AA with probability 0.49, we can say that $P(X = 0) = 0.49$, i.e., X is equal to 0 with probability of 0.49. Likewise, $P(X = 1) = 0.42$ and $P(X = 2) = 0.09$. Note that the total probability for the random variable is still 1. For the die rolling example, instead of saying that the probability of observing an odd number is $1/2$ when rolling a die, we can say $P(W = 1) = 1/2$, i.e., W is equal to 1 with probability of $1/2$.

In what follows, we write $P(X)$ to denote the probability of a random variable X in general without specifying any value or range of values. Since the probability of a random variable is defined based on the probability of its underlying random phenomenon, the probability rules we discussed in the previous chapter also apply to random variables. Specifically, concepts such as independence and conditional probability are defined similarly for random variables as they are defined for random events. For example, when two random variables do not affect each other's probabilities, we say that they are independent.

As mentioned above, this chapter focuses on random variables and their probability distributions.

> The probability distribution of a random variable specifies its possible values (i.e., its range) and their corresponding probabilities.

For the random variable X defined based on genotypes, the probability distribution can be simply specified as follows:

$$P(X = x) = \begin{cases} 0.49 & \text{for } x = 0, \\ 0.42 & \text{for } x = 1, \\ 0.09 & \text{for } x = 2. \end{cases}$$

Here, x denotes a specific value (i.e., 0, 1, or 2) of the random variable.

Probability distributions are specified differently for different types of random variables. In the following section, we divide the random variables into two major groups: **discrete** and **continuous**. Then, we provide several examples for each group.

5.2 Discrete vs. Continuous

The grouping of random variables into discrete and continuous is based on their range. Discrete random variables can take a countable set of values. These variables

can be categorical (nominal or ordinal), such as genotype, gender, disease status, or pain level. They can also be counts, such as the number of patients visiting an emergency room per day, or the number of lymph nodes containing evidence of cancer. For all these examples, we can count the number of possible values the random variable can take. In the above genotype examples, X is a discrete random variable since it can take 3 possible values only.

Continuous random variables can take an uncountable number of possible values. Examples include weight, body temperature, BMI, and blood pressure. Consider the random variable Y for birthweight, which is a random phenomenon. In this case, the values of the random variables (i.e., numbers it assigns to each possible outcome) are the same as the corresponding outcomes; $Y = 7.9$ if the birthweight is 7.9 pounds. In this example, we cannot count the possible values of Y. For any two possible values of this random variable, we can always find another value between them. Consider 7.9 pounds and 8.0 pounds as two possible values for Y. We can find another number such as 7.95 between these two values. Now consider 7.9 and 7.95 as possible values; we can still find another number between them, such as 7.93. For continuous random variables, we can continue this process for any two possible values no matter how close they are. This is not the case for discrete random variables. While you can find another possible value between 70 heart beets per minutes and 75 heart beats per minute, you cannot do so for 70 and 71; there is no other possible value between them.

5.3 Probability Distributions

The probability distribution of a random variable provides the required information to find the probability of its possible values. Recall that the total probability of all possible values is equal to 1. Therefore, probability distribution specifies how the total probability of 1 is allocated among all possible values.

While the focus of earlier chapters was on exploring observed data and their distribution, here, we are concerned about all the possible values a random variable can take and their corresponding probabilities (i.e., the chance of observing those values) as opposed to a sample of observations. Although we do not necessarily need to think about a population when talking about probability distributions, it is easier to discuss this concept with a population in mind. Therefore, throughout this book, we usually discuss random variables and their probability distributions in the context of a population. That is, a random variable represents a specific random characteristic of a population and the possible values for that characteristic, even though we might not see many of those values in our sample. The probability distribution of a random variable specifies its range of possible values and how often we expect to see those values in the population. In that sense, our discussion of random variables and their distributions in this chapter remains at population level and theoretical (since we almost never have access to the whole population).

The probability distributions discussed here are characterized by one or two **parameters**. (In general, probability distributions can depend on more than two parameters.) These are values that define the form of a probability distribution. The

parameters of probability distributions we assume for random variables are usually unknown. Typically, we use Greek alphabets such as μ and σ to denote these parameters and distinguish them from known values.

While the probability distribution of a random variable is treated as the theoretical distribution of that random variable in the population, our study of probability distributions involves some analogous concepts as those discussed in earlier chapters regarding the distribution of observed data. More specifically, we are interested in specifying the mean and variance (or standard deviation) of a random variable as its measures of location and dispersion, respectively. Note, however, that the mean and variance here refer to the **population mean** and **population variance**, as opposed to sample mean and sample variance; hence, they remain theoretical, which means that we never know their true values. The mean of a random variable is also called its *expected* value even. (Note that the mean of a random variable is not its typical value in general.) We usually use μ to denote the mean of a random variable and use σ^2 to denote its variance; the standard deviation of a random variable is therefore σ. For a population of size N, the mean and variance are calculated as follows:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N},$$

where x_i is the value of the random variable for the ith member of the population.

In the remaining parts of this chapter, we discuss probability distributions for discrete and continuous random variables separately. We also discuss some commonly used discrete and continuous probability distributions. For each probability distribution, we provide the mean and variance and interpret them as the population mean and population variance for the corresponding random variable. Further, we use R-Commander and R to plot probability distributions.

5.4 Discrete Probability Distributions

For discrete random variables, the probability distribution is fully defined by the **probability mass function (pmf)**. This is a function that specifies the probability of each possible value within range of random variable. For the genotype example, the pmf of the random variable X is

$$P(X = x) = \begin{cases} 0.49 & \text{for } x = 0, \\ 0.42 & \text{for } x = 1, \\ 0.09 & \text{for } x = 2. \end{cases}$$

The probabilities for all possible values of the random variable sum to one.

As another example, suppose Y is a random variable that is equal to 1 when a newborn baby has low birthweight, and is equal to 0 otherwise. We say Y is a *binary*

random variable. Further, assume that the probability of having a low birthweight for babies is 0.3. Then the pmf for the random variable Y is

$$P(Y = y) = \begin{cases} 0.7 & \text{for } y = 0, \\ 0.3 & \text{for } y = 1. \end{cases}$$

Note that in this case, the random variable can take two values only. Since $P(Y = 1) = 0.3$ and the total probability of all possible values is 1, then we know that $P(Y = 0) = 0.7$.

In the following sections, we introduce some of the most commonly used probability mass functions for distributions of discrete random variables. For each distribution, we will provide the mean and variance as measures of location and spread, respectively. Moreover, with R-Commander we will visualize these distributions, obtain probabilities, and simulate data (i.e., artificially generate a set of observed values).

5.4.1 Bernoulli Distribution

Binary random variables are abundant in scientific studies. Examples include disease status (healthy and diseased), gender (male and female), survival status (dead, survived), and a gene with two possible alleles (A and a). We usually regard one of the values as the outcome of interest and denote it as $X = 1$. The other outcome is denoted as $X = 0$. As before, the probabilities for all possible values sum to one: $P(X = 0) + P(X = 1) = 1$.

> The binary random variable X with possible values 0 and 1 has a **Bernoulli** distribution with parameter θ, where $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$. We denote this as $X \sim \text{Bernoulli}(\theta)$, where $0 \le \theta \le 1$.

Here, θ is the unknown parameter. If θ were known, we could fully specify the probability mass function:

$$P(X = x) = \begin{cases} 1 - \theta & \text{for } x = 0, \\ \theta & \text{for } x = 1. \end{cases}$$

Sometimes we use the notation $P(X|\theta)$ to emphasize the dependence of the probabilities on the value of θ (i.e., given θ). Here, we simply show these probabilities as $P(X = 0)$ and $P(X = 1)$, where the dependence on θ is implied.

For example, let X be a random variable representing the five-year survival status of breast cancer patient, where $X = 1$ if the patient survived and $X = 0$ otherwise. Suppose that the probability of survival is $\theta = 0.8$: $P(X = 1) = 0.8$. Therefore, the

Fig. 5.1 Plot of the pmf for
Bernoulli(0.8) distribution

Bernoulli(0.8) Distribution
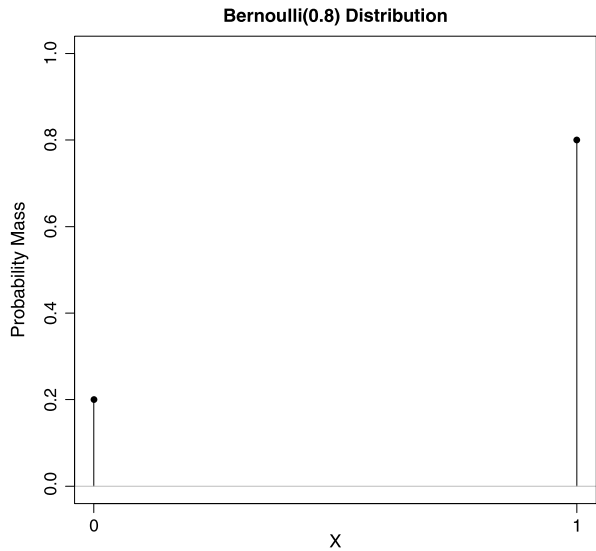
probability of not surviving is $P(X = 0) = 1 - \theta = 0.2$. Then X has a Bernoulli
distribution with parameter $\theta = 0.8$, and we denote this as

$$X \sim \text{Bernoulli}(0.8).$$

The pmf for this distribution is

$$P(X = x) = \begin{cases} 0.2 & \text{for } x = 0, \\ 0.8 & \text{for } x = 1. \end{cases}$$

Alternatively, we can plot pmf for visualizing the distribution. Figure 5.1 shows
the plot of pmf for the Bernoulli(0.8) distribution. The height of each bar is the
probability of the corresponding value on the horizontal axis. The height of the bar
is 0.2 at $X = 0$ and 0.8 at $X = 1$. Since the probabilities for all possible values of
the random variable add to 1, the bar heights also add up to 1.

The mean of a binary random variable, X, with Bernoulli(θ) distribution
is θ. We show this as $\mu = \theta$. In this case, the mean can be interpreted as
the proportion of the population who have the outcome of interest. Further-
more, the variance of a random variable with Bernoulli(θ) distribution is
$\sigma^2 = \theta(1 - \theta) = \mu(1 - \mu)$. The standard deviation is obtained by taking the
square root of variance: $\sigma = \sqrt{\theta(1 - \theta)} = \sqrt{\mu(1 - \mu)}$.

In the above example, $\mu = 0.8$. Therefore, we expect 80% of patients survive.
The variance of the random variable is $\sigma^2 = 0.8 \times 0.2 = 0.16$, and its standard de-
viation is $\sigma = 0.4$. This reflects the extent of variability in survival status from one

person to another. For this example, the amount of variation is rather small. Therefore, we expect to see many survivals ($X = 1$) with occasional death ($X = 0$). For comparison, suppose that the probability of survival for bladder cancer is $\theta = 0.6$. Then, the variance becomes $\sigma^2 = 0.6 \times (1 - 0.6) = 0.24$. This reflects a higher variability in the survival status for bladder cancer patients compared to that of breast cancer patients.

5.4.2 Binomial Distribution

A sequence of binary random variables $X_1, X_2, \ldots, X_n$ is called **Bernoulli trials** if they all have the same Bernoulli distribution (i.e., the same probability θ for the outcome of interest) and are independent (i.e., not affecting each other's probabilities). For example, suppose that we plan to recruit a group of 50 patients with breast cancer and study their survival within five years from diagnosis. We represent the survival status for these patient by a set of Bernoulli random variables $X_1, \ldots, X_{50}$. (For each patient, the outcome is either 0 or 1.) Assuming that all patients have the same survival probability, $\theta = 0.8$, and the survival status of one patient does not affect the probability of survival for another patient, $X_1, \ldots, X_{50}$ form a set of 50 Bernoulli trials.

Now we can create a new random variable Y representing the number of patients out of 50 who survive for five years. The number of survivals is the number of 1s in the set of Bernoulli trials. This is the same as the sum of Bernoulli trials, whose values are either 0 or 1:

$$Y = \sum_i^n X_i,$$

where $X_i = 1$ if the ith patient survive and $X_i = 0$ otherwise.

Since Y can be any integer number from 0 (no one survives) through 50 (everyone survives), its range is $\{0, 1, \ldots, 50\}$. The range is a countable set. Therefore, the random variable Y is discrete. The distribution of Y is a **binomial** distribution, shown as

$$Y \sim \text{Binomial}(50, 0.8).$$

The random variable representing the number of times the outcome of interest occurs in n Bernoulli trials (i.e., the sum of Bernoulli trials) has a Binomial(n, θ) distribution, where θ is the probability of the outcome of interest (a.k.a. the probability of success). A binomial distribution is defined by the number of Bernoulli trials n and the probability of the outcome of interest θ for the underlying Bernoulli trials.

The pmf of a binomial(n, θ) specifies the probability of each possible value (integers from 0 through n) of the random variable. For the breast cancer example,

Fig. 5.2 Plot of the pmf for
Binomial(50, 0.8) distribution
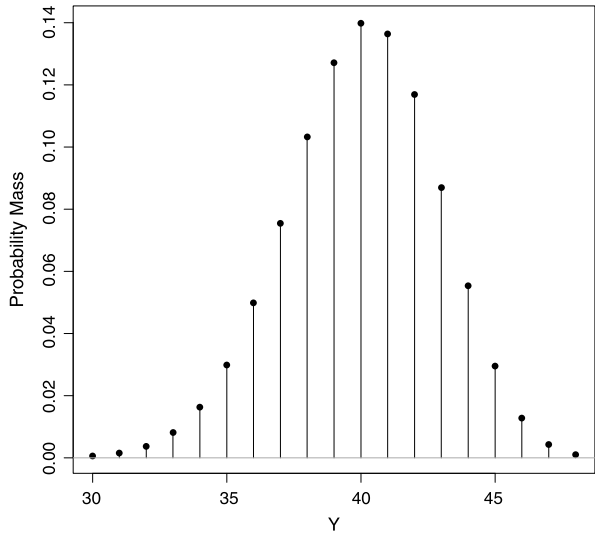
the pmf of Binomial(50, 0.8) distribution specifies the probability of 0 through 50 survivals.

The mathematical form of the pmf for binomial distributions are provided in Advanced section at the end of this chapter. Here, we use R-Commander to visualize this function and obtain the probability of each possible value. Click `Distri-butions → Discrete distributions → Binomial distribution → Binomial probabilities` and then set the number of `Binomial tri-als` (Bernoulli trials) to 50 and the `Probability of success` to 0.8. In the *Output* window, the result is shown as a table, where the first column shows the possible values for Y, and the second column shows their corresponding probabilities. For example, based on this table, the probability of 40 patients surviving is $P(Y = 40) = 0.14$.

We can also use R-Commander to plot the pmf for discrete distributions. Click `Distributions → Discrete distributions → Binomial distribution → Plot binomial distribution`. Specify the parameters of the distribution by entering 50 as the `Binomial trials` and 0.8 as the `Probability of success` (i.e., outcome of interest). Make sure the option `Plot probability mass function` is checked. The resulting graph illustrates the probabilities for different possible value of Y (Fig. 5.2). As before, the height of each bar is the probability of the corresponding value on the x-axis. For example, the probability of 35 survivals (out of 50) is 0.03, and the probability of 36 survivals is 0.05. Also, since the probabilities for all possible values of the random variable add to 1, the bar heights add up to 1. Note that even though Y can take integer values from 0 to 50, the plot does not show numbers below 30 and above 48 since the probability of these values is almost zero. For example, for the given survival probability, it is extremely unlikely to have only 10 survivals.

Now suppose that we are interested in the probability that either 34 or 35 or 36 patients survive. Since the underlying event include three possible outcomes, 34, 35, and 36, we obtain the probability by adding the individual probabilities for these outcomes:

$$P\big(Y \in \{34, 35, 36\}\big) = P(33 < Y \le 36)$$
$$= P(Y = 34) + P(Y = 35) + P(Y = 36)$$
$$= 0.02 + 0.03 + 0.05 = 0.1.$$

This is the same as adding the bar heights for $Y = 34$, $Y = 35$, and $Y = 36$. Further, suppose that we want to find the probability that the number of survivals (out of 50) is less than or equal to 36. That is, we are interested in the probability that either no one survives, or 1 patient survives, or 2 patients survive, . . . , or 36 patients survive. This is shown as $P(Y \le 36)$ and is called the **lower tail probability** of 36. As before, we can add up the individual probabilities to obtain

$$P(Y \le 36) = P(Y = 0) + P(Y = 1) + \cdots + P(Y = 36).$$

This is equivalent to adding the bar heights from $Y = 0$ through $Y = 36$. We can obtain the lower tail probability directly in R-Commander. Click `Distributions` $\rightarrow$ `Discrete distributions` $\rightarrow$ `Binomial distribution` $\rightarrow$ `Binomial tail probabilities`. Now enter 36 for the `Variable values`, 50 for the `Binomial trials`, 0.8 for the *Probability of success*, and make sure the option `Lower tail` is checked. The result, given in the *Output* window, is the probability that 36 or fewer patients survive: $P(Y \le 36) = 0.11$.

We can also obtain the **upper tail probability** of 36, which is the probability that more than 36 patients survive: $P(Y > 36)$. In R-Commander, repeat the above steps, but this time select the `Upper tail` option. The result is $P(Y > 36) = 0.89$. Since the lower tail and upper tail probabilities represent complementary events (i.e., either 36 people or fewer survival, or more than 36 people survive), we can obtain the upper tail probability by $P(Y > 36) = 1 - P(Y \le 36) = 1 - 0.11 = 0.89$.

In general, we use $P(Y \le y)$ to denote the lower tail probability for any specific value y. The upper tail probability can be obtained as $P(Y > y) = 1 - P(Y \le y)$.

In the above example, we add the individual probabilities for outcomes 34, 35, and 36 to obtain the interval probability $P(33 < Y \le 36)$. Note that by convention, the intervals include the upper limit (here 36) but not the lower limit (here 33). We can write this probability as follows:

$$P(33 < Y \le 36) = P(Y \le 36) - P(Y \le 33).$$

Here, we are subtracting the probability that Y is less than or equal to 33 from the probability that Y is less than or equal to 36. For this example, $P(Y \leq 36) = 0.11$ and $P(Y \leq 33) = 0.01$. Therefore,

$$P(33 < Y \leq 36) = 0.11 - 0.01 = 0.1.$$

In general, the probability of any interval from x_1 to x_2, where $x_1 < x_2$, can be obtained using the corresponding lower tail probabilities for these two points as follows:

$$P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1). \tag{5.1}$$

The above examples show that the pmf of a binomial distribution provides the required information to calculate all different probabilities. For a given pmf, we can also find the mean (expected value) and variance of the random variable.

The theoretical (population) mean of a random variable Y with Binomial(n, θ) distribution is $\mu = n\theta$. The theoretical (population) variance of Y is $\sigma^2 = n\theta(1 - \theta)$.

For the breast cancer example, the mean of the random variable is $50 \times 0.8 = 40$. (Note that in general the mean might not be an integer.) If we recruit 50 patients, we expect 40 people survive over five years. Of course, the actual number of survivals can change from one group to another (e.g., if we take another group of 50 patients). The variance of Y in the above example is $50 \times 0.8 \times 0.2 = 8$, which shows the extent of the variation of the random variable around its mean.

While in practice it is difficult to repeatedly recruit groups of 50 cancer patients, we can use computer programs such as R-Commander to **simulate** the sampling process. Click `Distributions` → `Discrete distributions` → `Binomial distribution` → `Sample from binomial distribution`. As in Fig. 5.3, name the simulated data set "BinomialSample1". Then specify the parameters n and θ by entering 50 for the `Binomial trials` and 0.8 for the `Probability of success`. Suppose that want to repeat the sampling procedure 10 times (i.e., 10 groups of 50). Enter 10 for the `Number of samples`. Lastly, set the `Number of observations` to 1, and uncheck `Sample means`. R-Commander then creates the data set `BinomialSample1`, which automatically becomes the active data set. This data set contains 10 randomly generated values for the random variable Y with Binomial$(50, 0.8)$ distribution. An example of the resulting data set is shown in Fig. 5.3. (Your simulated sample would be different.) In this simulated data set, 42 patients survive in the first group, 41 patients survive in the second group, and so forth. As expected, the number of survivals are generally close to the mean of the distribution, $\mu = 40$.
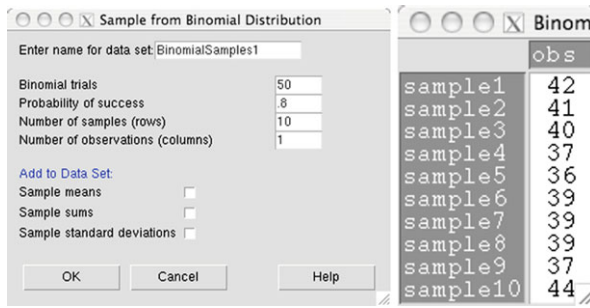
Fig. 5.3 *Left panel*: Simulating a random sample from the Binomial(50, 0.8) distribution in R-Commander. Name the data set and specify the number of trials n, the success probability θ, the number of samples, and the number of observations. *Right panel*: Viewing the simulated data set `BinomialSample1`, which was generated from the Binomial(50, 0.8) distribution

5.4.3 Poisson Distribution

So far, we have discussed the Bernoulli distribution for binary variables, and the binomial distribution for the number of times the outcome of interest (one of the two possible categories of the binary variable) occur within a set of n Bernoulli trials. While a random variable with a Binomial(n, θ) distribution is a count variable (e.g., number of people survived), its range is restricted to include integers from 0 through n only. For example, the number of survivals in a group of $n = 50$ cancer patients cannot exceed 50. Now, suppose that we are investigating the number of physician visits for each person in one year. Although very large numbers such as 100 are quite unlikely, there is no theoretical and prespecified upper limit to this random variable. Theoretically, its range is the set of all nonnegative integers. As another example, consider the number of trees per square mile in a certain region. Again, although spatial limits make very large numbers unlikely, there is no theoretical and prespecified limit for the possible values of the random variable.

Random variables representing counts within temporal and/or spacial limits but without prespecified upper limits are often assumed to have **Poisson** distributions. The range of these variables is the set of all nonnegative integers (i.e., the lower limit is zero, but there is no upper limit). A Poisson distribution is specified by a parameter λ, which is interpreted as the rate of occurrence within a time period or space limit. We show this as $X \sim \text{Poisson}(\lambda)$, where λ is a positive real number ($\lambda > 0$). The mean and variance of a random variable with Poisson(λ) distribution are the same and equal to λ. That is, $\mu = \lambda$ and $\sigma^2 = \lambda$.
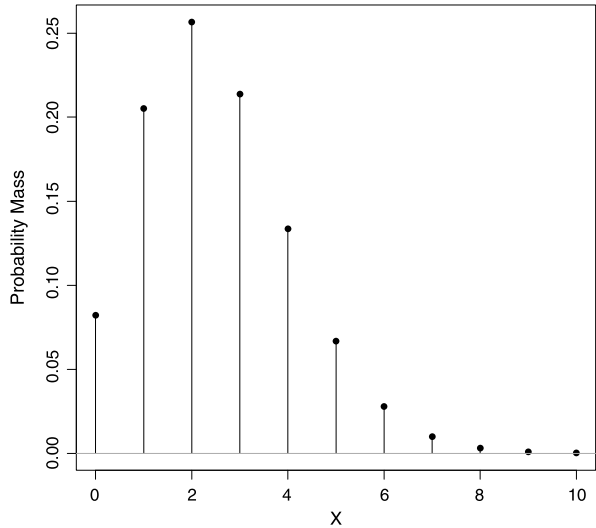
The pmf of a Poisson distribution specifies the probability of its possible values
(i.e., 0, 1, 2, . . .). The mathematical form of this function is provided in Advanced
section. Here, we use R-Commander to visualize the pmf and obtain the probability
of each possible value.

As an example, assume that the rate of physician visits per year is 2.5: $X \sim$
Poisson(2.5). The population mean and variance of this variable is therefore 2.5.
We can use R-Commander to plot the corresponding pmf for this distribution. Click
`Distributions` → `Discrete distributions` → `Poisson distri-`
`bution` → `Plot Poisson distribution` and then enter 2.5 for the `Mean`.
The resulting plot of the pmf shows the probability of each possible value, which is
any integer from 0 to infinity (Fig. 5.4). In this case, the probability of values above
8 becomes almost 0. According to this distribution, the probability of one visit per
year is $P(X = 1) = 0.21$. Also, the plot of pmf shows that it is very unlikely that a
person visits her physician 10 times per year.

To obtain the probability of specific values, click `Distributions` →
`Discrete distributions` → `Poisson distribution` → `Poisson`
`probabilities` and enter 2.5 as the `Mean`. The resulting probability table ap-
pears in the *Output* window. As before, the first column shows the possible values
of the random variable, and the second column shows their corresponding proba-
bilities. For this example, the probability that a person does not visit her physician
within a year is $P(X = 0) = 0.08$, while the probability of one visit per year in-
creases to $P(X = 1) = 0.21$.

Now suppose that we want to know the probability of up to three visits per year:
$P(X \leq 3)$. This is the probability that a person visit her physician 0, or 1, or 2,
or 3 times within one year. As before, we add the individual probabilities for the
corresponding outcomes:

$$P(X \leq 3) = 0.08 + 0.21 + 0.26 + 0.21 = 0.76.$$

This is the lower tail probability of $x = 3$. (Again, we use lower case for specific values of the random variable.) We can use R-Commander to obtain this probability. Click `Distributions` → `Discrete distributions` → `Poisson distribution` → `Poisson tail probabilities`. Enter 3 for the `Variable value`, 2.5 for the `Mean`, and make sure the option `Lower tail` is checked. The resulting probability, given in the *Output* window, matches the value calculated manually. The probability of more than three visits (i.e., the upper tail probability) is therefore $P(X > 3) = 1 - P(X \le 3) = 1 - 0.76 = 0.24$.

If the random variable representing the number of physician visits per year has in fact Poisson(2.5) distribution, its mean and variance are both 2.5. Use R-Commander to simulate (and plot) samples from this distribution. For example, we can sample the number of physician visits per year from Poisson(2.5) for $n = 1000$ people. Click `Distributions` → `Discrete distributions` → `Poisson distribution` → `Sample from Poisson distribution`. Then specify the parameter λ by setting the `Mean` to 2.5. Enter 1000 for the `Number of samples`, 1 for the `Number of observations`, and uncheck `Sample means`. View the newly created data set `PoissonSamples`. Now, find the sample mean and variance for this data set. They both should be close to 2.5, which is the theoretical mean and variance of the Poisson(2.5) distribution.

5.5 Continuous Probability Distributions

For discrete random variables, the pmf provides the probability of each possible value. For continuous random variables, the number of possible values is uncountable, and the probability of any specific value is zero. Intuitively, you can think about allocating the total probability of 1 among uncountable number of possible values. Therefore, instead of talking about the probability of any specific value x for continuous random variable X, we talk about the probability that the value of the random variable is within a specific interval from x_1 to x_2; we show this probability as $P(x_1 < X \le x_2)$. By convention, the interval includes the upper end of the interval but not the lower end.

For continuous random variables, we use **probability density functions** (pdf) to specify the distribution. Using the pdf, we can obtain the probability of any interval. As an example, consider the continuous random variable X representing the body mass index of the US population. Figure 5.5 shows the assumed probability density function for this variable. The mathematical form of this function is presented in the Advanced section. We refer to the corresponding curve shown in Fig. 5.5 as the **probability density curve**.

Note that the height of this curve at any specific value gives the *density* at that point. While we will use the density function to find probabilities for continuous random variables (discussed below), the value of the density function is not probability. Informally, however, the density curve shows the regions with high and low probability. In Fig. 5.5, for example, the region around 25 (e.g., between 20 to 30)

Fig. 5.5 The assumed
probability distribution for
BMI. The density curve
shown in this figure can be
used to find the probability
that the value of the random
variable falls within an
interval
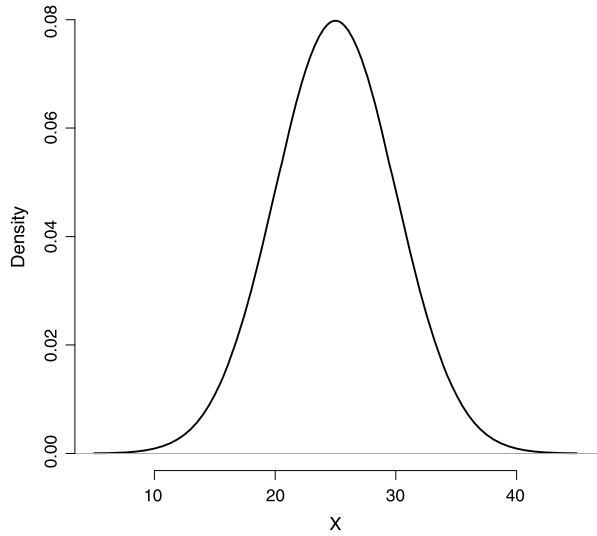
Fig. 5.6 The assumed
probability distribution for
BMI, which is denoted as X,
along with random sample of
100 values, which are shown
as *circles* along the horizontal
axis
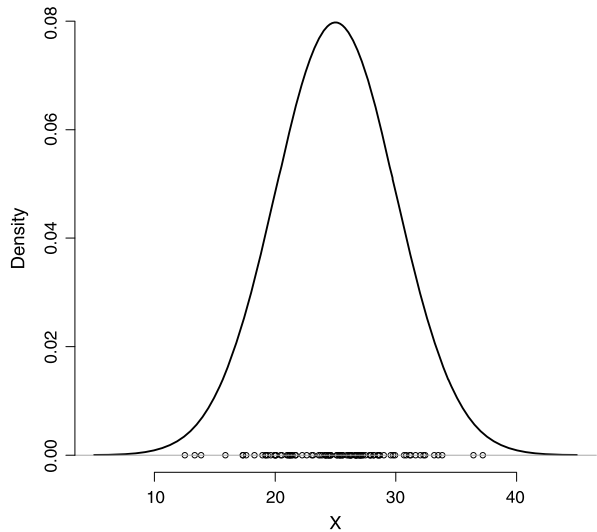
has relatively higher density compared to the region above 35 or below 15. If we ob-
serve a large number of values for this random variable, we expect many of them to
be around 25 (e.g., between 20 to 30). In Fig. 5.6, we show a random sample of 100
values for the random variable X. As we can see, many of these values fall within
the high-density region from 20 to 30. Only few observations fall in low-density
regions (e.g., the regions above 35 or below 15).

  As mentioned above, for continuous random variables, the probability of any
specific value is zero. For the above example, using the probability density curve
in Fig. 5.5, we can find the probability of that a person's BMI is between 25 and

Fig. 5.7 The *shaded area* is the probability that a person's BMI is between 25 and 30. People whose BMI is in this range are considered as overweight. Therefore, the *shaded area* gives the probability of being overweight
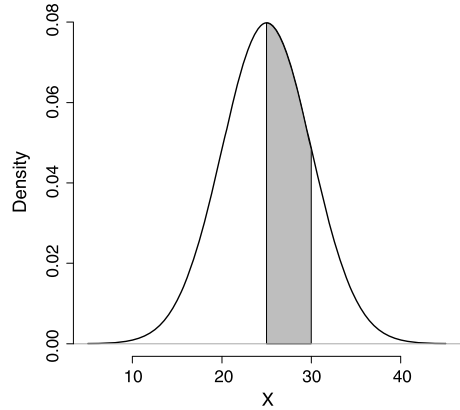
30: $P(25 < X \leq 30)$. (This is the probability of being overweight but not obese according to the Centers for Disease Control and Prevention.)

> The total area under the probability density curve is 1. The curve (and its corresponding function) gives the probability of the random variable falling within an interval. This probability is equal to the area under the probability density curve over the interval.

In Fig. 5.7, this probability is shown as the shaded area under the probability density curve between 25 and 30. Now suppose that we shrink the interval from $25 < X \leq 30$ to $28 < X \leq 30$. The shaded area under the curve would decrease, and the probability of the interval becomes smaller. If we continue shrinking the interval by moving the lower limit closer to 30, in limit the interval becomes a single number 30, and the shaded area, which is the probability of the interval, reduces to zero. Therefore, for this continuous random variable, the probability of 30 is zero: $P(X = 30) = 0$. In general, the probability of any specific value for continuous variables is zero: $P(X = x) = 0$. Note that the probability of zero for a specific value does not necessarily make it impossible. After all, the BMI of any person in the population is a specific value.

Similar to the discrete distributions, the probability of observing values less than or equal to a specific value x, is called the lower tail probability and is denoted as $P(X \leq x)$. This probability is found by measuring the area under the curve to the left of x. For example, the shaded area in the left panel of Fig. 5.8 is the lower tail probability of having a BMI less than or equal to 18.5 (i.e., being underweight), $P(X \leq 18.5)$. Likewise, the probability of observing values greater than x, $P(X > x)$, is called the upper tail probability and is found by measuring the area under the curve to the right of x. For example, the shaded area in the right panel of Fig. 5.8 is the upper tail probability of having a BMI greater than 30 (i.e., being obese).
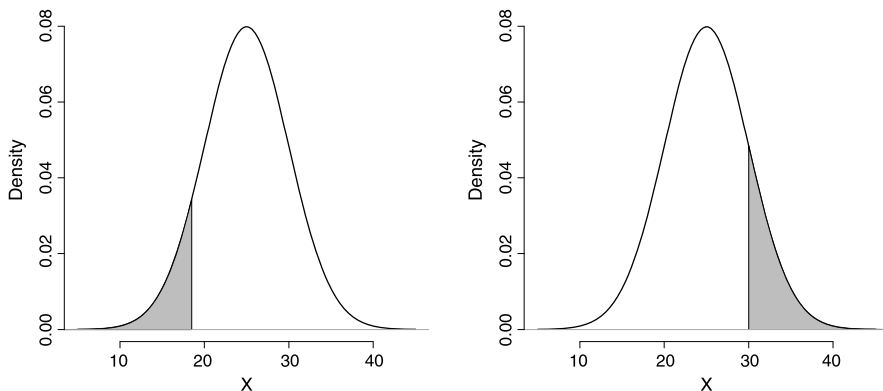
Fig. 5.8 *Left panel*: The lower tail probability of 18.8, $P(X \leq 18.8)$. *Right panel*: The upper tail probability 30, $P(X > 30)$

As before, the probability of any interval from x_1 to x_2, where $x_1 < x_2$, can be obtained using the corresponding lower tail probabilities for these two points as follows:

$$P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1). \tag{5.2}$$

For example, suppose that we wanted to know the probability of a BMI between 25 and 30. This probability $P(25 < X \leq 30)$ is obtained by subtracting the lower tail probability of 25 from the lower tail probability of 30:

$$P(25 < X \leq 30) = P(X \leq 30) - P(X \leq 25).$$

Again, we follow the general convention, where the intervals contain their upper limit (here, 30) but not their lower limit (here, 25).

In the following sections, we discuss some of the most common probability distributions for continuous random variables. These distributions depend on one or two unknown parameters and are specified by their probability density functions. As before, we will provide the mean μ and variance σ^2 of each distribution as measures of location and dispersion (spread), respectively. We also use R-Commander to plot the density curves and to obtain the probability of a given interval.

5.5.1 Probability Density Curves and Density Histograms

In the previous chapter, we discussed the interpretation of probability in terms of relative frequency. An analogous comparison can be made between density curves for probability distributions and density histograms for data.

Consider the density curve for the probability distribution of BMI shown Fig. 5.5. Now suppose that we observe the BMI values of 500 people selected from the population through simple random sampling. The left panel of Fig. 5.9 shows the density
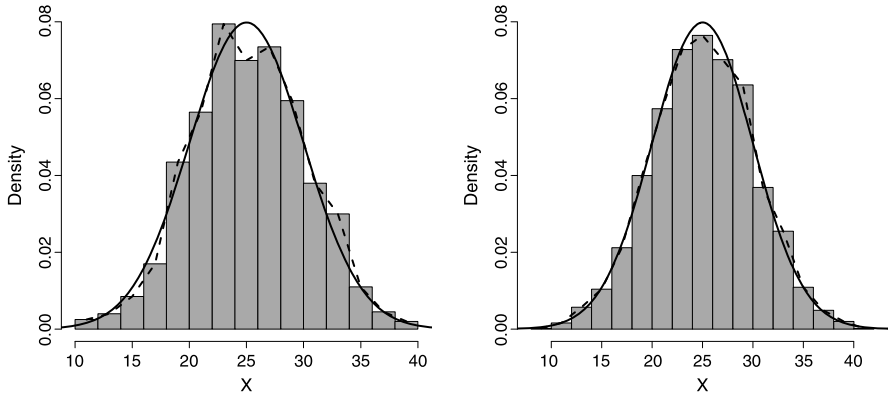
Fig. 5.9 *Left panel*: Histogram of BMI for 1000 observations. The *dashed line* connects the height of each bar at the midpoint of the corresponding interval. The *smooth solid curve* is the density curve for the probability distribution of BMI. *Right panel*: Histogram of BMI for 5000 observations. The histogram and its corresponding *dashed line* provide better approximations to the density curve

histogram of the observed data. Here, we have super-imposed a dashed line connecting the height of each bar at the midpoint of the corresponding interval. Recall that the height of each bar is the density for the corresponding interval, and the area of each bar is the relative frequency for that interval. We have also super-imposed the probability density curve for the random variable for comparison. As we can see, the density histogram and the dashed line, which shows the density for each interval based on the observed data, provide reasonable approximations to the density curve. Also, the area of each bar, which is equal to the relative frequency for the corresponding interval, is approximately equal to the area under the curve over that interval. The right panel of Fig. 5.9 shows the density histogram and the dashed line based on 5000 observations. The approximation to the density curve becomes much better as the sample size increases. If the assumed probability distribution for the random variable is in fact true, we expect that the dashed line reaches the density curve as the sample size n goes to infinity. (Note that as the sample size increases, we can increase the number of bins and decrease the bin width.)

5.5.2 Normal Distribution

Consider the probability distribution function and its corresponding probability density curve in Fig. 5.5 we assumed for BMI in the above example. As we move from left to right, the height of the density curve increases first until it reaches a point of maximum (peak), after which it decreases toward zero. We say that the probability distribution is **unimodal**. Because the height of the density curves reduces to zero symmetrically as we move away from the center, we say that the probability

Fig. 5.10 Examples of
density curves for the normal
distribution. The distribution
shown by the *solid curve* has
a mean of 1 and variance of 4.
The distribution shown by the
dashed curve has a mean of
−1 and variance of 1
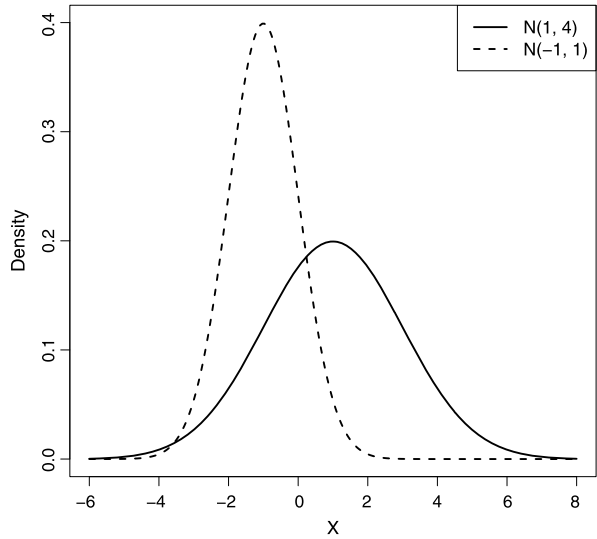
distribution is symmetric. We can use similar unimodal and symmetric probability distribution for many continuous random variables representing characteristics such as blood pressure, body temperature, atmospheric carbon dioxide concentration change, and so forth. This distribution is known as **normal** distribution, which is one of the most widely used distributions for continuous random variables. Random variables with this distribution (or very close to it) occur often in nature.

Figure 5.10 shows two examples of density curves for the normal distribution. Each curve is symmetric around its point of maximum. For normal distributions, the point where the density curve reaches its maximum is in fact the mean, denoted as μ. The mean is 1 for the distribution shown with the solid curve and −1 for the distribution shown with the dashed curve. The variance of a normally distributed random variable is denoted as σ^2 and determines the spread of the density curve; a higher variance means a more spread out curve. (The standard deviation is the square root of the variance and is denoted as σ.) The variance for the random variable with solid density curve is 4, whereas the variance of the random variable with dashed density curve is 1. Therefore, the former random variable is more dispersed than the latter one.

A **normal distribution** and its corresponding pdf are fully specified by the mean μ and variance σ^2. A random variable X with normal distribution is denoted $X \sim N(\mu, \sigma^2)$, where μ is a real number, but σ^2 can take positive values only. The normal density curve is always symmetric about its mean μ, and its spread is determined by the variance σ^2.

The range (set of possible values) for any normally distributed random variable is the set of real numbers from $-\infty$ to $+\infty$. However, it is not uncommon to assume

a normal distribution for a random variable that has only positive values. This is a reasonable assumption as long as the typical values of the random variable are far enough from zero, so that the probability of negative values (the area under the curve on the left hand side of zero) becomes negligible. For example, we can safely assume systolic blood pressure (SBP) is normally distributed since typical values are far from zero.

The 68–95–99.7% Rule   For normally distributed random variables, there is a simple rule, known as the **68–95–99.7% rule**, for finding the range of typical values.

The 68–95–99.7% rule for normal distributions specifies that

- 68% of values fall within 1 standard deviation of the mean:

$$P(\mu - \sigma < X \leq \mu + \sigma) = 0.68.$$

- 95% of values fall within 2 standard deviations of the mean:

$$P(\mu - 2\sigma < X \leq \mu + 2\sigma) = 0.95.$$

- 99.7% of values fall within 3 standard deviations of the mean:

$$P(\mu - 3\sigma < X \leq \mu + 3\sigma) = 0.997.$$

For example, suppose we know that the population mean and standard deviation for SBP are $\mu = 125$ and $\sigma = 15$, respectively. That is, $X \sim N(125, 15^2)$, where X is the random variable representing SBP. Therefore, the probability of observing an SBP in the range $\mu \pm \sigma$ is 0.68:

$$P(125 - 15 < X \leq 125 + 15) = P(110 < X \leq 140) = 0.68.$$

This probability corresponds to the central area shown in the left panel of Fig. 5.11. Likewise, the probability of observing an SBP in the range $\mu \pm 2\sigma$ is 0.95:

$$P(125 - 2 \times 15 < X \leq 125 + 2 \times 15) = P(95 < X \leq 145) = 0.95.$$

This probability is shown in the right panel of Fig. 5.11. Lastly, the probability of observing an SBP is in the range $\mu \pm 3\sigma$ is 0.997:

$$P(125 - 3 \times 15 < X \leq 125 + 3 \times 15) = P(80 < X \leq 170) = 0.997.$$

Therefore, we rarely (probability of 0.003) expect to see SBP values less than 80 or greater than 170.

Now suppose that we want to know the probability of being hypotensive, which is defined as a systolic blood pressure less than or equal to 90. Then we are interested
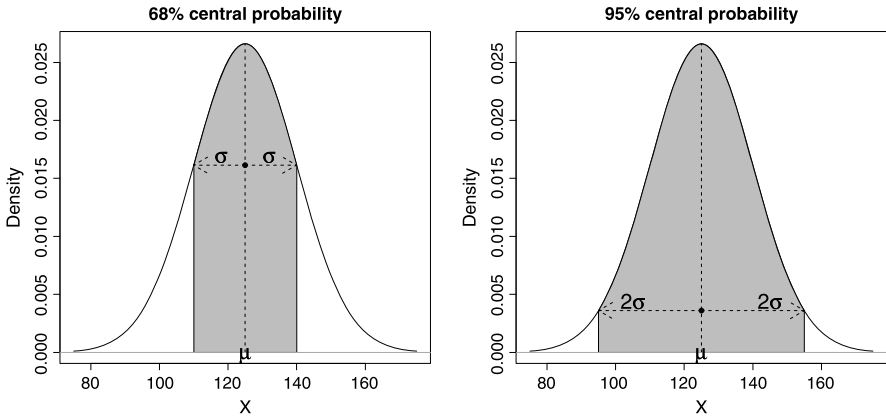
Fig. 5.11 Illustrating the *68–95–99.7% Rule* for systolic blood pressure, which is assumed to have a normal distribution: $X \sim N(125, 15^2)$. According to the rule, we would expect 68% of the observations to fall within 1 standard deviation of the mean (*left panel*) and 95% of the observations to fall within 2 standard deviations of the mean (*right panel*). Likewise, nearly 99.7% of observations to fall with in 3 standard deviations of the mean (not shown here)

in the lower tail probability $P(X \leq 90)$, which is equal to the area under the density curve to the left of $x = 90$. To obtain this probability in R-Commander, click `Dis-` `tributions` $\rightarrow$ `Continuous distributions` $\rightarrow$ `Normal distribu-` `tion` $\rightarrow$ `Normal probabilities`. Then enter 90 for `Variable value` and specify the parameter values as before. The result is given in the *Output* window. In this case, if SBP is normally distributed with $\mu = 125$ and $\sigma = 15$, the probability of being hypotensive is $P(X \leq 90) = 0.01$.

On the other hand, we can examine the probability of being hypertensive, which is defined as a systolic blood pressure over 140. In R-Commander, follow the same steps but now enter 140 for the `Variable value` and check the `Upper tail` option. The resulting upper tail probability is $P(X > 140) = 0.16$.

Using Eq. 5.1, we can find the probability of any given interval. For example, suppose that we consider a blood pressure from 110 to 130 as normal. The probability of having a normal blood pressure is

$$P(110 < X \leq 130) = P(X \leq 130) - P(X \leq 110) = 0.63 - 0.16 = 0.47.$$

R-Commander can be used to plot probability density curve for a normal distribution. Click `Distributions` $\rightarrow$ `Continuous distributions` $\rightarrow$ `Nor-` `mal distribution` $\rightarrow$ `Plot normal distribution`. Set the mean (μ) to 125 and the standard deviation (σ) to 15. Make sure the option `Plot density` `function` is checked. This creates a unimodal and symmetric probability density curve similar to those shown in Fig. 5.11. If the $N(25, 15^2)$ is in fact a good probability model for the distribution of BMI, then a large sample from the population would have a density histogram similar to this probability density curve, with the sample mean and sample standard deviation close to 125 and 15, respectively.

Fig. 5.12 Simulating 1000 observations from $N(125, 15^2)$ distribution for SBP and viewing the resulting `NormalSamples` data set
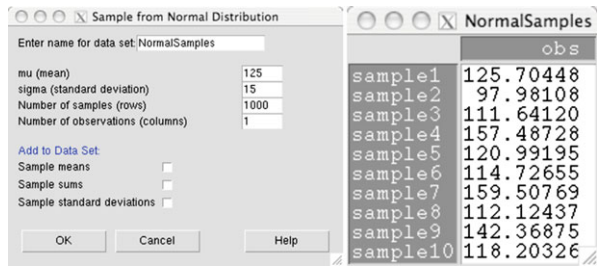
Fig. 5.13 Plots of the pdf for the standard normal distribution, $N(0, 1)$
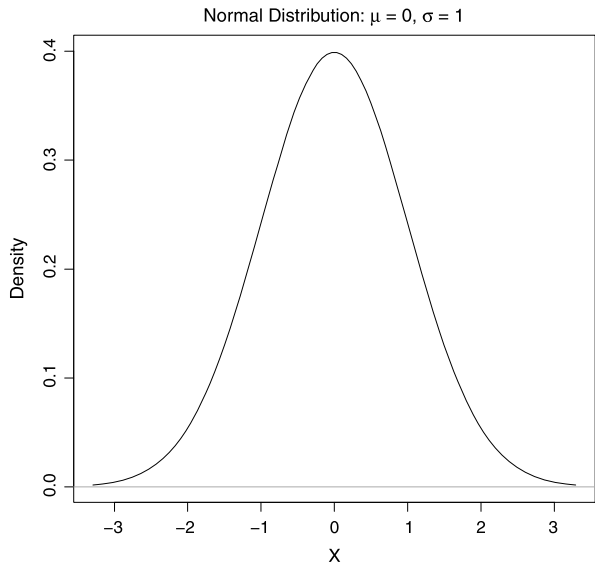
In R-Commander, let us try simulating 1000 values from $X \sim N(125, 15^2)$, which we used for the distribution of SBP in the population. Click `Distributions → Continuous distributions → Normal distribution → Sample from normal distribution`. Then specify the parameters by entering 125 for the `mu (mean)` and 15 for `sigma (standard deviation)`, as in Fig. 5.12. Set the `Number of samples` to 1000 and the `Number of observations` to 1. Lastly, uncheck the `Sample means` option. The first 10 observations (your sample will be different) of the resulting data set `NormalSamples` is shown in Fig. 5.12.

Plot the density histogram for the simulated data set and compare it to the probability density curve you previously created. They should be similar. Find the sample mean and sample standard deviation for this simulated data. They should be close to 125 and 15, respectively.

The Standard Normal Distribution

A normal distribution with a mean of 0 and a standard deviation (or variance) of 1 is called the **standard normal distribution** and denoted $N(0, 1)$.

Use R-Commander to plot density curve (i.e., the probability density function) of the standard normal curve $N(0, 1)$. The resulting plot is shown in Fig. 5.13. As expected, the density curve is symmetric around the mean of 0. Using the 68–95–99.7% rule, we expect 68% of the values to be between -1 and 1, 95% of values to be between -2 and 2, and nearly all (99.7%) of the values to be between -3 and 3.

5.5.3 Student's t-distribution

Another continuous probability distribution that is used very often in statistics is the **Student's t-distribution** or simply the t-**distribution**. As we will see in later chapters, the t-distribution especially plays an important role in testing hypotheses regarding the population mean. For example, testing the hypothesis that whether the average body temperature of healthy people is the widely accepted value of 98.6°F involves the t-distribution.

A t-distribution is specified by only one parameter called the **degrees of freedom** df. The t-distribution with df degrees of freedom is usually denoted as $t(df)$ or t_{df}, where df is a positive real number ($df > 0$). The mean of this distribution is $\mu = 0$, and the variance is determined by the degrees of freedom parameter, $\sigma^2 = df/(df - 2)$, which is of course defined when $df > 2$.

Similar to the standard normal distribution, the probability density curve for a t-distribution is unimodal and symmetric around its mean of $\mu = 0$. However, the variance of a t-distribution is greater than the variance of the standard normal distribution: $df/(df - 2) > 1$. As a result, the probability density curve for a t-distribution approaches zero more slowly than that of the standard normal. We say that t-distributions have *heavier tails* than the standard normal.

Figure 5.14 compares the pdf of a standard normal distribution to the t-distribution with 1 degree of freedom and then the t-distribution with 4 degrees of freedom. Both t-distributions have heavier tails than the standard normal distribution. However, as the degrees of freedom increase, the t-distribution approaches the standard normal. Try using R-Commander to plot the probability density function for various degrees as freedom. The steps to plot the pdf and obtaining probability of intervals based on a t-distribution is very similar to those of a normal distribution, except that we choose `t-distribution` under `Continuous distributions`.

Fig. 5.14 Comparing the pdf of a standard normal distribution to t-distributions with 1 degree of freedom and then with 4 degrees of freedom. The t-distribution has heavier tails than the standard normal; however, as the degrees of freedom increase, the t-distribution approaches the standard normal
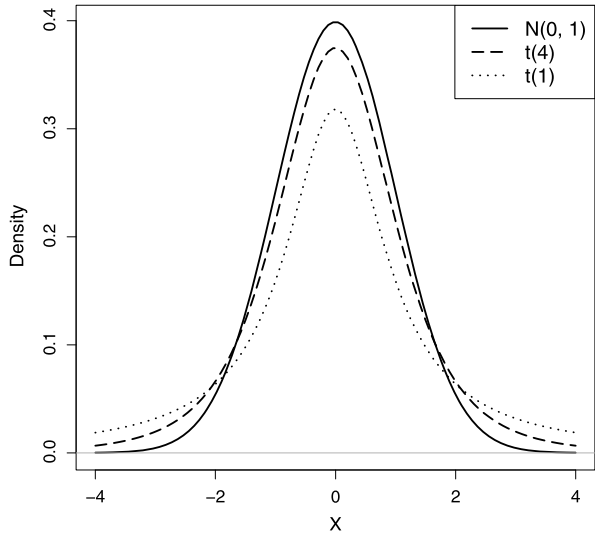
5.6 Cumulative Distribution Function and Quantiles

We saw that by using lower tail probabilities, we can find the probability of any given interval (see Eq. 5.1). This is true for all probability distributions (discrete or continuous). Indeed, all we need to find the probabilities of any interval is a function that returns the lower tail probability at any given value of the random variable. This function is called the **cumulative distribution function** (cdf) or simply the **distribution function**. For the value x of the random variable X, the cumulative distribution function returns $P(X \leq x)$.

Previously, we saw how we can use R-Commander to obtain lower tail probabilities. We can also use R-Commander to plot the cdf for all possible values of a random variable. Let us plot the cdf of the standard normal distribution, $N(0, 1)$. Click Distributions → Continuous distributions → Normal distribution → Plot normal distribution. The default parameters correspond to the standard normal. (By changing these parameters, we can plot the pdf of any normally distributed variable.) Check the option Plot distribution function. The resulting curve is shown in Fig. 5.15 and can be used to find the lower tail probability for all possible values of the random variable. For instance, in the left panel of Fig. 5.15, following the arrows starting from $x = 0$ gives us the lower tail probability of $P(X \leq 0) = 0.5$. Since the lower tail probability is a number between zero and one, the vertical axis of the cdf plot ranges from zero to one. Also, since the lower tail probability either remains the same or increases as we increase x, the cdf plot is always a nondecreasing function as x increases.

For the standard normal distribution in Fig. 5.15, we used the cdf plot to find the lower tail probability of zero: $P(X \leq 0) = 0.5$. For this, we followed the direction of arrows from the horizontal axis to the vertical axis. We can use the cdf plot in
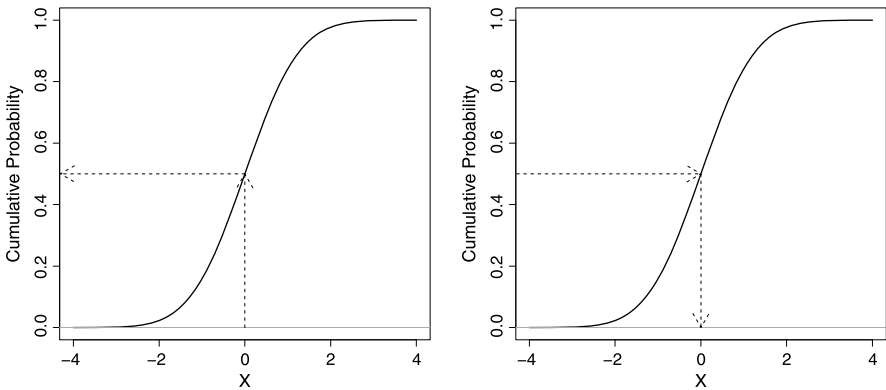
Fig. 5.15 *Left panel*: Plot of the cdf for the standard normal distribution, $N(0, 1)$. The cdf plot of the cdf can be used to find the lower tail probability. For instance, following the *arrow* from $x = 0$ (on the horizontal axis) to the cummulative probability (on the vertical axis) gives us the probability $P(X \leq 0) = 0.5$. *Right panel*: Given the lower tail probability of 0.5 on the vertical axis, we obtain the corresponding quantile $x = 0$ on the horizontal axis

the reverse direction to find the value of a random variable for a given lower tail probability. This is shown in right panel of Fig. 5.15. Here, we follow the direction of arrows from the vertical axis to the horizontal axis at the lower tail probability of 0.5 in order to find the value of the random variable whose lower tail probability is 0.5. In this case, the arrows start from 0.5 and points toward the value 0 on the horizontal axis (which represents the values of the random variable). We say zero is the *0.5 quantile* of the random variable X. In general, the 0.5 quantile of a random variable is called its *median*. For a normally distributed random variable, mean and median are the same.

As another example, suppose that we are interested in the probability of being underweight (BMI $\leq$ 18.5) assuming that the distribution of BMI is $N(25, 5^2)$. We can use R-Commander to find the lower tail probability 18.5 based on the normal distribution with mean 25 and standard deviation 5. This probability is 0.1. We could say that the BMI values of the 10% of the population are 18.5 or less. Now if we want to find the value of BMI that 10% of population fall below that, the answer is the 0.1 quantile, which is 18.5 in this case.

In R-Commander, we can obtain the quantiles for normal distribution (or other available distributions) by choosing `Distributions` $\rightarrow$ `Continuous distributions` $\rightarrow$ `Normal distribution` $\rightarrow$ `Normal quantiles`. We need to specify the parameters of the distribution (e.g., $\mu = 25$ and $\sigma = 5$ for the BMI example) and then input `Probabilities` (e.g., 0.1). This is shown in Fig. 5.16. Make sure that the option `Lower tail` is checked, so that R-Commander regards the given probability as a lower tail probability. The result is given in the *Output* window.

We can also use R-Commander to plot the cdf of discrete distributions and obtain the quantiles for a given lower tail probability. For the number of physician visits per year, we assumed that the random variable has Poisson(2.5) distribution. To

Fig. 5.16 Obtaining the normal quantile function in R-Commander. Here, we are finding the value corresponding to the lower tail probability of 0.1 (i.e., 0.1 quantile) for the standard normal distribution
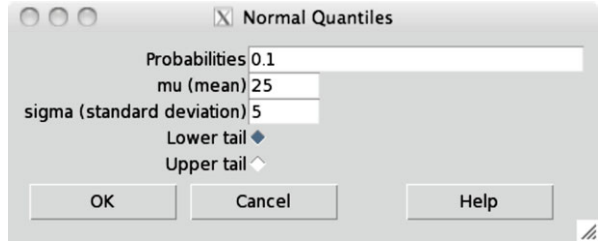
Fig. 5.17 Plot of the cdf for a Poisson distribution with $\mu = 2.5$. By following the arrows from $x = 3$ to the cumulative probability, we can obtain the lower tail probability of observing three or fewer physician visits per year: $P(X \leq 3) = 0.75$
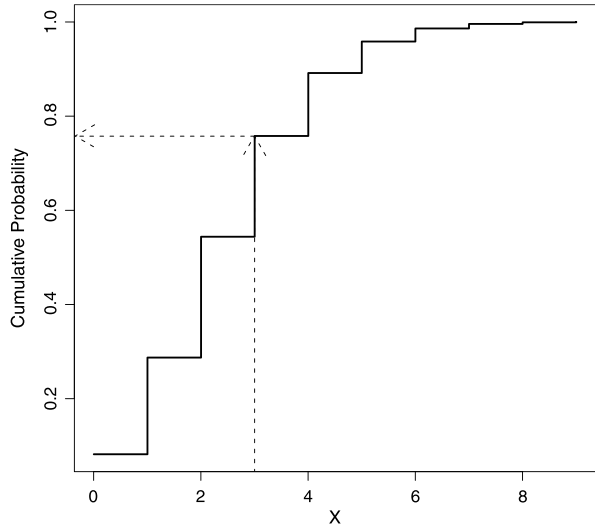
plot the corresponding cdf, click `Distributions → Discrete distribu-tions → Poisson distribution → Plot Poisson distribution`; then enter "2.5" for the `Mean` and select `Plot distribution function`. The resulting graph is shown in Fig. 5.17. Since this variable can only take non-negative integers, the cdf is a *step function* (i.e., its plot looks like a set of steps). Indeed, the cdf for all discrete distributions is a step function, whereas the cdf for all continuous distributions is a smooth curve. Note, however, that similar to the cdf of continuous distributions, the cdf plot of discrete distributions is a nondecreasing function as x increases.

As before, we can use the cdf plot to find the lower tail probability of each pos-sible value. For example, the lower tail probability of three physician visits or fewer per year is obtained in Fig. 5.17 by following the arrows from $x = 3$ to the cumu-lative probability $P(X \leq 3) = 0.75$. Note that the vertical arrow falls between two steps. However, to find the lower tail probability for discrete random variables, we always use the step on the right as shown in Fig. 5.17. We can use R-Commander to find the quantiles of discrete random variables by following similar steps as de-scribed for normal distributions above.

5.7 Scaling and Shifting Random Variables

In Chap. 2, we saw that if we multiply the observed values of a random variable by a constant a, its sample mean, sample standard deviation, and sample variance will be multiplied by a, $|a|$, and a^2, respectively. We also saw that if we add a constant b to the observed values of a random variable, that constant value will be added to the sample mean, but the sample standard deviation and sample variance remain unchanged. Similar rule applies to the theoretical (population) mean and variance of random variables. If $Y = aX + b$, then

$$\mu_Y = a\mu_X + b,$$
$$\sigma_Y^2 = a^2\sigma_X^2,$$
$$\sigma_Y = |a|\sigma_X.$$

Here, μ_X, σ_X^2, and σ_X are the mean, variance, and standard deviation of the random variable X, and μ_Y, σ_Y^2, and σ_Y are the mean, variance, and standard deviation of the random variable Y. Note that subtracting a constant b is the same as adding $-b$, and dividing by a is the same as multiplying by $1/a$.

For example, we assumed that the mean and variance of the random variable X representing SBP are $\mu_X = 125$ and $\sigma_X^2 = 15^2 = 225$, respectively. (The standard deviation is 15.) Now if we create a new random variable $Y = X - 125$, which is the original random variable minus its mean, we have

$$\mu_Y = \mu_X - 125 = 125 - 125 = 0,$$
$$\sigma_Y^2 = \sigma_X^2 = 225,$$
$$\sigma_Y = \sigma_X = 15.$$

Further, suppose that we create a new variable $Z = Y/15$, which is obtained by dividing the random Y by its standard deviation (i.e., multiplying by $1/15$). The mean, variance, and standard deviation of this new variable are

$$\mu_Z = \mu_Y/15 = 0/15 = 0,$$
$$\sigma_Z^2 = \sigma_Y^2/15^2 = 225/225 = 1,$$
$$\sigma_Z = \sigma_Y/|15| = 15/15 = 1.$$

Therefore, the newly created random variable Z has mean 0 and variance 1.

The process of shifting (changing the mean) and scaling (changing the variance) a random variable to create a new random variable with mean zero and variance one is called **standardization**. For this, we first subtract the mean μ and then divide the result by the standard deviation σ.

It can be shown that if a random variable has normal distribution, it will remain normally distributed if we add a constant to it or multiply it by a constant. For the above example, suppose that the distribution of the original variable is normal: $X \sim N(125, 15^2)$. Therefore, the distribution of Y and Z are also normal. More specifically, since the mean and variance of Z are 0 and 1, the distribution of Z is $N(0, 1)$, i.e., the standard normal distribution. In general,

$$Z = \frac{X - \mu}{\sigma}.$$

By standardizing a normally distributed random variable (i.e., subtracting μ and dividing the result by σ), we obtain a new random variable with the standard normal $N(0, 1)$ distribution.

5.8 Sum of Two Random Variables

Consider two random variables X and Y. By adding these two random variables, we obtain a new random variable $W = X + Y$. Regardless of the distribution of random variables X and Y, we can find the mean of W as follows:

$$\mu_W = \mu_X + \mu_Y.$$

If the random variables are independent (i.e., they do not affect each other probabilities), then we can find the variance of W as follows:

$$\sigma_W^2 = \sigma_X^2 + \sigma_Y^2.$$

The above results are true regardless of the distributions of the two random variables. If the two random variables have normal distributions, finding the mean and variance of W fully specifies its distribution. Therefore, we can use the following rules to find the distribution of W:

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then assuming that the two random variables are independent, we have

$$W = X + Y \sim N\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right).$$

In general, if two random variables are normally distributed, their sum is also normally distributed with the mean equal to the sum of the means and variance equal to the *sum* of the variances.

As an example, suppose that the distribution of weight for a population of people suffering from anorexia is $X \sim N(80, 5^2)$. Now suppose that a new treatment

(e.g., supplement plus diet) has shown promising results in helping patients gain weight. Of course, the effect of this treatment varies from one person to another. For illustration purposes, we assume that the amount of weight gain, denoted Y, is itself a normally distributed random variable with mean 5 and standard deviation 2: $Y \sim N(5, 2^2)$. If we apply this treatment to the population, we expect that it results in a healthier weight distribution. Denote the post-treatment weight of people in this population W, where $W = X + Y$. (The new weight is the pretreatment weight plus the weight gain due to the treatment.) The mean of this new variable is

$$\mu_W = \mu_X + \mu_Y = 80 + 5 = 85.$$

If the two variables are *independent* (i.e., the treatment effect Y does not depend on the original weight X), then the variance of W is

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 = 25 + 4 = 29.$$

Since we assumed that X and Y are normally distributed, W itself has normal distribution with mean 85 and variance 29: $W \sim N(85, 29)$.

If we subtract Y from X, then

$$W = X - Y.$$

Subtracting Y from X is the same as multiplying the random variable Y by -1 and adding the result to X. When we multiply Y by -1, its mean is multiplied by -1. Therefore,

$$\mu_W = \mu_X - \mu_Y.$$

If the two variables are independent,

$$\sigma_W^2 = \sigma_X^2 + \sigma_Y^2.$$

Note that we still *add* the variances, since multiplying a random variable Y by -1 does not change its variance; the variance is multiplied by $(-1)^2 = 1$.

As before, if the two random variables have normal distributions, finding the mean and variance of W fully specifies its distribution. Therefore,

$$W = X - Y \sim N\left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2\right).$$

Again, notice that the variances are added even though we are subtracting the two random variables.

As an example, suppose that there is a diet that helps people lose 10 pounds on average in one month. Of course, the results vary among different people. We denote the amount of weight reduction due to the diet Y and assume that Y has the normal $N(10, 2)$ distribution. We denote the initial weight (i.e., before starting the diet) X and assume that for the population of interest, $X \sim N(250, 15)$. For individuals in

this population, weight after one month of dieting is $W = X - Y$, which has the following distribution:

$$W = X - Y \sim N(250 - 10, 15 + 2).$$

That is, the population mean reduces to 240, but the population variance increases to 17.

5.9 Advanced

In this section, we continue our discussion of probability distributions. We provide the mathematical form of some of the distributions discussed in previous sections, and introduce some other important probability distributions. We also show a simple approach for comparing the assumed probability distribution of a random variable with the distribution of its observed values. Finally, we provide some useful R functions for working with probability distributions.

5.9.1 More on Probability Distributions

Thus far, we have studied several probability distributions for discrete and continuous random variables. While we showed how to plot the pmf for discrete random variables and the pdf for continuous random variables, and use them to obtain probabilities, we did not explicitly specified the mathematical forms of these functions. Here, we specify the corresponding mathematical forms of pmf and pdf for four widely used distributions.

The mathematical form of the probability mass function for the Bernoulli(μ) distribution is

$$P(X = x) = \mu^x (1 - \mu)^{1-x}.$$

If, for example, $\mu = 0.8$, the probability of observing $X = 1$ and the probability of observing $X = 0$ are

$$P(X = 1) = 0.8^1 (1 - 0.8)^0 = 0.8,$$
$$P(X = 0) = (0.8)^0 (1 - 0.8)^1 = 0.2.$$

The mathematical form of the probability mass function for the Binomial(n, μ) distribution is

$$P(X = x) = \binom{n}{x} \mu^x (1 - \mu)^{n-x},$$

$$\text{where } \binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

For any positive integer a, $a! = a \times (a-1) \times \cdots \times 2 \times 1$ and $0! = 1$. We can use the above formula to answer questions such as "what is the probability that 6 out of 10 patients survive when the probability of the survival is 0.8?" Plugging in $x = 6$, $n = 10$, and $\mu = 0.8$, we find

$$P(X = 6) = \binom{10}{6} 0.8^6 (1 - 0.8)^{10-6} = 0.09.$$

The mathematical form of the probability mass function for a random variable with a Poisson(μ) distribution is

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!},$$

where $e \approx 2.72$ is the base of the natural logarithm. For example, if the rate of physician visits is $\mu = 2.5$ per year, the probability of three visits per year is

$$P(X = 3) = \frac{2.5^3 e^{-2.5}}{3!} = 0.21.$$

For continuous random variables, we denote the density function as $f(x)$. For the $N(\mu, \sigma^2)$ distribution, the probability density function, denoted as $f(x)$, is

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

For the standard normal distribution, where $\mu = 0$ and $\sigma^2 = 1$, the density simplifies to

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

Plugging a specific value of x into the density function gives us the height of the probability density curve at that point. This is NOT the probability $P(X = x)$, which is always 0 for any specific value of x. The pdf $f(x)$ only specifies the function that can be used to calculate the probability of any given interval.

5.9.2 Some Other Commonly Used Probability Distributions

The discrete and continuous probability distributions we discussed in previous sections are by far the most commonly used probability distributions. Later, we will discuss two other important distributions (especially for hypothesis testing), namely, the F-distribution in Chap. 9 and the χ^2-distribution (chi-square distribution) in Chap. 10. Here, we discuss some other important probability distributions, which are also commonly used in statistical analysis of data.

Discrete Uniform Distribution   Suppose a random variable X can take one of k integers $1, \ldots, k$ with equal probabilities such that

$$P(X = x) = \begin{cases} \frac{1}{k} & \text{if } x \in \{1, \ldots, k\}, \\ 0 & \text{otherwise.} \end{cases}$$

We say that X has the discrete uniform distribution on integers $1, \ldots, k$. For example, the probability distribution for the outcome of rolling a fair die is a discrete uniform distribution on integers $1, \ldots, 6$,

$$P(X = x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, \ldots, 6\}, \\ 0 & \text{otherwise.} \end{cases}$$

The sequence of integers could represent k distinct categories (e.g., 50 states, different flu viruses). In general, the distribution can be defined for any finite sequence of integers (e.g., $\{-5, -4, \ldots, 1, 2\}$).

Hypergeometric Distribution   For the Binomial(n, μ) distribution, we mentioned that the random variable can be considered as the sum of n *independent* Bernoulli variables, where the probability for the outcome of interest is the same, μ, for each Bernoulli variable. Now we consider situations where the Bernoulli variables are dependent; that is, they affect each others' probabilities. More specifically, we consider situations where the dependency among Bernoulli variables is due to sampling without replacement from a finite population of size N.

As an example, suppose we have performed medical tests on $N = 200$ people and found $m = 10$ of them are affected by a specific disease, and the remaining $n = 190$ people are not affected. We decide to repeat the tests again, but this time on a smaller sample of size $k = 50$, who are randomly selected one-by-one and without replacement from the finite population of 200 people. To this end, we randomly select the first person. The outcome of a test is a Bernoulli variable with two possible values: zero for non-diseased, and one for diseased. The probability of having the disease is 10/200 for this person. After performing the medical test, we remove the person from the population. (Note that the sampling is without replacement.) If the person had the disease, the probability of having the disease for the next person randomly selected from the population of 200 people is 9/199 because there are 199 people left in the population, where 9 of them have the disease. However, if the first person did not have the disease, the probability of having the disease for the second person is 10/199. The outcome for the second person is also a Bernoulli variable. Unlike the Bernoulli variables for binomial distributions, the Bernoulli variables in this case are dependent since the outcome of the first Bernoulli variable affects the probability of the outcome of interest for the second Bernoulli variable. Following the same logic, all $k = 50$ Bernoulli variables are dependent.

We use the random variable X to denote the number of people who are affected by the disease within the sample of size $k = 50$ randomly selected (without replacement) from the finite population of size $N = 200$. The probability distribution of X is called the **hypergeometric distribution**. In general, a hypergeometric distribu-

tion has the following probability mass function:

$$P(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}, \quad \max(0, k - n) \leq x \leq \min(m, k).$$

Here, m is the number of cases with the outcome of interest, n is the number of cases that do not have the outcome of interest, and k is the number of Bernoulli trials (i.e., sample size). (We use these notations to be consistent with R functions.) The hypergeometric distribution is usually explained in terms of drawing k balls without replacement from an urn with m white balls and n black balls. In this case, the random variable X represents the number of while balls among the k drawn balls. Note that the specific value x for the random variable must be greater than or equal to zero. Further, x cannot exceed m (the total number of white balls in the urn) or k (the sample size). Therefore, $x \leq \min(m, k)$. Also, the number of black balls in our sample, $k - x$, cannot exceed n the total number of black balls in the urn. Therefore, $k - x \leq n$, which means $k - n \leq x$. The mean and variance of a random variable with hypergeometric distribution are as follows:

$$\mu = \frac{km}{m + n},$$

$$\sigma^2 = \frac{kmn}{(m + n)^2} \frac{m + n - k}{m + n - 1}.$$

Using R-Commander, we can plot hypergeometric distributions, find the probability of a specific value of the random variable, generate random samples, find the quantile for a given lower tail probability, and find the lower tail probability for a given value of the random variable. (The steps are similar to those we discussed for other probability distributions.) For example, suppose we want to find the probability of $x = 4$ diseased people (white balls) when we take a sample of size $k = 50$ from a population (urn) of size $N = 200$, which includes $m = 10$ diseased people and $n = 190$ non-diseased people (black balls). In R-Commander, click `Distribu-tions` → `Discrete distributions` → `Hypergeometric distri-butions` → `Hypergeometric probabilities`. Then, set $m = 10$, $n = 190$, and $k = 50$. In the *Output* window, R-Commander provides the probabilities for values from 0 to 7. Theoretically, x could be any integer from 0 to 10 for this example. In this case, however, the probabilities are almost zero for numbers from 8 to 10 so they are not provided by R-Commander. From the table provided in the *Output* window, we find $P(X = 4) = 0.146$.

The Uniform Distribution   Similar to the discrete uniform distribution on a sequence of integers, we can define a continuous uniform distribution on an interval $[\alpha, \beta]$, which is shown as Uniform(α, β). This is in fact one of the simplest continuous distributions, where the probability density function is a straight horizontal line over the interval $[\alpha, \beta]$,

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the value of the density function is constant (i.e., it does not depend on x).

Within all possible uniform distributions, Uniform(0, 1) is the most widely used. For this probability distribution, the pdf has the following form:

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

According to this distribution, the probability of any interval is equal to the length of the interval. (Since the hight of the pdf for this distribution is equal to one, the area under the line for each interval is the same as the length of that interval.) Therefore, $P(0.1 < X \leq 0.2) = P(0.7 < X \leq 0.8) = 0.1$. That is, while the first interval includes small values of the random variable, and the second interval includes large values of the random variable, both intervals are equally probable since they have the same length. To plot uniform distributions, in R-Commander click `Distributions` → `Continuous distributions` → `Uniform distributions` → `Plot uniform distribution` and specify the interval limits α and β. The default values these parameters are 0 and 1 for Uniform(0, 1). Use R-Commander to plot the pdf of Uniform(0, 1).

The Beta Distribution   The beta distribution is commonly used as the probability distribution of random variables whose range is the set of real numbers from 0 to 1. As we discuss in Chap. 13, this distribution is often used in Bayesian statistical inference.

A beta distribution is specified by two parameters, α, which is called *shape 1*, and β, which is called *shape 2*. We should a beta distribution with parameters α and β as Beta(α, β). Both parameters must be positive. If $X \sim$ Beta(α, β), then the mean and variance of X are as follows:

$$\mu = \frac{\alpha}{\alpha + \beta},$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

To plot beta distributions, in R-Commander click `Distributions` → `Continuous distributions` → `Beta distributions` → `Plot beta distribution` and specify `Shape1` and `Shape 2`. As an example, plot the probability density function for Beta(1, 1) by setting both parameters to 1. Compare the resulting pdf with that of Uniform(0, 1) you plotted previously. Indeed, the Beta(1, 1) distribution is the same as the Uniform(0, 1) distribution.

The Lognormal Distribution   Using normal distributions could be problematic for random variables that can take positive real numbers only. (Theoretically, normally distributed random variables can take negative and positive values.) This is especially true if the possible values of the random variable tend to be close to zero and/or the distribution of observed values is heavily right-skewed. For example, assuming normal distribution for tumor size (in millimeter) might not be

appropriate since the values for this variable are positive and typically close to zero. For such random variables, we often use the log-transformation, $Y = \log(X)$, and assume a normal distribution for the log-transformed variable, Y. That is, $Y = \log(X) \sim N(\mu, \sigma^2)$. While X can be positive only, Y can take negative and positive values.

When we assume $N(\mu, \sigma^2)$ for $\log(X)$, we say that X itself has a *lognormal* distribution with parameters μ and σ^2. Note that in this case, while the mean and variance of Y (i.e., the random variable after log-transformation) are μ and σ^2 respectively, the mean and variance of X (i.e., the original random variable before log-transformation) are $e^{\mu + 0.5\sigma^2}$ and $(e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$ respectively. To plot lognormal distributions, in R-Commander click `Distributions` → `Continuous distributions` → `Lognormal distributions` → `Plot lognormal distribution`.

5.9.3 Quantile–Quantile Plots

When we assume a distribution for a random variable, we should evaluate the appropriateness of our assumption. We do this by assessing how well the theoretical distribution with its estimated parameters fits the observed data. That is, how close the two distributions are. The statistical methods used for this purpose are called the **goodness-of-fit** tests. A common approach for assessing the assumption of a specific probability distribution is based on the **quantile–quantile** (Q–Q) plots. In general, Q–Q plots compare two distributions by plotting the quantiles of one distribution against the quantiles of the other distribution. Therefore Q–Q plots can be used to compare the theoretical distribution of a random variable to the data distribution (i.e., distribution of observed values). Specifically, if the quantiles of the data distribution exactly match those of the theoretical distribution, the points on the Q–Q plot will fall in straight line.

Figure 5.18 shows the Q–Q plot to test the normality assumption for the BMI variable in the `Pima.tr` data set. To create this plot using R-Commander, make sure `Pima.tr` is the active data set, then click `Graphs` → `Quantile-comparison plot`. Choose `bmi` as the `Variable` and `Normal` for the `Distribution`. For each point, the horizontal axis represents the quantile value for the standard normal distribution, and the vertical axis represents the quantile values based on the observed data. Notice that the points are very close to the straight line confirming the appropriateness of the normal assumption. As we move further from the center, the points tend to deviate from the line. However, the points remain between the dashed lines, which indicates that the deviations from the straight solid line remain within an acceptable range.

For comparison, repeat the above steps to create the Q–Q plot for the `age` variable in the `Pima.tr` data set. The resulting plot is shown in the right panel of Fig. 5.18. The assumption of normal distribution is clearly not appropriate for this variable: the points do not follow a linear trend and very often fall outside of the
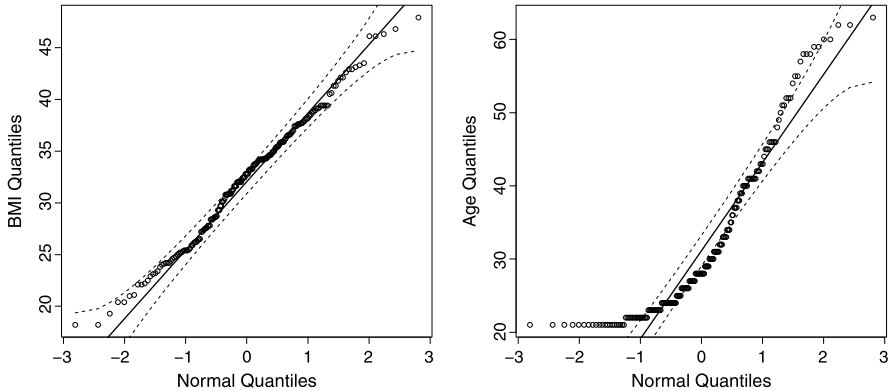
Fig. 5.18 Using the quantile–quantile plot to evaluate the normality assumption for the distribution of BMI (*left panel*) and the distribution of age (*right panel*). Each point represents a quantile. The horizontal axis represents the quantile value for the theoretical distribution (normal), and the vertical axis represents the quantile values based on the observed data

dashed lines. This was expected since the distribution of observed values for `age` is right skewed.

5.9.4 Probability Distributions with R Programming

As in R-Commander, it is possible to plot probability distributions and obtain probabilities directly from the command line.

Binomial Distribution Assume that we want to examine 10 people for a disease that has probability of 0.2 in the population of interest. The number of people (out of 10) who are affected, denoted as Y, has Binomial(10, 0.2) distribution. Let us first simulate five random samples from this distribution (i.e., examine five groups each with 10 people):

```
> rbinom(5, size = 10, prob = 0.2)

[1] 0 0 1 2 3
```

where the first argument to the `rbinom()` functions specifies the number of random samples. The `size` option is the number of Bernoulli trials (here, $n = 10$), and the `prob` option is the probability for the outcome of interest. Each randomly generated number represents the number of people affected by the disease out of 10 people. If we set `size=1`, we will be simulating random samples from the corresponding Bernoulli distribution. For example, we can simulate the disease status for a group of 10 people:

```
> rbinom(10, size = 1, prob = 0.2)
```

```
 [1] 1 0 0 0 0 0 1 1 0 0
```

Now suppose that we want to know the probability of observing 3 out of 10 people affected by the disease: $P(X = 3)$. Then we need probability mass function dbinom(), which returns the probability of a specific value:

```
> dbinom(3, size = 10, prob = 0.2)
```

```
[1] 0.2013266
```

Along with the value of the random variable, 3, the other arguments of the dbinom() function are the number of Bernoulli trials (size=10) and the probability (prob=0.2) for the event of interest.

We can also create a vector x of the possible values of X and then use this vector as input to dbinom() function:

```
> x <- 0:10
> Px <- dbinom(x, size = 10, prob = 0.2)
> round(Px, 2)
```

```
 [1] 0.11 0.27 0.30 0.20 0.09 0.03 0.01 0.00
 [9] 0.00 0.00 0.00
```

Using vectors x and Px, we can plot the probability mass function (pmf), similar to the one shown in Fig. 5.2:

```
> plot(x, Px, type = "h", xlab = "Number of Successes",
+      ylab = "Probability Mass",
+      main = "Binomial(10, 0.2)")
> points(x, Px, pch = 16)
> abline(h = 0, col = "gray")
```

In the plot() function, the first argument provides the values for the horizontal axis, and the second argument provides the values for the vertical axis. We use the type="h" option to create "histogram-like" vertical lines. The points at the top of the lines are added with the points() function, whose option pch=16 gives filled-in circles. Similar to the plot() function, the first and second arguments provide the coordinates of points. Lastly, the gray horizontal line at 0 is added with abline(h=0, col="gray").

The functions points() and abline() only add points and lines to an existing plot; they cannot be used alone. The abline() function can be used to add a straight line to an existing plot. (You first need to create a plot before using abline.) For example, abline(h=2) draws a horizontal line two units above the origin, abline(v=-1) draws a vertical line one unit to the left of origin, and abline(a=-5, b=2) draws a line with intercept −5 and slope 2. By default, abline() draws a solid line. We can set the line type to dashed line by using the option lty=2:

```
> abline(h = 0, lty = 2)
```

Try other values, such as 3 and 4, for the `lty` option. To add additional points to an existing plot, you can use the `points()` function. To learn more about this function, enter the command `?points` or `help(points)`.

Now suppose that we are interested in the probability of observing three or fewer affected people in a group of 10. We could of course sum the values of pmf: $P(Y \leq 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3)$. However, it is easier to use the *cumulative distribution function* for a binomial random variable, `pbinom()`, to obtain the lower tail probability:

```
> pbinom(3, size = 10, prob = 0.2, lower.tail = TRUE)

[1] 0.8791261
```

As before, the arguments `size=10` and `prob=0.2` specify the parameters of the binomial distribution. The option `lower.tail=TRUE` tells R to find the lower tail probability. By changing the `lower.tail` option to false (`FALSE`), we can find the upper tail probability $P(Y > 3)$.

On the other hand, to obtain the 0.879 quantile, we use the `qbinom()` function:

```
> qbinom(0.879, size = 10, prob = 0.2,
+       lower.tail = TRUE)

[1] 3
```

Poisson Distribution   Suppose that on average 4 people visit the hospital each hour. Then we can represent the hourly number of hospital visitation as $X \sim$ Poisson(4) and simulate 12 samples from this distribution:

```
> rpois(12, 4)

 [1] 3 3 3 3 3 1 2 5 5 4 3 5
```

These randomly generated numbers can be regarded as the number of people visiting the hospital at different hours. Similar to the `rbinom()` function, the first parameter to the `rpois()` function is the number of samples, and the remaining argument specifies the distribution parameter.

Suppose that we want to know the probability that six people visit the hospital in an hour. Then we would use the probability mass function `dpois()`:

```
> dpois(6, 4)

[1] 0.1041956
```

Here, 6 is the specific value of the random variable, and 4 is the distribution parameter. As before, we can create a plot of the pmf by first creating a vector of possible values and finding their corresponding densities.

To find the probability of six or fewer people visiting the hospital (as opposed to the probability that exactly six people visit), we need to find the lower tail probability of $x = 6$. For this, we use the `ppois()` function:

```
> ppois(6, 4)

[1] 0.889326
```

The 0.889 quantile of the distribution is

```
> qpois(0.889, 4)

[1] 6
```

Normal Distribution   Suppose that BMI in a specific population has a normal distribution with mean of 25 and variance of 16: $X \sim N(25, 16)$. Then we can simulate 5 values from this distribution using the `rnorm()` function:

```
> rnorm(5, mean = 25, sd = 4)

[1] 26.71568 32.66739 26.99269
[4] 30.27329 29.58406
```

These numbers can be regarded as BMI values for five randomly selected people from this population. In the *rnorm()* function, the first parameter is the number of samples, the second parameter is the mean, and the third parameter is the standard deviation (not the variance).

Now let us plot the pdf of this distribution. A normal random variable can take any value from $-\infty$ to ∞. However, according to the *68–95–99.7% rule* approximately 99.7% of the values fall within the interval [13, 37] (i.e., within 3 standard deviations of the mean). Therefore, the interval [10, 40] is wide enough to plot the distribution:

```
> x <- seq(from = 10, to = 40, length = 100)
```

Here, vector `x` is a sequence of length 100 from 10 to 40. We can then find and plot the density for each point in the vector `x`:

```
> fx <- dnorm(x, mean = 25, sd = 4)
> plot(x, fx, type = "l", xlab = "BMI",
+     ylab = "Density", main = "N(25, 16)")
> abline(h = 0, col = "gray")
```

The `dnorm()` function returns the height of the density curve at a specific point and requires the parameters of the `mean` and the standard deviation `sd`. In the `plot()` function, we are using `type="l"` to plot the points as a continuous line (curve).

Recall that for continuous variables, the probability of a specific value is always zero. Instead, for continuous variables, we are interested in the probability of observing a value in a given interval. For instance, the probability of observing a BMI less than or equal to 18.5 is the area under the density curve to the left of 18.5. In R, we find this probability with the cumulative distribution function `pnorm()`:

```
> pnorm(18.5, mean = 25, sd = 4,
+      lower.tail = TRUE)

[1] 0.05208128
```

Once again, we can find the upper tail probability $P(X > 22)$ by setting the option `lower.tail=FALSE`. The `qnorm()` returns the quantile for normal distributions is. For example, the 0.05 quantile for the above distribution is

```
> qnorm(0.05, mean = 25, sd = 4,
+      lower.tail = T)

[1] 18.42059
```

We can find the probability of a BMI between 25 and 30 by subtracting their lower tail probabilities, $P(25 < X \le 30) = P(X \le 30) - P(X \le 25)$:

```
> pnorm(30, mean = 25, sd = 4) -
+      pnorm(25, mean = 25, sd = 4)

[1] 0.3943502
```

We can also create a plot of the cdf by using vector `x` as input to `pnorm()` function:

```
> Fx <- pnorm(x, mean = 25, sd = 4)
> plot(x, Fx, type = "l", xlab = "BMI",
+      ylab = "Cumulative Probability",
+      main = "N(25, 16)")
> abline(h = 0, col = "gray")
```

In general, for each distribution, the random number generating function starts with r, the probability mass function or probability density function starts with d, the distribution function (i.e., cdf) starts with p, and the quantile function starts with q.

For the t-distribution, these functions are `rt()`, `dt()`, `pt()`, and `qt()`.
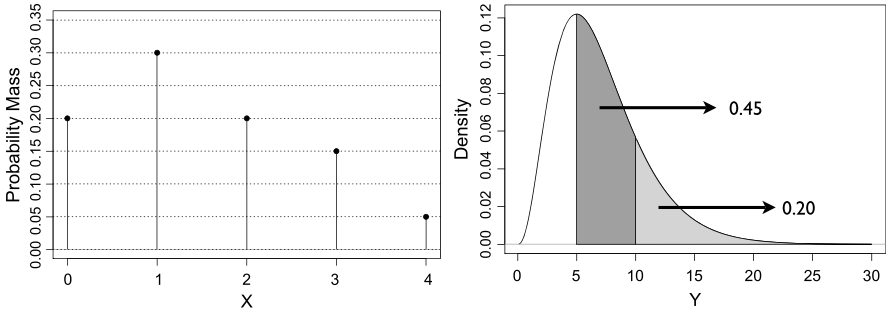
Fig. 5.19 *Left panel*: Probability mass function of random variable X. *Right panel*: Probability density function of variable Y

5.10 Exercises

1. What would be the most appropriate probability distribution for each of the following random variables:
   (a) Whether a tumor is benign or malignant.
   (b) Number of people with a malignant tumor out of 10 patients with tumor.
   (c) Size of tumors.
   (d) Number of people diagnosed with malignant tumor in California every year.
2. Consider the two plots of Fig. 5.19. In the right panel, the dark-gray area is 0.45, and the light-gray area is 0.20. Write down these probabilities:
   (a) $P(X < 3)$.
   (b) $P(1 < X \leq 4)$.
   (c) $P(Y > 5)$.
3. Consider Binomial(10, 0.3) distribution. Do the following tasks:
   (a) Plot the probability mass function and cumulative distribution function.
   (b) Write down the mean and standard deviation of each distribution.
   (c) Find the lower tail probability of 4.
   (d) What is the probability that the value of the random variable is 2?
   (e) What is the probability that the value of the random variable is greater than 2 and less than or equal to 4?
4. Consider $N(3, 2.1)$ distribution. Do the following tasks:
   (a) Plot the probability density function and cumulative distribution function.
   (b) Write down the mean and standard deviation of each distribution.
   (c) Find the lower tail probability of 4.
   (d) What is the probability that the value of the random variable is 2?
   (e) What is the probability that the value of the random variable is bigger than 2 and less than or equal to 4?
5. For the probability distributions Binomial(100, 0.3) and $N(30, 21)$, find the lower tail probability of 35 and the upper tail probability of 27. Compare the results based on the two distributions.
6. Suppose that X has the t-distribution with 6 degrees of freedom.

  (a) Find the lower tail probabilities of -1 and 1.5.
  (b) Find the 0.95 and 0.9 quantiles.
7. National Heart, Lung, and Blood Institute defines the following categories based on Systolic Blood Pressure (*SBP*):

  • Normal: $SBP \leq 120$.
  • Prehypertension: $120 < SBP \leq 140$.
  • High blood pressure: $SBP > 140$.

  If *SBP* in the US has a normal distribution such that $SBP \sim N(125, 15^2)$,
  (a) Use R-Commander to find the probability of each group.
  (b) Find the intervals that include 68, 95, and 99.7% of the population.
  (c) What are the lower and upper tail probabilities for *SBP* equal to 115?
8. Assume that BMI in US has the $N(27, 6^2)$ distribution. Following the recommendation by National Heart, Lung, and Blood Institute, we define the following BMI categories:

  • Underweight: $BMI \leq 18.5$.
  • Normal weight: $18.5 < BMI \leq 25$.
  • Overweight: $25 < BMI \leq 30$.
  • Obesity: $BMI > 30$.

  (a) Use R-Commander to find the probability of each group.
  (b) Find the intervals that include 68, 95, and 99.7% of the population.
  (c) What is the probability of being either underweight OR obese?
  (d) What are the lower and upper tail probabilities for BMI equal to 29.2?
9. For the above question, we denote BMI as X. Find the value x such that $P(X \leq x) = 0.2$. Next, find the value x such that $P(X > x) = 0.2$.
10. If the height (in inches) of newborn babies has the $N(18, 1)$ distribution, what is the probability that the height of a newborn baby is between 17 and 20 inches? What is the distribution of height in centimeters (1 inch $= 2.54$ cm)? Using this distribution, what is the probability that the height of a newborn baby is between 43.18 cm (17 inches) and 50.80 cm (20 inches)?
11. Suppose that the distribution of systolic blood pressure, X, among people suffering from hypertension is $N(153, 4^2)$. Further, suppose that researchers have found a new treatment that drops systolic blood pressure by 4 points on average. The effect of drug, Y, varies among patients randomly and does not depend on their current blood pressure level. If the variance of Y is 1, what is the mean (expectation) and variance of systolic blood pressure if every person in the population starts using the drug? What is the distribution of systolic blood pressure in this case if we assume that Y has a normal distribution?

Chapter 6
Estimation

6.1 Parameter Estimation

In the previous chapter, we discussed using random variables to represent character-istics of a population (e.g., BMI, disease status). Furthermore, we discussed some commonly used probability distributions for discrete and continuous random vari-ables. As we mentioned, we are specifically interested in population mean and pop-ulation variance of a random variable. These quantities are unknown in general. We refer to these unknown quantities as **parameters**. Here, we use parameters μ and σ^2 to denote the unknown population mean and variance respectively. Note that for all the distributions we discussed in the previous chapter, the population mean and variance of a random variable are related to the unknown parameters of probabil-ity distribution assumed for that random variable. Indeed, for normal distributions $N(\mu, \sigma^2)$, which are widely used in statistics, the population mean and variance are exactly the same parameters used to specify the distribution.

In this chapter, we discuss statistical methods for parameter **estimation**. Estima-tion refers to the process of guessing the unknown value of a parameter (e.g., pop-ulation mean) using the observed data. For this, we will use an **estimator**, which is a **statistic**. A statistic is a function of the observed data only. That is, it does not depend on any unknown parameter, and given the observed data, we should be able to find its value. For example, the sample mean is a statistic. Given a sample of data, we can find the sample mean by adding the observed values and dividing the result by the sample size. No unknown parameter is involved in this process.

Sometimes we only provide a single value as our estimate. This is called **point estimation**. The point estimate for μ is denoted $\hat{\mu}$, and the point estimate for σ^2 is denoted $\hat{\sigma}^2$. (In general, we use the "hat" notation for point estimates.) Point estimates do not reflect our uncertainty when estimating a parameter. We always remain uncertain regarding the true value of the parameter when we estimate it using a sample from the population. To address this issue, we can present our estimates in terms of a range of possible values (as opposed to a single value). This is called **interval estimation**.

Unless stated otherwise, we assume that the population size N is large, so we can consider the number of individuals in the population as infinite for all prac-

tical purposes. Suppose that the random variable X represents a specific population characteristic we are investigating (e.g., BMI, height, blood pressure). We use $X_1, X_2, \ldots, X_n$ to denote n possible values of X obtained from a sample randomly selected from the population. The values of $X_1, X_2, \ldots, X_n$ themselves cannot be determined with certainty before they are observed, and they can change every time we take a different sample of size n from the population. Therefore, we treat $X_1, X_2, \ldots, X_n$ themselves as n random variables, and hence we reserve the use of capital letters for random variables. We typically assume that the samples are taken independently from each other and that they all have the same probability distribution, which is the probability distribution we assume for X. In this case, we say that the samples are *independent and identically distributed* (IID).

While theoretically we can have many different samples of size n, we usually have only one such sample in practice. We use $x_1, x_2, \ldots, x_n$ as the specific set of values we have observed in our sample. That is, x_1 is the observed value for X_1, x_2 is the observed value X_2, and so forth.

6.2 Point Estimation

In this section, we discuss the point estimations for the population mean, μ, and the population variance, σ^2. As mentioned above, the point estimate for μ is denoted $\hat{\mu}$, and the point estimate for σ^2 is denoted $\hat{\sigma}^2$.

6.2.1 Population Mean

For a population of size N, μ is calculated as

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},$$

where x_i is the value of the random variable for the ith member of the population.

> Given n observed values, $X_1, X_2, \ldots, X_n$, from the population, we can estimate the population mean μ with the sample mean:
>
> $$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$
>
> In this case, we say that $\bar{X}$ is an estimator for μ.

As our sample (the n representative members from the population) changes, the value of this estimator (sample mean) can also change. Therefore, the estimator itself is considered as a random variable and is denoted $\bar{X}$ in accordance to the general convention of capital letters for random variables we follow in this book.

We usually have only one sample of size n from the population $x_1, x_2, \ldots, x_n$. Therefore, we only have one value for $\bar{X}$, which we denote $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

where x_i is the ith observed value of X in our sample, and $\bar{x}$ is the observed value of $\bar{X}$.

As an example, consider the study by Mackowiak et al. [19] aimed at estimating the population mean for body temperature among healthy people. From a sample of $n = 148$ people, they estimated the unknown population mean with the sample mean $\hat{\mu} = \bar{x} = 98.25$. This estimate is lower than the commonly believed value of 98.6°F.

The sample size for this study was relatively small. We would expect that as the sample size increases, our point estimate based on the sample mean would become closer to the true population mean.

> The **Law of Large Numbers (LLN)** indicates that (under some general conditions such as independence of observations) the sample mean converges to the population mean ($\bar{X}_n \to \mu$) as the sample size n increases ($n \to \infty$). Informally, this means that the difference between the sample mean and the population mean tends to become smaller and smaller as we increase the sample size. The LLN provides a theoretical justification for the use of sample mean as an estimator for the population mean.

As an illustrative example, suppose that the true population mean for normal body temperature is 98.4°F. As we gradually increase the sample size from 100 to 5000, the plot of the sample means (i.e., the point estimates of the population mean) might look like Fig. 6.1. Here, the estimate of the population mean is plotted as a function of the sample size. As the sample size increases, the sample means converge to the true (but unknown) population mean $\mu = 98.4$.

The Law of Large Numbers is true regardless of the underlying distribution of the random variable. Therefore, it justifies using the sample mean $\bar{X}$ to estimate the population mean for continuous random variables, discrete random variables, whose values are counts (i.e., nonnegative integers), and for discrete binary variables, whose possible values are 0 and 1 only. For count variables, the mean is usually referred to as the *rate* (e.g., rate of traffic accidents). For binary random variables, the mean is usually referred to as the *proportion* of the outcome of interest (denoted as 1). Hence, we sometimes use the notation p instead of $\bar{x}$ for the sample mean of binary random variables.

As an example, suppose that we are interested in estimating the rate of physician visits during the first trimester for pregnant women. Using the `birthwt` data set discussed in previous chapters, our estimate of this rate is $\hat{\mu} = \bar{x} = 0.79$.

We can also use this sample to estimate the proportion of mothers who smoke during their pregnancy. The sample proportion of smoking mothers is 0.39: $\hat{\mu} =$

Fig. 6.1 Illustrating the Law
of Large Numbers. As the
sample size is increased, the
sample mean $\bar{X}$ converges to
the population mean μ. For
the temperature example, by
increasing n, $\bar{X} \to \mu = 98.4$
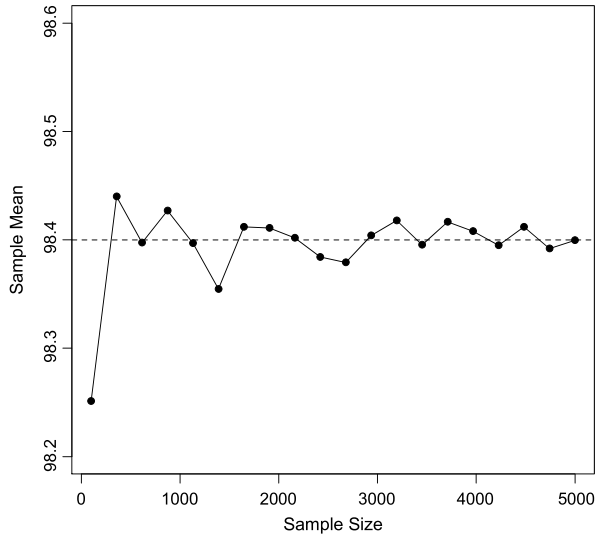
$\bar{x} = p = 0.39$. (The current estimate is much smaller according to CDC.) Now suppose that we want to estimate the number of smoking mothers in the whole population assuming that there are currently $N = 4$ million pregnant women in the US. To estimate this number, we can simply use our point estimate for the population proportion $p = 0.39$ as follows:

$$\text{estimated number of smoking pregnant women} = pN$$

$$= 0.39 \times 4{,}000{,}000$$

$$= 1{,}560{,}000.$$

6.2.2 Population Variance

The population variance is denoted as σ^2 and calculated as

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

This is the average of squared deviations of each observation x_i from the population mean μ.

Given n randomly sampled values $X_1, X_2, \ldots, X_n$ from the population and their corresponding sample mean $\bar{X}$, we can estimate the variance. A natural estimator for variance is

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

However, this estimator tends to underestimate the population variance. (On average, the values obtained by the above estimator are smaller than the true value of σ^2.)

To address this issue, a more commonly used estimator for σ^2 is the sample variance,

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

This is the sum of squared deviations from the sample mean divided by $n - 1$ instead of n. Dividing by $n - 1$ instead of n increases the value of the estimator by a small amount, which is enough to avoid underestimation associated with the more natural estimator. Therefore, the sample variance is the usual estimator of the population variance. Likewise, the sample standard deviation S (i.e., square root of S^2) is our estimator of the population standard deviation σ.

Again, we regard the estimator S^2 as a random variable (hence the capital-letter notation) since it changes as we change the sample. However, in practice, we usually have one set of observed values, $x_1, x_2, \ldots, x_n$, and therefore, only one value for S^2, which we denote as s^2:

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

For example, using the `Pima.tr` data set, our estimate of the population variance for BMI among Pima Indian women is $\hat{\sigma}^2 = s^2 = 37.6$.

For binary random variables with 0 and 1 values, we can show that the population variance σ^2 is equal to $\mu(1 - \mu)$, where μ is the population mean (proportion). (See the Bernoulli distribution discussed in the previous chapter.) Therefore, after we estimate the population mean μ using the sample mean (proportion) $\bar{x} = p$, we can use it to estimate the population variance instead of estimating σ^2 separately:

$$s^2 = p(1 - p).$$

For example, using the `birthwt` data set, we estimated that the proportion of mothers who smoke during their pregnancy is $p = 0.39$. Our estimate for the population variance is therefore

$$s^2 = 0.39 \times 0.61 = 0.24.$$

6.3 Sampling Distribution

As we emphasized, the value of estimators discussed so far (and all estimators in general) depend on the specific sample selected from the population. Indeed, if we repeat our sampling, we are likely to obtain a different value for an estimator. Therefore, we regard the estimators themselves as random variables. As a result, similar to any other random variable, we can talk about their probability distribution. Probability distributions for estimators are called **sampling distributions**. In this section, we focus on the sampling distribution of the sample mean $\bar{X}$. (For binary random variables, this is the same as the sample proportion.)

We start by assuming that the random variable of interest, X, has a normal $N(\mu, \sigma^2)$ distribution. Further, we assume that the population variance σ^2 is known, so the only parameter we want to estimate is μ. To this end, we use the sample mean $\bar{X}$ as our estimator for μ. We need to find the sampling distribution of $\bar{X}$ under these assumptions. Later, we discuss situations when σ^2 is not known and the random variable of interest is not normally distributed. As a running example, consider the random variable $X \sim N(125, 15^2)$ representing systolic blood pressure, whose population mean $\mu = 125$ is unknown to us, but we know the population variance $\sigma^2 = 15^2$. (The population standard deviation is $\sigma = 15$.)

Suppose that we take a sample of size $n = 2$ from the population. We denote the corresponding values obtained from this sample as X_1 and X_2. Following our general assumption, X_1 and X_2 are identically distributed. We write this as

$$X_1, X_2 \sim N(\mu, \sigma^2).$$

Further, we assume that X_1 and X_2 are independent; That is, they are independent and identically distributed (IID). In the previous chapter, we mentioned that for two independent and normally distributed random variables, their sum is also normally distributed, and its mean and variance are obtained by adding the means and variances of the two original random variables. Therefore,

$$X_1 + X_2 \sim N(\mu + \mu, \sigma^2 + \sigma^2) = N(2\mu, 2\sigma^2).$$

We can easily generalize this to the sum of n random variables:

$$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2).$$

We can rewrite this as

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2).$$

That is, the sum of n IID random variables with $N(\mu, \sigma^2)$ distribution is itself normally distributed with mean $n\mu$ and variance $n\sigma^2$.

If we divide $\sum_{i=1}^{n} X_i$ by n, we obtain the sample mean

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

This is the same as multiplying $\sum_{i=1}^{n} X_i$ by $1/n$. From the previous chapter we know that when we multiply a random variable by a constant (here, $1/n$), its mean
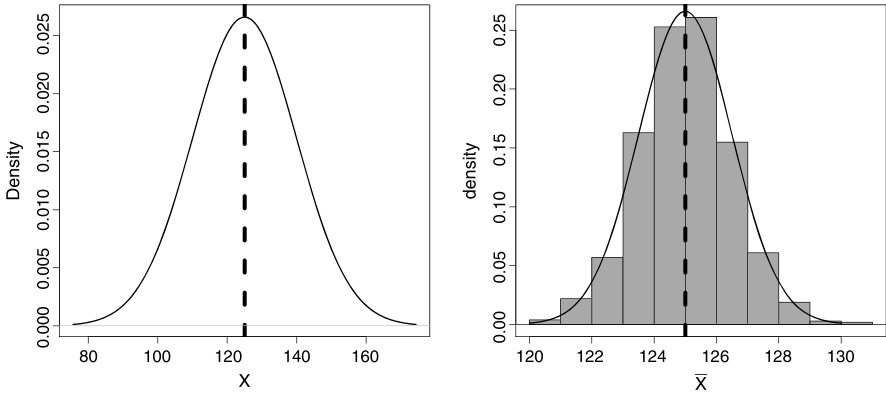
Fig. 6.2 *Left panel*: The (unknown) theoretical distribution of blood pressure, $X \sim N(125, 15)$. *Right panel*: The density curve for the sampling distribution $\bar{X} \sim N(125, 15^2/100)$ along with the histogram of 1000 sample means. The distribution of sample means is centered on the population mean (shown with the a *vertical line*), but its variance is much less than that of blood pressure itself. Note the different scales on the x-axis

is multiplied by that constant, and its variance is multiplied by the square of that constant. When we multiply $\sum_{i=1}^{n} X_i$ by $1/n$ to obtain the sample mean $\bar{X}$, the mean becomes $n\mu/n = \mu$, and the variance becomes $n\sigma^2/n^2 = \sigma^2/n$:

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

This is the sampling distribution of $\bar{X}$. The standard deviation of $\bar{X}$ can be obtained by taking the square root of its variance: $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$. The standard deviation of the sampling distribution in this case reflects the extent of the variability of the sample mean as an estimator for the population mean.

For the above blood pressure example, if we take a sample of size $n = 100$ from the population and use $X_1, X_2, \ldots, X_{100}$ to denote the 100 possible values obtained from this sample, we have

$$X_1, X_2, \ldots, X_{100} \sim N(125, 15^2),$$
$$\bar{X} \sim N(125, 15^2/100).$$

Figure 6.2 (left panel) shows the (unknown) theoretical probability distribution of blood pressure: $X \sim N(125, 15^2)$. The density curve in the right panel shows the probability distribution of the sample mean $\bar{X}$. The distribution of sample means is centered on the population mean (shown with the a vertical line), but its variance is much less ($n = 100$ times smaller) than that of blood pressure itself. Suppose that we could repeat this process (selecting 100 people randomly, measuring their blood pressure) many times. Each time, we obtain a different value for the sample mean. If we were to repeat this process one thousand times, we would obtain 1000 different values for the sample mean. The right panel of Fig. 6.2 shows the histogram of 1000 sample means for blood pressure. As we can see, the histogram is very similar to the sampling distribution of $\bar{X}$.

Fig. 6.3 Using
R-Commander for simulating
samples from the population,
measuring X, and calculating
$\bar{x}$. Here, we are generating
1000 samples of size $n = 100$
from $N(\mu = 125, \sigma = 15)$
distribution
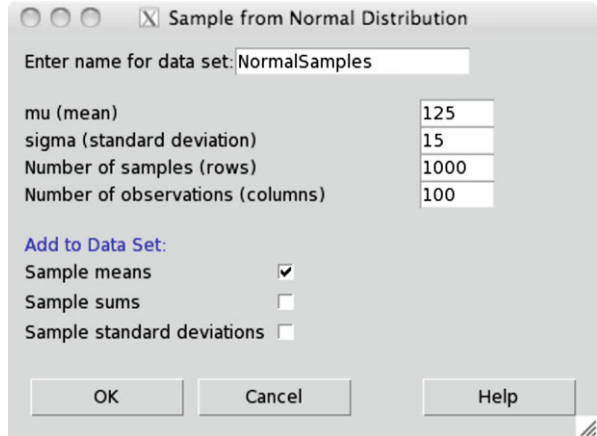
While in practice it is difficult to perform the above procedure 1000 times, we
can simulate it using R-Commander. Click `Distributions` → `Continuous`
`distributions` → `Normal distribution` → `Sample from normal`
`distribution`. Then enter 125 for the `mu (mean)`, 15 for `sigma` (stan-
dard deviation). Set the `Number of samples` to 1000 and the `Number`
`of observations` (i.e., n) to 100, as in Fig. 6.3. This creates 1000 different
samples, where the size of each sample is $n = 100$. Keep the option `Sample`
`means` checked; this will store the sample means in a variable called `mean`.

We can now plot the histogram of the 1000 sample means. (`NormalSamples`
should be the active data set.) Click `Graphs` → `Histograms`; choose `mean` as
the `Variable` and check `Densities`. The resulting histogram will be similar to
the one shown in the right panel of Fig. 6.2.

6.4 Confidence Intervals for the Population Mean

It is common to express our point estimate along with its standard deviation to show
how much the estimate could vary if different members of population were selected
as our sample. Alternatively, we can use the point estimate and its standard deviation
to express our estimate as a range (interval) of possible values for the unknown
parameter.

Consider the estimation of the population mean μ in the systolic blood pressure
example. We know that $\bar{X} \sim N(\mu, \sigma^2/n)$. Since the sampling distribution is normal,
the *68–95–99.7%* rule applies. Therefore, approximately 95% of the values of $\bar{X}$ fall
within the 2 standard deviations of the mean. We assumed that the variance of X
is $\sigma^2 = 15^2$ and sample size is $n = 100$. The standard deviation of $\bar{X}$ is therefore
$\sigma/\sqrt{n} = 1.5$. Following the *68–95–99.7%* rule, with 0.95 probability, the value of $\bar{X}$
is within 2 standard deviations from its mean, μ,

$$\mu - 2 \times 1.5 \leq \bar{X} \leq \mu + 2 \times 1.5.$$

Fig. 6.4 This graph shows
the 95% confidence intervals
obtained from 20 different
samples (indexed on the
vertical axis) each of size
$n = 100$. We expect
$19/20 = 0.95$ of these
intervals to include the true
population mean of $\mu = 125$
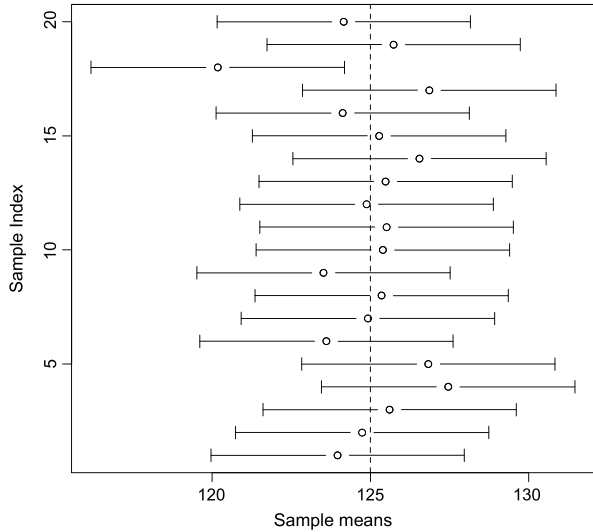
In other words, with probability 0.95,

$$\mu - 3 \leq \bar{X} \leq \mu + 3.$$

We are, however, interested in estimating the population mean μ (instead of the sample mean $\bar{X}$). By rearranging the terms of the above inequality (see Sect. 6.9), we find that with probability 0.95,

$$\bar{X} - 3 \leq \mu \leq \bar{X} + 3.$$

This means that with probability 0.95, the population mean μ is in the interval $[\bar{X} - 3, \bar{X} + 3]$.

The sample mean $\bar{X}$ is itself a random variable and changes from one sample to another. Therefore, the above interval is not fixed. With every new sample, we have a new value for $\bar{X}$, and as the result, we have a new interval. Theoretically, we could repeatedly sample $n = 100$ people, find the sample mean, and determine the interval. Then, the true population mean μ would fall within these intervals with probability 0.95.

Suppose, for example, that we repeated this process twenty times to obtain twenty such intervals, as shown in Fig. 6.4. In this figure, each sample mean is shown as a circle and the true (but unknown) population mean $\mu = 125$ as the dashed vertical line. Of twenty intervals, nineteen (i.e., 95%) include (cover) the true mean.

In reality, however, we usually have only one sample of n observations, one sample mean $\bar{x}$, and one interval $[\bar{x} - 3, \ \bar{x} + 3]$ for the population mean μ. For the blood pressure example, suppose that we have a sample of $n = 100$ people and that the sample mean is $\bar{x} = 123$. Therefore, we have one interval as follows:

$$[123 - 3, \ 123 + 3] = [120, 126].$$

We refer to this interval as our 95% **confidence interval** for the population mean μ.

In general, when the population variance σ^2 is known, the 95% confidence interval for the unknown population mean μ is obtained as follows:

$$\left[\bar{x} - 2 \times \sigma/\sqrt{n},\ \bar{x} + 2 \times \sigma/\sqrt{n} \right],$$

where $\bar{x}$ is the specific value of the sample mean (i.e., observed sample mean) we obtain based on our sample. Alternatively, we say that the **confidence level** or **confidence coefficient** for the above interval is 0.95.

Note that the above interval is only one of many possible intervals we could see. (In reality, we usually do not see more than one.) While we could assign a probability to all possible intervals based on $\bar{X}$ and say that 95% of them include the true value of the population mean, we cannot say the same thing for this specific interval based on $\bar{x}$. This specific interval is either one of the those intervals that includes the true value of the population mean, or it is one of those intervals that do not. However, we are 95% confident that it belongs to the former set of intervals and includes the true value of the population mean. The 95% confidence refers to our degree of confidence in the *procedure* that generated this interval. If we could repeat this procedure many times, 95% of intervals it creates would include the true population mean.

The multiplier 2 we used to obtain the above interval was derived from the *68–95–99.7* rule for normal distributions, which states that for a normally distributed random variable (in this case, $\bar{X}$), 95% of the observations fall within 2 standard deviations of the mean. If we want to increase our confidence level to 0.997, we use the multiplier 3 since 99.7% of observations fall within 3 standard deviations of the mean. Therefore, our 99.7% CI for the population mean is

$$\left[\bar{x} - 3 \times \sigma/\sqrt{n},\ \bar{x} + 3 \times \sigma/\sqrt{n} \right].$$

For the blood pressure example, the 99.7% CI is

$$[123 - 3 \times 1.5,\ 123 + 3 \times 1.5] = [118.5,\ 127.5].$$

Note that in this case, we have expanded the interval in order to be more confident about our estimate.

For estimates at lower confidence level of 0.68, we use the multiplier 1 instead. Our 68% CI for the population mean is

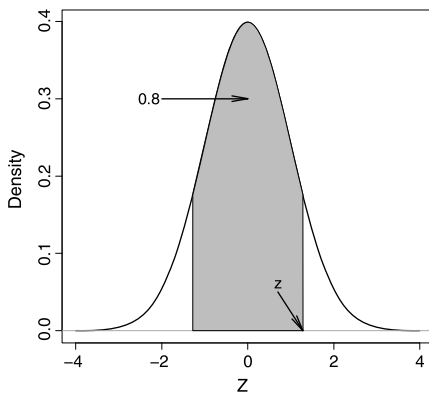$$\left[\bar{x} - \sigma/\sqrt{n},\ \bar{x} + \sigma/\sqrt{n} \right].$$

For the blood pressure example, the 68% CI is

$$[123 - 1.5,\ 123 + 1.5] = [121.5,\ 124.5].$$

Note that the length of this interval is smaller than the two previous interval estimates.

Fig. 6.5 Finding the
z-critical value for 0.8
confidence level.
Approximately 80% of the
values of the random variable
fall between -1.28 and 1.28

z-critical Values   We now discuss the process of finding the interval for any con-
fidence level other than 0.68, 0.95, and 0.997. For this, we need to find the corre-
sponding multiplier for a given confidence level. Recall that the multipliers 1, 2,
and 3 we previously used are the number of standard deviations we need to move
from the mean of a normally distributed random variable (here, $\bar{X}$) on each side to
find intervals whose probabilities are 0.68, 0.95, and 0.997, respectively. Also, recall
that these rules apply to all normal distributions regardless of the mean and standard
deviation. It is easier to work with the standard normal distribution, $N(0, 1)$. For
this distribution, the standard deviation is 1. So we can simplify the above rules as
follows. The multipliers 1, 2, and 3 are the numbers of *units* we need to move from
the mean zero on each side to find intervals whose probabilities are 0.68, 0.95, and
0.997, respectively.

Suppose that we want to set the confidence level of our interval estimate for the
population mean to 0.8. To find the corresponding multiplier, we need to find the
number of units we need to move from 0 on each side so that the probability of the
resulting interval becomes 0.8 based on the standard normal distribution. Figure 6.5
shows the probability density curve of $N(0, 1)$, which is known as the *Z-curve*. The
shaded area is 0.8, which is the probability of the corresponding interval on the x-
axis. The upper end of this interval is shown as z. Here, z is the number of units
we need to move away from 0 so that the probability of the resulting interval is
0.8. That is, z is the multiplier needed to use to obtain 80% confidence intervals for
population mean.

Since the total area under the curve is 1, the unshaded area is $1 - 0.8 = 0.2$.
Moreover, because of the symmetry of the curve around the mean, the two unshaded
areas on the left and the right of the plot are equal, which means that the unshaded
area on the right-hand side is $0.2/2 = 0.1$. Therefore, the upper-tail probability of z
is 0.1, which is equal to $(1 - 0.8)/2$.

Now we can use R-Commander to find the value of z. Click `Distribu-`
`tions` $\rightarrow$ `Continuous distributions` $\rightarrow$ `Normal distribution` $\rightarrow$
`Normal quantiles`. Enter 0.1 for the `Probabilities` and select `Upper`

tail. (The default parameters correspond to the standard normal.) The result, shown in the *Output* window, is 1.28. Therefore, we need to move $z = 1.28$ standard deviations from the mean on each side so that the probability of the resulting interval becomes 0.8. The 80% confidence interval for the for the population mean is

$$\left[\bar{x} - 1.28 \times \sigma/\sqrt{n},\ \bar{x} + 1.28 \times \sigma/\sqrt{n}\,\right].$$

For the systolic blood pressure example, where $\bar{x} = 123$ and $\sigma/\sqrt{n} = 1.5$, we are 80% confident that the true mean blood pressure is in the interval

$$[123 - 1.28 \times 1.5, 123 + 1.28 \times 1.5] = [122.8, 123.2].$$

We call the multiplier 1.28 the z-critical value, denoted as z_{crit}, for the 80% confidence interval. We can follow similar steps to find the z-critical values for any other confidence level. For 0.9 confidence level, for example, $z_{\text{crit}} = 1.64$. For 0.95 confidence level, so far we have been using $z_{\text{crit}} = 2$. Following the above steps, you will find that a more accurate value is $z_{\text{crit}} = 1.96$, which is sometimes used instead of 2 to be more precise.

In general, for a given confidence level, c, we use the standard normal distribution to find the value whose upper tail probability is $(1 - c)/2$. We refer to this value as the z-critical value for the confidence level of c. Then with the point estimate $\bar{x}$, the confidence interval for the population mean at c confidence level is

$$\left[\bar{x} - z_{\text{crit}} \times \sigma/\sqrt{n},\ \bar{x} + z_{\text{crit}} \times \sigma/\sqrt{n}\,\right].$$

6.5 Confidence Interval When the Population Variance Is Unknown

So far, we have assumed the population variance, σ^2, of the random variable is known. Hence, we assumed that $\sigma/\sqrt{n}$, i.e., the standard deviation of the sample mean, is known. This is an unrealistic assumption. Almost always, we need to estimate σ^2 along with the population mean μ. For this, we use our sample of n observations to obtain the sample variance s^2 and sample standard deviation s. As a result, the standard deviation for $\bar{X}$ is estimated to be $s/\sqrt{n}$. We refer to $s/\sqrt{n}$ as the **standard error** of the sample mean $\bar{X}$ to distinguish it from $\sigma/\sqrt{n}$. In general, we refer to the standard deviation of an estimator (e.g., $\bar{X}$) as its standard error if we have to use the data to estimate it. We use SE to denote the standard error of an estimator.

To find confidence intervals for the population mean when the population variance is unknown, we follow similar steps as described above, but instead of $\sigma/\sqrt{n}$ we use $SE = s/\sqrt{n}$, and instead of z_{crit} based on the standard normal distribution,

we use t_{crit} obtained from a t-distribution with $n - 1$ degrees of freedom. The confidence interval for the population mean at c confidence level is

$$\left[\bar{x} - t_{crit} \times s/\sqrt{n}, \ \bar{x} + t_{crit} \times s/\sqrt{n}\,\right],$$

where t_{crit} is the value with an upper tail probability of $(1 - c)/2$ based on a t-distribution with $n - 1$ degrees of freedom.

For example, suppose that we have randomly selected seven newborn babies and recorded their heights (in inches) at the time of birth as follows:

$$\text{Height:} \quad 18, 22, 19, 17, 20, 18, 15.$$

We use X to denote the height of newborn babies and assume that $X \sim N(\mu, \sigma^2)$. Based on the above observed data, the point estimates for μ and σ are $\bar{x} = 18.4$ and $s = 2.2$, respectively. The standard error (estimated standard deviation) for the sample mean is $SE = 2.2/\sqrt{7} = 0.83$.

Suppose that we want to find the 90% confidence interval for the population mean, μ. Then, using the t-distribution with $7 - 1 = 6$ degrees of freedom, we need to find the t-critical value, t_{crit}, whose upper tail probability is $(1 - 0.9)/2 = 0.05$.

In R-Commander, click `Distributions` → `Continuous distributions` → `t distribution` → `t quantiles`. Set the `Probabilities` to 0.05, the `Degrees of Freedom` to 6, and check `Upper tail` option. The result, shown in *Output* window, is $t_{crit} = 1.94$, which is greater than $z_{crit} = 1.64$ based on the standard normal distribution. The 90% CI, therefore, is

$$\left[18.4 - 1.94 \times \frac{2.2}{\sqrt{7}}, \ 18.4 + 1.94 \times \frac{2.2}{\sqrt{7}}\right] = [16.8, \ 20.0].$$

That is, at 0.9 confidence level, we estimate the mean of height for newborn babies to be between 16.8 and 20.0 inches.

In this example, if we knew $\sigma = 2.2$ instead of estimating it to be $s = 2.2$, we would have used $z_{crit} = 1.64$ instead of $t_{crit} = 1.94$, and the interval would have been smaller. Everything else the same, using t-distribution instead of the standard normal leads to wider intervals. This is the price we pay for the additional uncertainty due to the estimation of population variance (and standard deviation) from the data.

As discussed previously, the t-distribution approaches the standard normal distribution as the sample size increases (i.e., the degree of freedom increase). Therefore, the difference between the z-critical values and the t-critical values becomes negligible for very large sample sizes.

6.6 Using Central Limit Theorem for Confidence Interval

So far, we have assumed that the random variable has normal distribution, so the sampling distribution of $\bar{X}$ is normal too. If the random variable is not normally distributed, the sampling distribution of $\bar{X}$ is still *approximately* normal as long as the sample size is large. The larger the sample size, the better the approximation. This concept is known as the **central limit theorem** (CLT) in statistics.

For large sample sizes, the CLT indicates that (under certain conditions such as independence of observations) if the random variable X has the population mean μ and the population variance σ^2, then the sampling distribution of $\bar{X}$ is approximately normal with mean μ and variance σ^2/n:

$$\bar{X} \sim N\left(\mu,\ \sigma^2/n\right).$$

As before, the mean of the sampling distribution of $\bar{X}$ is the population mean μ of the random variable, and its variance is the population variance σ^2 of the random variable divided by the sample size n

The CLT is applicable regardless of the random variable's distribution. Therefore, even if the random variable has other distributions such as Bernoulli, binomial, or Poisson, the sampling distribution of its mean will be approximately normal and will be centered on the true population mean if the sample size is large.

For example, suppose that we are investigating the number of physician visits per year. Further, suppose that the true but unknown population mean (rate) is $\mu = 2.5$. For illustrative purposes, we assume that the random variable has a Poisson(2.5) distribution. The pmf of this distribution is shown in the left panel of Fig. 6.6. Recall that the variance of a Poisson distribution is the same as its mean. Therefore, the theoretical variance of the random variable is $\sigma^2 = 2.5$. To estimate μ, we can randomly select 200 people, record the number of times they have visited their physicians last year, and calculate the average number of visits. If we repeat this process a thousand times, the distribution of the sample means will be similar to Fig. 6.6. (Follow the above steps for simulating data in R-Commander, but this time
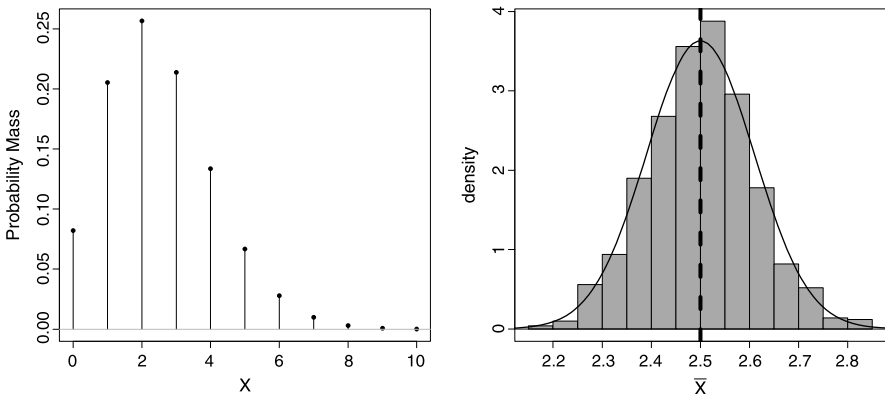
Fig. 6.6 *Left panel*: Plot of the pmf for a Poisson(2.5) distribution. *Right panel*: Histogram of sample means for the number of physician visits per year. This distribution was generated from 1000 groups each with 200 people. While the distribution of the random variable itself is not normal, the sampling distribution of the mean is approximately normal and is centered on the population mean 2.5 (shown with the *dashed vertical line*)

Fig. 6.7 *Left panel*: Plot of the pmf for a Bernoulli(0.8) distribution. *Right panel*: Histogram of sample mean (proportion) for survival of breast cancer patients. This distribution was generated from 1000 groups each with 220 people. While the distribution of the random variable itself is not normal, the sampling distribution of the mean is approximately normal and is centered on the population mean 0.8 (shown with the *dashed vertical line*)
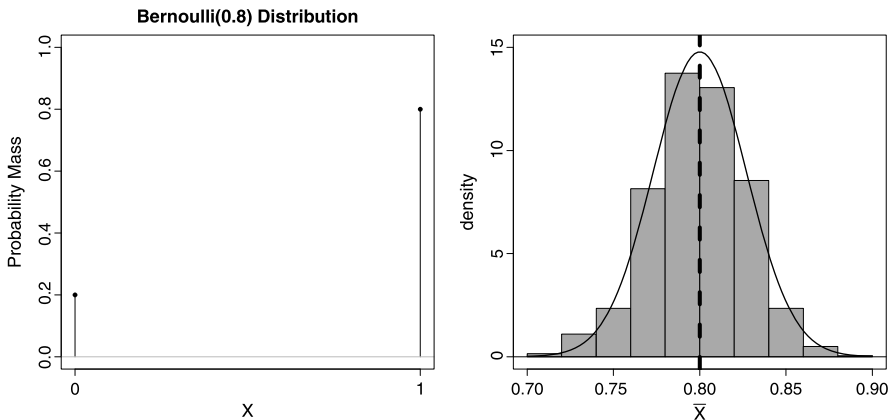
use `Poisson` distribution instead of normal and set the number of observations to 200.)

While the distribution of the random variable itself is not normal, the sampling distribution of the sample mean $\bar{X}$ is approximately normal and is centered on the population mean $\mu = 2.5$. Since the variance of the random variable is 2.5, the standard deviation of $\bar{X}$ in this case is $\sqrt{2.5/200} = 0.11$. The right panel in Fig. 6.6 shows the density curve of $N(2.5, 0.11^2)$ along with the histogram of 1000 sample means, each based on 200 people. As we can see, the distribution presented by the histogram is closely approximated by the normal distribution.

As the second example, suppose that the 5-year survival status, X, of breast cancer patients has Bernoulli(0.8) distribution. That is, the probability of survival ($X = 1$) is 0.8. Recall that the population variance for Bernoulli distributed random variables is $\mu(1 - \mu)$. In this case, the population variance is $\sigma^2 = 0.8 \times (1 - 0.8) = 0.16$. The left panel in Fig. 6.7 shows the plot of the pmf for the Bernoulli(0.8) distribution. Now suppose that we take a sample of $n = 220$ from the population to obtain the sample mean, $\bar{X}$. The right panel shows approximate normal distribution for $\bar{X}$ with mean 0.8 (i.e., the population mean) and variance $0.16/220 = 0.0007$. The right panel also shows the histogram of 1000 sample means (i.e., sample proportions) for survival of breast cancer patients. To obtain each sample mean, 220 patients were randomly selected and the proportion of people who survived within five years was calculated. (We used R-Commander to simulate data.) While the distribution of the random variable itself is not normal, the sampling distribution of the mean is approximately normal and is centered on the population mean (shown with the dashed vertical line).

We can use the central limit theorem to find confidence intervals for the population mean of a random variable without assuming that it is normally distributed. In

this case, since the sample mean is still approximately normal when the sample size is large (e.g., at least 15 if there are no outliers and the distribution is roughly symmetric), we can still use the methods we discussed above to find confidence intervals for the population mean. For example, we want to estimate the population mean of BMI (denoted as X) among Pima Indian women without making any assumption about the probability distribution of X. Using a data set with $n = 200$ observations (available from the MASS package), our point estimates for the distribution parameters are $\hat{\mu} = \bar{x} = 32.3$ and $\hat{\sigma}^2 = s^2 = 6.1^2$. We are interested in 90% confidence interval for μ.

The standard error for the sample mean is $6.1/\sqrt{200} = 0.43$, and the t-critical value for 0.9 confidence level (obtained from a t distribution with $n - 1 = 199$ degrees of freedom) is 1.65. Therefore, the 90% CI for μ is

$$[32.3 - 1.65 \times 0.43, 32.3 + 1.65 \times 0.43] = [31.6, 33.0],$$

which means that at 0.9 confidence level, the mean of the distribution (population mean of the BMI) falls between 31.6 and 33.0.

Note that when the sample size is large, the t distribution reaches the standard normal distribution, and t-critical values become very close to z-critical values. In the above example, $t_{\text{crit}} = 1.65$ is almost the same as $z_{\text{crit}} = 1.64$ for 0.9 confidence level.

6.7 Confidence Intervals for the Population Proportion

Suppose that we want to find the 95% CI for the population proportion of mothers who smoke during their pregnancy in the year 1986. Using the birthwt data set with $n = 189$, our estimate for this proportion is $\bar{x} = p = 0.39$. (Note that the data are collected during 1986.) Using p, we estimate the population variance $p(1 - p) = 0.39 \times 0.61 = 0.24$.

> For binary random variables, we use the sample proportion to estimate the population proportion as well as the population variance. That is, the sample variance depends on the data through p and n only. Therefore, estimating the population variance does not introduce an additional source of uncertainty to our analysis, so we do not need to use a t-distribution instead of the standard normal distribution.

The standard error (i.e., estimated standard deviation) for the sample mean is

$$SE = \sqrt{p(1 - p)/n} = \sqrt{0.39 \times 0.61/189} = 0.03.$$

The 95% CI is then

$$[p - z_{\text{crit}} \times SE, \, p + z_{\text{crit}} \times SE].$$

For 0.95 confidence level, $z_{\text{crit}} = 1.96$, which we usually round off to 2. Therefore, the 95% CI for the population proportion is

$$[0.39 - 2 \times 0.03, 0.39 + 2 \times 0.03] = [0.33, 0.45],$$

which means that we are 95% confident that the true population proportion is between 0.33 and 0.45. (The current estimate provided by CDC is much lower than this.)

From the above confidence interval, we can find the confidence interval for the number of smoking pregnant women in the US during 1986. As before, we suppose that there are currently $N = 4$ million pregnant women in the US. We find the 95% confidence interval for the number of smoking pregnant women as follows:

$$[0.33 \times 4000000, 0.45 \times 4000000] = [1320000, 1800000].$$

6.8 Margin of Error

For the above example, we can write the 95% CI for the population proportion of women who smoke during their pregnancy as follows:

$$0.39 \pm 2 \times 0.03.$$

In this case, the term $2 \times SE = 2 \times 0.03 = 0.06$ is called the **margin of error** for 0.95 confidence level. In general, it is common to present interval estimates for c confidence level as

$$\text{Point estimate} \pm \text{Margin of error}.$$

That is, the margin of error of an estimate is the half-width of the confidence interval. For the smoking during pregnancy example, our interval estimate can be written as

$$0.39 \pm 0.06.$$

When the population variance σ^2 is known, the margin of error e is calculated as follows:

$$e = z_{\text{crit}} \frac{\sigma}{\sqrt{n}}.$$

Here, z_{crit} is the multiplier obtained for the given confidence level c from the standard normal distribution. When the population variance is not known and we need to use the data to estimate it using the sample standard deviation, s, the margin of error is calculated as follows:

$$e = t_{\text{crit}} \frac{s}{\sqrt{n}}.$$

Here, t_{crit} is the multiplier obtained for the given confidence level c from the t distribution with $n-1$ degrees of freedom.

If the variable is binary so we can use the sample proportion p to estimate $\hat{\sigma} = \sqrt{p(1-p)}$, then the margin of error is

$$e = z_{crit}\sqrt{\frac{p(1-p)}{n}}.$$

Often, the media reports a margin of error to accompany the results of a poll. Generally, what they report is the margin of error for a 95% CI, although they do not specify that.

6.9 Advanced

In this section, we show the derivation of confidence interval and discuss finding the required sample size for a given margin of error.

6.9.1 Deriving Confidence Intervals

In Sect. 6.4, we derived confidence intervals for the sample mean based on the Central Limit Theorem and the *68–95–99.7%* rule for normal distributions. Specifically, we saw that the sampling distribution of $\bar{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Using the *68–95–99.7%* rule, we know that with probability 0.95, $\bar{X}$ falls within 2 standard deviations from its mean,

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

(Note that since the normal distribution is continuous, using $\leq$ instead of $<$, which is what we usually use for the lower end of the interval, doses not change the probability.) Therefore, 95% of the values of $\bar{X}$ fall in the interval

$$\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}.$$

To produce a range of possible values for the population mean μ, we subtract μ from all three terms of the inequalities:

$$\mu - 2\frac{\sigma}{\sqrt{n}} - \mu \leq \bar{X} - \mu \leq \mu + 2\frac{\sigma}{\sqrt{n}} - \mu,$$

$$-2\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 2\frac{\sigma}{\sqrt{n}}.$$

Now, subtract $\bar{X}$ from all terms:

$$-\bar{X} - 2\frac{\sigma}{\sqrt{n}} \le -\bar{X} + \bar{X} - \mu \le -\bar{X} + 2\frac{\sigma}{\sqrt{n}},$$

$$-\bar{X} - 2\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X} + 2\frac{\sigma}{\sqrt{n}}.$$

Lastly, multiply all three terms by -1 (multiplying by -1 changes the directions of inequalities) to obtain the interval for the population mean μ:

$$\bar{X} + 2\frac{\sigma}{\sqrt{n}} \ge \mu \ge \bar{X} - 2\frac{\sigma}{\sqrt{n}},$$

$$\bar{X} - 2\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 2\frac{\sigma}{\sqrt{n}}.$$

Therefore, the true value of μ falls within the following interval with the probability of 0.95:

$$\left[\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right].$$

Given a specific value $\bar{x}$ for the sample mean $\bar{X}$, we can construct the 95% confidence interval as follows:

$$\left[\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right].$$

For the above confidence interval, 0.95 represents our confidence level in the procedure that produced this interval.

6.9.2 Sample Size Estimation

Suppose that we want to estimate the population mean. Further, suppose that we have an acceptable margin of error e in mind and want to find the required sample size so that the margin of error for our estimate is e. In this situation, since empirical data is not yet available, we should either know what σ is or make a conservative guess. For latter, if the variable is numerical, we can guess its range. Then, assuming that the distribution is approximately normal, we know that 4 standard deviations (2 standard deviations on each side of the mean) includes the values for 95% of the population. Therefore, we use *range*/4 as a rough estimate for σ. If the variable is binary, we can use the standard deviation for $\mu = 0.5$, which is $\sigma = \sqrt{0.5(1 - 0.5)} = 0.5$, as our guess. This is a quite conservative guess since 0.5 is the highest possible standard deviation for a binary random variable. (Try using other values for μ to see that this is true; remember that μ must be between 0 and 1.)

Using the following equation for the margin of error:

$$e = z_{\text{crit}}\frac{\sigma}{\sqrt{n}},$$

we can estimate the required sample size n for the assumed acceptable margin of error e as follows:

$$n = \left(\frac{z_{crit}\sigma}{e} \right)^2 .$$

For example, let us find the appropriate sample size to estimate population mean for BMI. Suppose that we decide that the acceptable margin of error at confidence level 0.95 is 3. That is, we want to be confident (at 0.95 level) that the true population mean falls within 3 units from its point estimate. We want the population mean) Further, suppose that, based on previous experience, we know that the BMI is roughly between 10 to 50. Therefore, we assume that σ is approximately $(50 - 10)/4 = 10$. Then the required sample size is

$$n = \left(\frac{2 \times 10}{3} \right)^2 \approx 45.$$

Therefore, we need to measure the BMI of 45 people.

As another example, suppose that we want to test a new drug for breast cancer, and we would like to estimate the 5-year survival mean (proportion) with the margin of error of 0.1 at 0.8 confidence level. (We want the true survival rate to fall within 10% from its point estimate.) Using R-Commander, the z-critical value for 0.8 confidence level is 1.28. Then we need

$$n = \left(\frac{1.28 \times .5}{0.1} \right)^2 = 41.$$

Note that we set σ to 0.5 to be conservative since 0.5 is the highest possible standard deviation for binary random variable, and the above equation results in the highest value of n for the given margin of error and confidence level. Therefore, we need to test the drug on 41 people to achieve the required margin of error at the given confidence level.

6.10 Exercises

1. We assume that the probability distribution of blood pressure, X, is $N(\mu, \sigma^2)$ distribution. Suppose we know that $\sigma = 6$. To estimate μ, we randomly selected 9 people and measured their blood pressure. The sample mean is $\bar{x} = 110$.
   (a) Write down the sampling distribution of the sample mean $\bar{X}$ and find its standard deviation.
   (b) Find the 80% confidence interval estimation for μ.
2. For the above question, suppose that we did not know σ and estimated it using the sample standard deviation $s = 6$.

(a) Find the standard error for the sample mean as the estimator of the population mean.

(b) Find the 80% confidence interval estimation for μ based on this sample.

3. Using the `birthwt` data set, find the point estimate and the 85% confidence interval estimate for the population proportion of babies with low birthweight and the population proportion of mothers who have hypertension history.

4. Using the `birthwt` data set, find the point estimate and the 90% confidence interval estimate for the population mean of the number of physician visits during the first trimester.

5. Using the "BodyTemperature.txt" data set, find the point estimate and the 80% confidence interval estimate for the population means of heart rate and normal body temperature.

6. Suppose that we want to estimate the population mean of birthweight. The acceptable margin error at 0.9 confidence level is 0.5 pounds. If the range of birthweight is from 2 pounds to 11 pounds, what is the required sample size?

7. We would like to estimate the proportion of people who smoke regularly. For this, we decide to interview a sample of people from the population. If the accepted margin of error at 0.95 confidence level is 0.02, how many people should we interview?

8. Suppose that we interviewed a random sample of 2000 people and found that 320 of them smoke regularly. Find the 90% confidence interval for the population proportion of smokers.

9. Read the paper "A Critical Appraisal of 98.6°F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich" by Mackowiak et al. [19]. (The paper is available online at http://jama.ama-assn.org/cgi/reprint/268/12/1578.) What is their point estimate along with its margin of error for the population mean of normal body temperature?

Chapter 7
Hypothesis Testing

7.1 Introduction

In the previous chapter, we focused on estimating parameters such as the population mean and variance. In this chapter, we rely on estimators, their sampling distributions, and their specific values from observed data to evaluate **hypotheses**.

In general, many scientific investigations start by expressing a hypothesis. For example, Mackowiak et al. [19] hypothesized that the average normal (i.e., for healthy people) body temperature is less than the widely accepted value of 98.6°F. If we denote the population mean of normal body temperature as μ, then we can express this hypothesis as $\mu < 98.6$.

When we state our hypothesis, we are mainly proposing an explanation for an observed phenomenon. For the above example, we might have observed that the body temperature of many healthy people is less than 98.6°F. Typically, we can find another explanation, also expressed as a hypothesis, that invalidates (annuls) our proposed hypothesis. For this example, one might hypothesize that $\mu \geq 98.6$. We refer to this hypothesis as the **null hypothesis** and denote it as H_0. The null hypothesis usually reflects the "status quo" or "nothing of interest". In contrast, we refer to our hypothesis (i.e., the hypothesis we are investigating through a scientific study) as the **alternative hypothesis** and denote it as H_A.

It is common to express the null hypothesis in the simplest form possible. For the above example, to annul the alternative hypothesis, $H_A : \mu < 98.6$, it suffices to show that $H_0 : \mu = 98.6$. This makes the task of evaluating a hypothesis easier.

The procedure for evaluating a hypothesis is called **hypothesis testing**, and it rises in many scientific problems. A common approach for hypothesis testing is to focus on the null hypothesis, which is usually simpler than the alternative hypothesis, and decide whether or not to reject it. To this end, we examine the evidence that the observed data provide against the null hypothesis H_0. If the evidence against H_0 is strong, we reject H_0. If not, we state that the evidence provided by the data is not strong enough to reject H_0, and we fail to reject it.

With respect to our decision regarding the null hypothesis H_0, we might make two types of errors:

- Type I error: we reject H_0 when it is true and should not be rejected.
- Type II error: we fail to reject H_0 when it is false and should be rejected.

We denote the probability of making type I error as α and the probability of making type II error as β.

We hope to avoid both type I and type II errors as much as possible. However, there is a trade-off between them. To minimize α (the probability of making type I error), we might be tempted not to reject H_0 unless there is extremely strong evidence against it. This, however, increases the probability β of failing to reject H_0 when it is false. Likewise, to reduce β (the probability of making type II error), we might be tempted to reject H_0 based on weak evidence against it. This, of course, increases the probability α of rejecting H_0 by mistake.

Now suppose that we have a hypothesis testing procedure that fails to reject the null hypothesis when it should be rejected with probability β. This means that our test correctly rejects the null hypothesis with probability $1 - \beta$. (Note that the two events are complementary.) We refer to this probability (i.e., $1 - \beta$) as the **power** of the test. In practice, it is common to first agree on a tolerable type I error rate α, such as 0.01, 0.05, and 0.1. Then try to find a test procedure with the highest power among all reasonable testing procedures.

In this chapter, we discuss some commonly used testing procedures when the hypothesis is related to the population mean, μ, without explicit discussion of type I and type II errors. Throughout this chapter, we follow similar assumptions we used for estimation in the previous chapter; namely, we assume that $X_1, \ldots, X_n$ (i.e., values of the random variable obtained from a random sample of size n) are IID (unless stated otherwise) and that the sample size n is large enough for the CLT to hold.

7.2   Hypothesis Testing for the Population Mean

To decide whether we should reject the null hypothesis, we quantify the empirical support (provided by the observed data) against the null hypothesis using some statistics. Recall that a statistic is what we calculate based on the observed data only (i.e., it does not depend on any unknown parameter). Since these statistics are used to evaluate our hypotheses, we refer to them as **test statistics**. To evaluate hypotheses regarding the population mean, we use the sample mean $\bar{X}$ as the test statistic.

For a statistic to be considered as a test statistic, its sampling distribution must be fully known (exactly or approximately) under the null hypothesis. That is, we should know the distribution of the test statistic if we assume that the null hypothesis is true. For the sample mean, the CLT states that the sampling distribution is approximately

normal when the sample size is large. (The distribution is exactly normal if the variable itself is normal and the population variance is known.)

Now consider the body temperature example, where we want to examine the null hypothesis $H_0 : \mu = 98.6$ against the alternative hypothesis $H_A : \mu < 98.6$. To start, suppose that $\sigma^2 = 1$ is known. (Later, we will discuss situations where σ^2 is unknown.) Further, suppose that we have randomly selected a sample of 25 healthy people from the population and measured their body temperature.

Using the Central Limit Theorem, the sampling distribution of $\bar{X}$ is approximately normal as follows:

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

In this case,

$$\bar{X} \sim N(\mu, 1/25).$$

Now, suppose that the null hypothesis is true and the population mean is $\mu = 98.6$. By setting μ to 98.6, the sampling distribution of $\bar{X}$ becomes

$$\bar{X}|H_0 \sim N(98.6, 0.04).$$

Note that the distribution of $\bar{X}$ is obtained conditional (hence the notation for conditional probability) on the assumption that the null hypothesis is true. The distribution is fully specified if we assume that H_0 is true. Therefore, we can use the sample mean $\bar{X}$ as a test statistic for the population mean μ.

In what follows, we refer to the distribution of test statistics under the null hypothesis as the **null distribution**. For the above example, the null distributions is $N(98.6, 0.04)$. Use R-Commander to plot this distribution. (Note that you need to enter the standard deviation instead of the variance in R-Commander). The left panel in Fig. 7.1 shows this distribution.
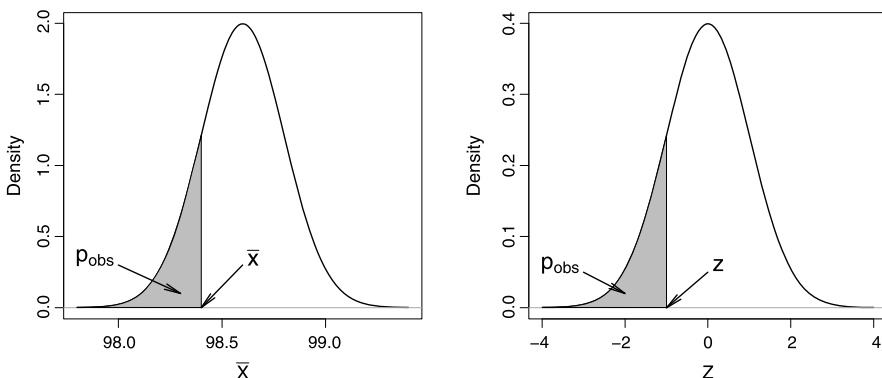

Fig. 7.1 For the normal body temperature example, we are examining the hypotheses $H_0 : \mu = 98.6$ against $H_A : \mu < 98.6$. *Left panel*: The *shaded area* shows the lower-tail probability of the observed sample mean, $\bar{x} = 98.4$. This is the observed significance level, p-value, which is denoted as p_{obs}. *Right panel*: After standardizing, the p-value corresponds to the lower tail probability of $z = -1$ based on the standard normal distribution

7.3 Statistical Significance

The test statistic $\bar{X}$ is itself a random variable, so its value can change every time we take a new sample of size n from the population. However, if the null hypothesis is indeed true, then we would expect to see these values to be close to the mean of the null distribution (here, 98.6). In contrast, if the null hypothesis is false, then the null distribution does not represent the sampling distribution of the test statistics, and we would expect to see the values of $\bar{X}$ to be far from the mean of the null distribution. In reality, we have only one value, $\bar{x}$, for the sample mean. We can use this value to quantify the evidence of departure from the null hypothesis. The further $\bar{x}$ is from the value stated by the null, the stronger the evidence against it.

Suppose that from our sample of 25 people we find that the sample mean is $\bar{x} = 98.4$. A very common method to measure the amount of evidence for the departure from the null hypothesis $H_0 : \mu = 98.6$ versus the alternative $H_A : \mu < 98.6$ is the lower tail probability of this value from the null distribution. This probability is highlighted in the left panel of Fig. 7.1. Note that as the observed sample mean moves away from 98.6 (e.g., $\bar{x} = 98.3$), the lower tail probability decreases. For values of the sample mean far away from 98.6 (the population mean according to H_0) we would be more inclined to reject the null hypothesis. In contrast, as the observed sample mean moves closer to 98.6 (e.g., $\bar{x} = 98.5$), the lower tail probability increases. In this case, we would be more reluctant to reject the null hypothesis. That is, values close to 98.6 make the null hypothesis that $\mu = 98.6$ more believable.

For the observed sample mean $\bar{x} = 98.4$, the lower tail probability is the probability of observing values equal to or less than 98.4. The values less than 98.4 provide more evidence, compared to 98.4, *against* the null hypothesis, and are considered more extreme than 98.4 if we were to believe the null hypothesis. Therefore, the lower tail probability at 98.4 is the probability of observing values as or more extreme than 98.4. We refer to this probability as the **observed significance level** for the test statistic.

> The observed significance level for a test is the probability of values as or more extreme than the observed value, based on the null distribution (i.e., the sampling distribution of the test statistic assuming the null hypothesis is true), in the direction supporting the alternative hypothesis. This probability is also called the p-**value** and denoted p_{obs}.

For the above example

$$p_{obs} = P(\bar{X} \leq \bar{x}|H_0),$$

since the alternative is $H_A : \mu < 98.6$. That is, the direction of the alternative hypothesis is towards values smaller than what is specified by H_0. Note that the probability is found conditional on the assumption that the null hypothesis is true, hence the conditional probability notation. In what follows, we drop H_0 for simplicity, but we should always remember that *p-value is obtained conditional on* H_0.
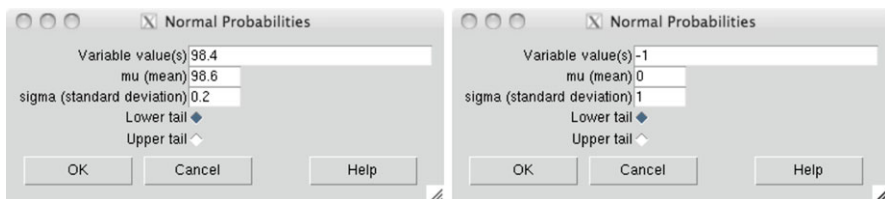
Fig. 7.2 *Left panel*: Obtaining the lower tail probability $P(\bar{X} \leq 98.4)$ where the null distribution is $N(98.6, 0.2)$. The resulting probability is the observed significance level for testing the hypothesis about the population mean of body temperature: $H_0 : \mu = 98.6$. *Right panel*: Obtaining the p-value using the z-test using the standard normal, $N(0, 1)$, distribution. Here, the z-score (i.e., the standardized value of $\bar{x} = 98.4$) is -1

We can regard p_{obs} as a measure of agreement between the observed data and the null hypothesis. As p_{obs} becomes smaller, we would be more confident to reject the null hypothesis.

Given the observed sample mean $\bar{x} = 98.4$, we calculated the p-value for the above example as follows:

$$p_{obs} = P(\bar{X} \leq 98.4).$$

To find the p-value in R-Commander, click `Distributions` → `Continuous distributions` → `Normal distribution` → `Normal probabilities`. Then set the `Variable value` to 98.4 and the parameters for the null distribution ($\mu = 98.6$ and $\sigma = \sqrt{0.04} = 0.2$), as in the left panel of Fig. 7.2. Make sure the option `lower tail` is selected since we are interested in $P(\bar{X} \leq 98.4)$. The result, given in the *Output* window, is the probability of seeing as or more extreme values than $\bar{x} = 98.4$ (in the lower direction) under the null hypothesis: $p_{obs} = 0.16$.

7.3.1 z-Tests of the Population Mean

In practice, it is more common to use the standardized version of the sample mean as our test statistic. For the body temperature example, we standardize the test statistic $\bar{X}$ by subtracting the mean $\mu = 98.6$ (under the null) and dividing the result by the standard deviation $\sqrt{0.04} = 0.2$. We denote the resulting random variable Z:

$$Z = \frac{\bar{X} - 98.6}{0.2}.$$

In Chap. 5, we saw that if a random variable is normally distributed (as it is the case for $\bar{X}$), subtracting the mean and dividing by standard deviation creates a new random variable with standard normal distribution. Therefore,

$$Z \sim N(0, 1).$$

This way, the null distribution becomes the standard normal distribution.

The observed value of $\bar{X}$ in the body temperature example was $\bar{x} = 98.4$. To find the corresponding value for the random variable Z, we standardize $\bar{x}$ in the same way we standardized $\bar{X}$. We denote this value as z:

$$z = \frac{98.4 - 98.6}{0.2} = -1.$$

Now, instead of finding the p-value based on the lower tail probability of $\bar{x} = 98.4$, we can find it based on the lower tail probability of $z = -1$:

$$p_{\text{obs}} = P(Z \le -1).$$

This probability is shown as the shaded area in the right panel of Fig. 7.1. The p-value obtained based on the standardized test statistic Z is exactly the same as the p-value obtained based on the original test statistic $\bar{X}$. That is, the shaded areas in the left and right panels of Fig. 7.1 are the same. To see this, we can start by the definition of p-value based on $\bar{X}$ and show that it is equivalent to its definition based on Z. For the above example,

$$p_{\text{obs}} = P(\bar{X} \le 98.4).$$

We can subtract 98.6 from both sides of the inequality and divide the results by 0.2:

$$p_{\text{obs}} = P\left(\frac{\bar{X} - 98.6}{0.2} \le \frac{98.4 - 98.6}{0.2}\right) = P(Z \le -1).$$

We can use R-Commander to find the p-value based on Z. See the right panel of Fig. 7.2. As expected, the result would be the same as before: $p_{\text{obs}} = 0.16$. Therefore, using either the unstandardized version or the standardized version, we will reach the same conclusion.

> We refer to the standardized value of the observed test statistic as the z-**score** and the corresponding hypothesis test of the population mean as the z-**test**, or more specifically, **single-sample** z-**test**. In a z-test, instead of comparing the observed sample mean $\bar{x}$ to the population mean according to the null hypothesis, we compare the z-score to 0. Therefore, the p-value becomes the probability of seeing values as extreme or more extreme than the observed z-score under the standard normal distribution.

One advantage of using the z-test is that the null distribution remains the same for different tests (e.g., different null values and different σ^2), and we can easily compare z-scores from two different tests. Using the z-test to find p-values was crucial when computer programs such as R-Commander were not widely available, and statisticians needed to use probability tables to calculate probabilities.

7.3.2 Interpretation of p-value

The p-value is the conditional probability of extreme values (as or more extreme than what has been observed) of the test statistic assuming that the null hypothesis

is true. When the p-value is small, say 0.01 for example, it is rare to find values as extreme as what we have observed (or more so). This means that the observed value of the test statistic is quite extreme if we were to believe the null hypothesis. As the p-value increases, it indicates that there is a good chance to find more extreme values (for the test statistic) than what has been observed. Then, the observed value does not seem that extreme any more. In this case, we think it is quite reasonable to believe that what we have observed was generated according to the null hypothesis, so we would be more reluctant to reject the null hypothesis.

Based on the above description of the p-value, we can interpret it as a measure of agreement between the observed data and the null hypothesis. Smaller p-values mean less agreement and provide stronger evidence against the null hypothesis.

A common **mistake** is to regard the p-value as the probability of the null hypothesis given the observed value of the test statistic: $P(H_0|\bar{x})$. This is because intuitively it makes more sense to evaluate the null hypothesis by finding its probability given the data we have observed. However, this is not what the p-value provides.

In order to use the p-value to decide whether we should reject the null hypothesis, a convenient approach is to prespecify a cutoff for the p-value and reject the null hypothesis if p_{obs} is below the cutoff (i.e., when the measure of agreement between the null hypothesis and observed data is less than an acceptable level). This cutoff is called the **significance level** or the **size** of the test. This is the acceptable type I error probability, i.e., the probability of rejecting the null hypothesis when it is true. As mentioned above, we denote this probability α. The common significance levels are 0.01, 0.05, and 0.1. If p_{obs} is less than the assumed cutoff, we say that the data provides **statistically significant** evidence against H_0, and we call the results statistically significant; that is, the difference between the observed value of the test statistic (here, 98.4) and the value specified by the null hypothesis (here, 98.6) is statistically significant. When we find the observed difference between the sample mean and the population mean according to the null as statistically significant, we believe that it is unlikely that the difference is due to chance alone. In practice, it is common to interpret p-values close to 0.1 as small amount of evidence against H_0, p-values around 0.05 as modest evidence against H_0, and p-values below 0.01 as strong evidence against H_0.

Note that while the above approach provides a convenient framework for testing a hypothesis, one should be always cautious against relying on such significance tests as the only tool for making decisions regarding acceptance or rejection of a hypothesis [5].

For the body temperature example where $p_{obs} = 0.16$, if we set the significance level at 0.05, we say that there is not significant evidence against the null hypothesis

$H_0 : \mu = 98.6$ at the 0.05 significance level, so we do not reject the null hypothesis. It is clear that when we *cannot* reject the null hypothesis at 0.05, we cannot reject it at 0.01 or any other lower significance level since for lower significance levels, it becomes even harder to reject H_0 (i.e., we need stronger evidence against it). On the other hand, when we *can* reject the null hypothesis at 0.05 level, we can reject it at higher significance levels such as 0.1.

> When there is not significant evidence to reject H_0 at any acceptable level, we fail to reject the null hypothesis. This could happen because the null hypothesis is in fact true. However, it is also possible that the null hypothesis is false but we do not have enough evidence to reject it. For example, this could happen when the sample size is too small. When we fail to reject the null hypothesis, the result of our test is *inconclusive* since we do not know which one of the above two possible scenarios is true.

In the above example, if we had set the cutoff at 0.05 and we had observed $\bar{x} = 98.25$ instead of 98.4, then $p_{\text{obs}} = 0.04$, and we could reject the null hypothesis. In this case, we say that the result is statistically significant and the data provide enough evidence against $H_0 : \mu = 98.6$. However, if we had set the cutoff at 0.01, we would fail to reject the null hypothesis and say that the result is not statistically significant. One simple solution to resolve this issue, i.e., profoundly different conclusion based on arbitrary cutoff, is to choose several reference levels (e.g., 0.01, 0.05, 0.1) as opposed to one and comment on the significance of the data with respect to these reference levels [5]. In this example, when $\bar{x} = 98.25$ and $p_{\text{obs}} = 0.04$, we could say that the amount of evidence is statistically significant at 0.05 level but not at 0.01 level. (When the result is significant at 0.05 level, it is also significant at 0.1 level or any other larger levels, and so we do not need to mention them.)

> We should be cautious about interpreting the results when they are statistically significant leading to the rejection of H_0. In general, a *statistically significant* result might not be considered as significant in practice. In the above example, if we obtain $\bar{x} = 98.25$ and $p_{\text{obs}} = 0.04$, we can reject the null hypothesis at 0.05 level and conclude that the difference between 98.25 and 98.6 is *statistically significant*. However, for some practical purposes, we might not consider the 0.35 difference as *biologically significant*. In general, even an extremely small difference between the observed sample mean and the population mean according to H_0 could eventually result in z-scores far away from zero (hence, leading to statistically significant results) as we increase the sample size n (i.e., decrease the denominator of z-score).

Finally, we should emphasize that the interpretation of p-values and their cutoffs discussed throughout this book is specific to the classical hypothesis-testing framework, where we evaluate only one hypothesis at a time. Situations where we need

to test multiple hypotheses simultaneously (i.e., using the same data) require more advanced statistical inference methods, which are not discussed in this book.

7.3.3 One-Sided Hypothesis Testing

For the body temperature example, we tested the null hypothesis $H_0 : \mu = 98.6$ against the alternative $H_A : \mu < 98.6$. We refer to such tests as **one-sided hypothesis testing**, where the departure from the null is in one direction (here, in the direction of lower values than 98.6).

Let us denote the population mean according to the null hypothesis as μ_0. Then, for the above examples, we can express our alternative hypothesis that the population mean is less than a certain value as $H_A : \mu < \mu_0$. Likewise, our null hypothesis is $H_0 : \mu = \mu_0$. In this case, we quantified the support for the null hypothesis by finding the probability of test statistic values as small or smaller than the observed value if the null hypothesis is true. The values more extreme than $\bar{x}$ (here, smaller than the observed mean 98.4) represent a larger departure from μ_0 and provide stronger evidence against the null.

In some situations, we might hypothesize that the population mean is greater than a specific value and express our hypothesis as $H_A : \mu > \mu_0$. Our null hypothesis is still $H_0 : \mu = \mu_0$. This is also a one-sided test since the departure from the null is still in one direction: toward values larger than μ_0.

For example, suppose that we have observed that many Pima Indian women suffer from diabetes. We know that obesity and diabetes are related; we might therefore hypothesize that this population is obese on average, where obesity is defined as BMI higher than 30. If we denote the population mean of BMI for Pima Indian women, we can then express our hypothesis as $\mu > 30$. In this case, the null hypothesis is $H_0 : \mu = 30$; that is, $\mu_0 = 30$.

As before, we use the sample mean as the test statistic. For illustrative purposes, suppose that we have obtained a sample of size $n = 100$ from the population of Pima Indian women. Further, suppose we know that the population variance is $\sigma^2 = 6^2$. If the null hypothesis is true and the population mean is $\mu = 30$, then the sampling distribution is

$$\bar{X}|H_0 \sim N\left(30, 6^2/100\right).$$

This distribution is shown in the left panel of Fig. 7.3. If the null hypothesis is indeed true, then we would expect to see the value of sample mean near the population mean according to the null distribution (here, 30). In contrast, if the null hypothesis is false, then the null distribution does not represent the sampling distribution of the test statistics, and we would expect to see the value of the sample mean away from 30, in this case, larger than 30 according to the alternative hypothesis.

Suppose that from our sample of 100 Pima Indian women we find that the sample mean is $\bar{x} = 31$. As before, we find the observed significance level, p-value, to measure the amount of evidence provided by the data in support for H_0. Recall
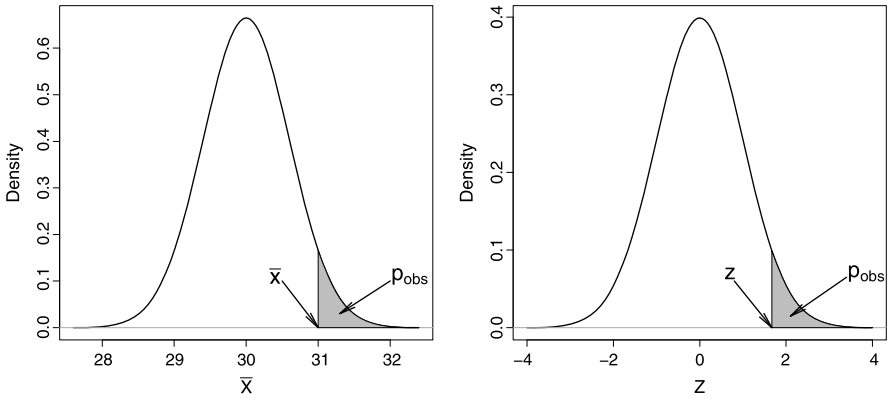
Fig. 7.3 *Left panel*: The sampling distribution for the test statistic $\bar{X}$ under the null hypothesis $H_0 : \mu = 30$. The *p*-value, which is the probability of values as or more extreme than the observed value of the test statistic $\bar{x} = 31$, is shown as the *shaded area*. *Right panel*: Obtaining the upper tail probability using one-sided *z*-test

that we defined *p*-value as the probability of values as or more extreme than the observed value of the test statistic (here, $\bar{x} = 31$) based on the null distribution, in the direction specified by the alternative hypothesis. If the null distribution is in fact true and $\mu = 30$, then values larger than $\bar{x} = 31$ would seem more extreme than what we have observed. Therefore,

$$p_{\mathrm{obs}} = P(\bar{X} \geq \bar{x}|H_0),$$

since $H_A : \mu > \mu_0$. Again, we drop H_0 for simplicity. For the above example,

$$p_{\mathrm{obs}} = P(\bar{X} \geq 31).$$

This probability is shown as the shaded area in the left panel of Fig. 7.3.

As before, we can standardize the test statistic by subtracting the mean and dividing the result by the standard deviation:

$$Z = \frac{\bar{X} - 30}{0.6} \sim N(0, 1).$$

The corresponding *z*-score is obtained as follows:

$$z = \frac{31 - 30}{0.6} = 1.67.$$

Now, to find the *p*-value, we can find the upper tail probability of $z = 1.67$ from the null distribution $N(0, 1)$:

$$p_{\mathrm{obs}} = P(Z \geq 1.67).$$

This probability is shown as the shaded area in the right panel of Fig. 7.3. We can use R-Commander to find this probability. This is the upper tail probability at 1.67 based on the standard normal distribution. Note that the upper tail probability is

by convention $P(Z > 1.67)$. However, as discussed previously, for continuous random variables, $P(Z > 1.67) = P(Z \geq 1.67)$ since the probability of any specific value (here, 1.67) is zero. For this example, $p_{\text{obs}} = 0.048$. We can reject the null hypothesis at 0.05 level but not at 0.01 level. At 0.05 level, we can conclude that the population mean of BMI for Pima Indian women is higher than 30 and the difference is statistically significant.

In general, for one-sided hypothesis testing, we evaluate the null hypothesis $H_0 : \mu = \mu_0$ by using the following standardized test statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

To this end, we find the sample mean $\bar{x}$ and calculate the observed value of Z called z-score (assuming σ is known):

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

We then use the standard normal distribution to find the p-value. If the alternative hypothesis regarding the population mean is $H_A : \mu < \mu_0$, we use the standard normal distribution to find lower tail probability of the z-score: $P(Z \leq z)$. If the alternative hypothesis regarding the population mean is $H_A : \mu > \mu_0$, we use $P(Z \geq z)$ instead. The resulting probability, p_{obs}, is the observed significance level, which can be compared to several significance levels such as 0.01, 0.05, and 0.1.

7.3.4 Two-Sided Hypothesis Testing

For many hypothesis testing problems, we might be indifferent to the direction of departure from the null value. In such cases, we can express the null and alternative hypotheses as $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$, respectively. Then we consider both large positive values and small negative values of z-score as evidence against the null hypothesis, and our alternative hypothesis is referred to as **two-sided**.

For example, suppose we believe that the average normal body temperature is different from the accepted value 98.6°F, but we are not sure whether it is higher or lower than 98.6. Then the null hypothesis remains $H_0 : \mu = 98.6$, but the alternative hypothesis is expressed as $H_A : \mu \neq 98.6$. As before, we calculate the sample mean $\bar{x} = 98.4$ and standardize it to obtain the z-score, which is -1. The p-value is still calculated as the probability of values as or more extreme than the observed z-score. However, in this case, extreme values are those whose distance from 0 is

Fig. 7.4 Illustrating the
p-value for a two-sided
hypothesis test of average
normal body temperature,
where $H_0 : \mu = 98.6$ and
$H_A : \mu \neq 98.6$. After
standardizing, $p_{obs} =$
$P(Z \leq -1) + P(Z \geq 1) =$
$2 \times 0.16 = 0.32$
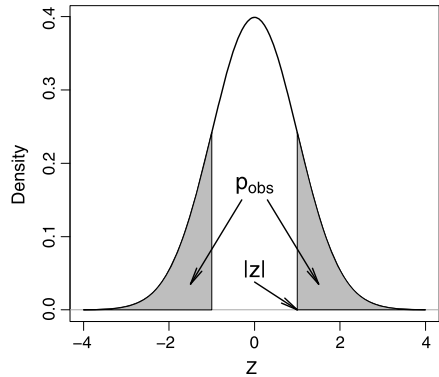
more than the distance of -1 from zero. These are values that are either less than
-1 or greater than 1. Therefore, to find the observed significance level, we need to
add the probabilities for $Z \leq -1$ and $Z \geq 1$:

$$p_{obs} = P(Z \leq -1) + P(Z \geq 1).$$

This probability is equal to the shaded area in Fig. 7.4.

To obtain the p-value for this example, we can use R-Commander to find the
lower tail probability of -1 and the upper tail probability of 1, and then add the
two probabilities. However, because of the symmetry of the standard normal distri-
bution, these two probabilities are equal. Therefore, we can just find the upper tail
probability of 1 and multiply the results by 2 to obtain the p-value:

$$p_{obs} = 2 \times P(Z \geq 1) = 2 \times 0.16 = 0.32.$$

The p-value is greater than typical significance levels such as 0.01, 0.05, and 0.1,
so we cannot reject it at these levels. Therefore, we conclude that the observed
difference is not statistically significant, and could be due to chance alone.

> In general, when we are evaluating the null hypothesis $H_0 : \mu = \mu_0$ against
> the alternative hypothesis is $H_A : \mu \neq \mu_0$, the p-value for the two-sided hy-
> pothesis test is calculated as follows (assuming σ is known):
>
> 1. Determine the observed z-score: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.
> 2. Take the absolute value of the z score: $|z|$.
> 3. Obtain the upper tail probability: $P(Z \geq |z|)$.
> 4. Double the resulting probability: $p_{obs} = 2 \times P(Z \geq |z|)$.

7.4 Hypothesis Testing Using t-tests

So far, we have assumed that the population variance σ^2 is known. Therefore, eval-
uating a hypothesis regarding the population mean did not involve estimating σ^2.

In reality, σ^2 is almost always unknown, and we need to estimate it from the data. As before, we estimate σ^2 using the sample variance S^2. We would be of course uncertain about our estimate of σ^2, and our hypothesis testing procedure should take this additional source of uncertainty into account. Similar to our approach for finding confidence intervals, we account for this additional source of uncertainty by using the t-distribution with $n - 1$ degrees of freedom instead of the standard normal distribution. The hypothesis testing procedure is then called the t-**test**.

To perform a t-test , we use the following test statistic (instead of Z):

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

where $\bar{X}$ is the sample mean, n is the sample size, S is the sample standard deviation, and μ_0 is the null value. The test statistic, T, has a t-distribution with $n - 1$ degrees of freedom under the null.

$$T \sim t(n - 1).$$

Using the observed values of $\bar{X}$ and S, the observed value of the test statistic is obtained as follows:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

We refer to t as the t-**score**.

Suppose we hypothesize that the population mean of BMI among Pima Indian women is above 30: $H_A : \mu > 30$. The corresponding null hypothesis is $H_0 : \mu = 30$. To test this hypothesis, we use the `Pima.tr` data set from the `MASS` package. (Follow the steps described in earlier chapters to upload this data set into R-Commander.) The sample size is $n = 200$. The sample mean and standard deviation are $\bar{x} = 32.31$ and $s = 6.13$, respectively. The t-score is

$$t = \frac{32.31 - 30}{6.13/\sqrt{200}} = 5.33.$$

To assess the null hypothesis $H_0 : \mu = \mu_0$ using the t-test, we first calculate the t-score based on the observed sample mean $\bar{x}$ and sample standard deviation. We then calculate the corresponding p-value as follows:

$$
\begin{aligned}
&\text{if } H_A : \mu < \mu_0, && p_{\text{obs}} = P(T \leq t), \\
&\text{if } H_A : \mu > \mu_0, && p_{\text{obs}} = P(T \geq t), \\
&\text{if } H_A : \mu \neq \mu_0, && p_{\text{obs}} = 2 \times P(T \geq |t|),
\end{aligned}
$$

where T has a t-distribution with $n - 1$ degrees of freedom, and t is our observed t-score. This is known as the **single-sample** t-**test**.

For the above example, $p_{\text{obs}} = P(T \geq 5.33)$, which we obtain from the t-distribution with $200 - 1 = 199$ degrees of freedom. To obtain this probability
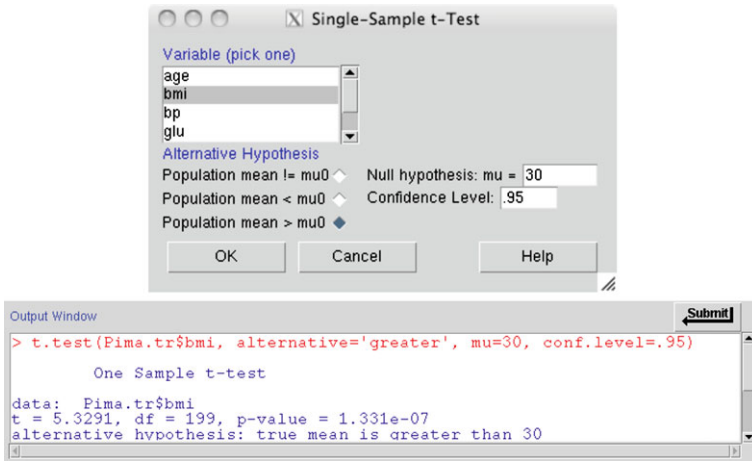
Fig. 7.5 Single-sample *t*-test using R-Commander of the null hypothesis $H_0 : \mu = 30$ against the alternative hypothesis $H_A : \mu > 30$. Based on a sample of $n = 200$ people, the observed *t*-score is 5.33 and *p*-value is $p_{obs} = P(T > 5.33) = 1.33 \times 10^{-7}$

in R-Commander, click `Distributions` → `Continuous Distributions` → `t distribution` → `t probabilities`. Then enter 5.33 for `Variable value` and 199 for `Degrees of freedom`, and select `Upper tail`. The resulting probability is 1.33×10^{-07}, which is shown as 1.33e–07. This is quite small and leads us to conclude that the result is statistically significant. At any reasonable significance level, there is strong evidence to reject the null hypothesis and conclude that the population mean of BMI among Pima Indian women is in fact greater than 30. Therefore, on average, the population is obese.

We can use R-Commander to perform a *t*-test directly. For example, let us consider the `Pima.tr` data set and test the hypothesis that $H_0 : \mu = 30$ against $H_A : \mu > 30$ for the BMI of Pima Indian women.

In R-Commander, make sure that `Pima.tr` is the active data set, then click `Statistics` → `Means` → `Single-sample t-test`. Select `bmi` as the `Variable`, select `Population mean > mu0` for the `Alternative Hypothesis`, and enter the value 30 for the `Null Hypothesis` as shown in Fig. 7.5. Note that for a two-sided test, we would have used the option `Population mean != mu0`, where the sign "!=" means not equal.

The results are given in the *Output* window in Fig. 7.5. Based on the sample of $n = 200$ people, the *t*-score is 5.33 and the degrees of freedom are `df = 200 - 1 = 199`. Therefore, the *p*-value is $P(T > 5.33) = 1.33 \times 10^{-7}$, which is exactly the same as what we found before.

7.5 Hypothesis Testing for Population Proportion

For a binary random variable X with possible values 0 and 1, we are typically interested in evaluating hypotheses regarding the population proportion of the outcome

of interest, denoted as $X = 1$. As discussed before, the population proportion is the same as the population mean for such binary variables. So we follow the same procedure as described above. More specifically, we use the z-test for hypothesis testing. Note that we do not use t-test, because for binary random variable, population variance is $\sigma^2 = \mu(1 - \mu)$. Therefore, by setting $\mu = \mu_0$ according to the null hypothesis, we also specify the population variance as $\sigma^2 = \mu_0(1 - \mu_0)$ so we do not need to estimate the population variance separately.

Now, if we assume that the null hypothesis is true, we have

$$\bar{X}|H_0 \sim N\big(\mu_0, \mu_0(1 - \mu_0)\big).$$

This means that

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\mu_0(1 - \mu_0)/n}} \sim N(0, 1).$$

As a result, we obtain the z-score as follows:

$$z = \frac{p - \mu_0}{\sqrt{\mu_0(1 - \mu_0)/n}},$$

where p is the sample proportion (mean).

Consider the Melanoma example. The data set `Melanoma` is available from the `MASS` package. Suppose that we hypothesize that less than 50% of cases ulcerate: $\mu < 0.5$. Then the null hypothesis can be expressed as $H_0 : \mu = 0.5$. Using the `Melanoma` data set, we can test the above null hypothesis. The number of observations in this data set is $n = 205$, of which 90 patients had ulceration. Therefore, $p = 90/205 = 0.44$.

Next, we can find the z-score for our test statistic as follows:

$$z = \frac{0.44 - 0.5}{\sqrt{0.5(1 - 0.5)/205}} = -1.72.$$

Because $H_A : \mu < 0.5$, the observed significance level based on this z-score is the lower tail probability $P(Z \le -1.72)$. Using R-Commander, we find the p-value to be $p_{\text{obs}} = 0.043$. Therefore, we can reject the null hypothesis at 0.05 level but not at 0.01 level.

In general, to assess the null hypothesis $H_0 : \mu = \mu_0$, where μ is the population proportion (mean) of a binary random variable, we first calculate z-score based on the observed sample proportion p:

$$z = \frac{p - \mu_0}{\sqrt{\mu_0(1 - \mu_0)/n}}.$$

Then we determine the support for the null hypothesis as:

$$\text{if } H_A : \mu < \mu_0, \quad p_{\text{obs}} = P(Z \le z),$$
$$\text{if } H_A : \mu > \mu_0, \quad p_{\text{obs}} = P(Z \ge z),$$
$$\text{if } H_A : \mu \ne \mu_0, \quad p_{\text{obs}} = 2 \times P\big(Z \ge |z|\big),$$

where Z has the standard normal distribution, and z is the observed z-score.

7.6 Advanced

In statistics, it is often convenient to assume normal distributions for random variables. In this section, we discuss a formal test of normality. We also discuss some useful R functions for hypothesis testing.

7.6.1 Test of Normality

Previously, we used Q–Q plots to visually assess the appropriateness of normality assumption for random variables. The appropriateness of the normality assumption can be evaluated formally using a testing procedure such as the **Shapiro–Wilk** test of normality. More specifically, this test evaluates the null hypothesis that the distribution of the random variable is normal. As usual, we then either reject this hypothesis and conclude that the normality assumption is not appropriate, or fail to reject it and conclude that there is no strong evidence of deviation from normality.

Suppose we assume that the `bmi` variable in `Pima.tr` has normal distribution. Let us now evaluate this assumption. In R-Commander, click `Statistics` → `Summaries` → `Shapiro-Wilk test of normality`, then select the `bmi`. The *p*-value for this test is 0.25. Therefore, we do not reject the null hypothesis (which states that the distribution is normal) and conclude that the deviation of the distribution from normality is not statistically significant.

For comparison, repeat the above steps to test the normality assumption for the `age` variable in the `Pima.tr` data set. Using the Shapiro–Wilk test, the *p*-value is 1.853×10^{-12}, which is quite small. Therefore, we can comfortably reject the null hypothesis and conclude that the deviation from normality is statistically significant.

7.6.2 Hypothesis Testing with R Programming

To perform the *z*-test in R, we can use the function *pnorm()* in order to find the *p*-value. For the body temperature example discussed at the beginning of this chapter, the *z*-score was -1. For the one-sided hypothesis of the form $H_0 : \mu < \mu_0$, we find the lower tail probability of -1 as follows:

```
> pnorm(-1, mean = 0, sd = 1, lower.tail = TRUE)

[1] 0.1586553
```

For the two-sided hypothesis, we multiply the above probability by 2. Similar approach is used for testing one-sided or two-sided hypothesis regarding population proportion.

For the BMI example, *z*-score was 1.67. For the one-sided hypothesis of the form $H_0 : \mu > \mu_0$, we need to find the upper tail probability of 1.67 as follows:

```
> pnorm(1.67, mean = 0, sd = 1, lower.tail = FALSE)
```

```
[1] 0.04745968
```

Remember to specify the option `lower.tail=FALSE` to get the upper tail probability.

When σ^2 is unknown and we need to use the data to estimate it separately, we use the t-test to evaluate hypotheses regarding the mean of a normal distribution. For the BMI example in Sect. 7.4, we found t-score was $t = 5.33$. For the one-sided hypothesis of the form $H_0 : \mu > \mu_0$, we need to find the upper tail probability of 5.33 from a t distribution with $n - 1$ degrees of freedom, where $n = 200$ in this example. We use the $pt()$ function:

```
> pt(5.33, df = 199, lower.tail = FALSE)
```

```
[1] 1.324778e-07
```

Alternatively, instead of calculating the t-score and finding the appropriate tail probabilities to obtain the p-value, we can use the function `t.test()`. For the BMI example, we use this function as follows:

```
> t.test(x = Pima.tr$bmi, mu = 30,
+    alternative = "two.sided")

          One Sample t-test

data:   Pima.tr$bmi
t = 5.3291, df = 199, p-value = 2.661e-07
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 31.45521 33.16479
sample estimates:
mean of x
     32.31
```

Here, the argument x is a (nonempty) numeric vector of data values, and `mu` is the population mean according to the null hypothesis. For one-sided t-tests, set the argument `alternative` to either "greater", or "less". Notice that the output provides the t-score (t), the degrees of freedom (df), and the p-value. Additionally, it provides the sample mean $\bar{x} = 32.31$ and the 95% confidence interval for the population mean, [31.46, 33.16]. We can estimate the interval at other confidence levels (instead of 0.95) by using the option `conf.level`:

```
> t.test(x = Pima.tr$bmi, mu = 30, conf.level = 0.9)

          One Sample t-test
```

```
data:   Pima.tr$bmi
t = 5.3291, df = 199, p-value = 2.661e-07
alternative hypothesis: true mean is not equal to 30
90 percent confidence interval:
 31.59367 33.02633
sample estimates:
mean of x
    32.31
```

Note that only the confidence interval estimate changes; the parts that are related to hypothesis testing remain as before.

Finally, to perform the Shapiro–Wilk test of normality in R, we use the function `shapiro.test()`. For example, if we assume that BMI among Pima Indian women is normally distributed, we can evaluate our assumption as follows:

```
> shapiro.test(x = Pima.tr$bmi)

        Shapiro-Wilk normality test

data:   Pima.tr$bmi
W = 0.991, p-value = 0.2524
```

In this case, the *p*-value is large, so we do not reject the null hypothesis, which states the distribution is normal, at commonly used significance levels. In other words, the test confirms our normality assumption.

7.7 Exercises

1. Suppose that the population mean of systolic blood pressure in the US is 115. We hypothesize mean systolic blood pressure is lower than 115 among people who consume a small amount (e.g., around 3.5 ounces) of dark chocolate every day. Assume that systolic blood pressure, X, in this population has a $N(\mu, \sigma^2)$ distribution. To evaluate our hypothesis, we randomly selected 100 people, who include a small amount of dark chocolate in their daily diet, and measured their blood pressure. If the sample mean is $\bar{x} = 111$ and the sample variance is $s = 32$, can we reject the null hypothesis at 0.1 confidence level?
2. Use the `Pima.tr` data set to evaluate the hypothesis that the population mean of diastolic blood pressure for Pima Indian women is not 70.
3. Consider the problem of estimating the proportion of people who regularly smoke. We use X to denote smoking status and μ to denote the population proportion of people who smoke. We hypothesize that the population proportion is less than 0.2. Write down the null and alternative hypotheses. Suppose that we interview 150 people and find that 27 of them smoke regularly. Evaluate the null hypothesis.

4. We believe that the population mean of normal body temperature is less than the widely accepted value of 98.6°F. Write down the null hypothesis and evaluate it using the "BodyTemperature.txt" data.
5. Download the "BodyTemperature.txt" data set from the book website (http://extras.springer.com). For the heart rate variable, we want to evaluate the following hypotheses. We set the significance level (cutoff) to 0.01.

   - Evaluate the hypothesis that the population mean is less than 75. Write down the null and alternative hypotheses and discuss your findings.
   - Evaluate the hypothesis that the population mean is different from 75. Write down the null and alternative hypotheses and discuss your findings.

6. We hypothesize that more than 5% of pregnant women have history of hypertension. Write down the null and alternative hypotheses. Use the `birthwt` data set (available from the MASS package) to evaluate this hypothesis (with discussion). We set the significance level (cutoff) to 0.05. (In `birthwt` data set, the variable `ht` shows the hypertension history: `ht=1` when women have history of hypertension, `ht=0` otherwise.)

Chapter 8
Statistical Inference for the Relationship Between Two Variables

8.1 Introduction

In the previous two chapters, we discussed estimation and hypothesis testing regarding the population mean of a random variable. For instance, using sample data, we estimated the population mean of normal body temperature and tested the hypothesis that the population mean is less than the accepted value of 98.6°F. Often, however, the goal of scientific studies is to investigate the relationship between two (or more) variables. For example, we might be interested in investigating the relationship between gender and body temperature. In this chapter, we discuss estimation and hypothesis testing with respect to the relationship between two random variables. We start by discussing problems where we are investigating the relationship between one binary categorical variable (e.g., gender) and one numerical variable (e.g., body temperature). (More general situations where the categorical variable could take more than two possible values are discussed later.) We then discuss some statistical inference methods to examine relationship when both random variables are binary. Finally, we review situations where both random variables are numerical.

Throughout this chapter, we assume that the individuals from which we collect data are sampled randomly and independently from the population (unless stated otherwise). This will be the case if we use simple random sampling. Also, we assume that the sample size n is large enough for the CLT to hold.

8.2 Relationship Between a Numerical Variable and a Binary Variable

In this section, we discuss situations where we investigate possible relationship between a binary random variable and a numerical random variable. In these situations, the binary variable typically represents two different groups (e.g., smoking vs. nonsmoking, male vs. female, cancer cells vs. normal cells) from the population or two

different experimental conditions (e.g., treatment A vs. treatment B). In this section, we treat the binary variable as the explanatory variable in our analysis. The binary variable is also known as the **factor**. The numerical variable, on the other hand, is regarded as the response (target) variable (e.g., body temperature).

As a running example, suppose that we believe that gender and normal body temperature are related. That is, we believe that healthy men and women are different with respect to their body temperature. We can interpret this as difference in the distributions of body temperature between men and women. The two distributions (for men and women) can of course be different in many ways. (For example, one distribution could have higher variance than the other one.) For simplicity, we focus on the means of the two distributions. If we denote the population mean of body temperature μ_1 for women and μ_2 for men, the hypothesis that the two groups are different in terms of their body temperature can be specified as $H_A : \mu_1 \neq \mu_2$. In other words, H_A states that gender is an important factor with respect to body temperature and that the two characteristics are related. The corresponding null hypothesis is that the two means are equal: $H_0 : \mu_1 = \mu_2$.

In general, we can denote the means of the two groups as μ_1 and μ_2. The null hypothesis indicates that the population means are equal, $H_0 : \mu_1 = \mu_2$. In contrast, the alternative hypothesis is one the following:

$H_A : \mu_1 > \mu_2$ if we believe the mean for group 1 is greater than the mean for group 2.

$H_A : \mu_1 < \mu_2$ if we believe the mean for group 1 is less than the mean for group 2.

$H_A : \mu_1 \neq \mu_2$ if we believe the means are different but we do not specify which one is greater.

We can also express these hypotheses in terms of the *difference* in the means: $H_A : \mu_1 - \mu_2 > 0$, $H_A : \mu_1 - \mu_2 < 0$, or $H_A : \mu_1 - \mu_2 \neq 0$. Then the corresponding null hypothesis is that there is no difference in the population means, $H_0 : \mu_1 - \mu_2 = 0$.

More generally, we can express the null hypothesis in terms of the difference between the population means as $H_0 : \mu_1 - \mu_2 = \mu_0$. However, in most cases, $\mu_0 = 0$.

Previously, we used the sample mean $\bar{X}$ to perform statistical inference regarding the population mean μ. To evaluate our hypothesis regarding the difference between two means, $\mu_1 - \mu_2$, it is reasonable to choose the difference between the sample means, $\bar{X}_1 - \bar{X}_2$, as our statistic. Here, $\bar{X}_1$ is the sample mean of the random variable of interest (e.g., body temperature) in the first group, and $\bar{X}_2$ is the sample mean in the second group. We use μ_{12} to denote the difference between the population means μ_1 and μ_2, and use $\bar{X}_{12}$ to denote the difference between the sample means $\bar{X}_1$ and $\bar{X}_2$:

$$\mu_{12} = \mu_1 - \mu_2,$$
$$\bar{X}_{12} = \bar{X}_1 - \bar{X}_2.$$

A specific value of the test statistic $\bar{X}_{12}$ based on a sample of data is denoted $\bar{x}_{12}$ and calculated as

$$\bar{x}_{12} = \bar{x}_1 - \bar{x}_2,$$

where $\bar{x}_1$ and $\bar{x}_2$ are the observed sample means for group 1 and group 2, respectively. In this case, $\bar{x}_{12}$ is our point estimate for $\mu_1 - \mu_2$, the difference between population means.

For the above example, suppose that our sample includes $n_1 = 25$ women and $n_2 = 27$ men. The sample mean of body temperature is $\bar{x}_1 = 98.2$ for women and $\bar{x}_2 = 98.4$ for men. Then, our point estimate for the difference between population means is $x_{12} = -0.2$.

As discussed before, point estimates do not reflect our uncertainty of our guess for unknown values (here, the difference between population means). To address this issue, we use interval estimates. Finding confidence intervals for the difference between two means is quite similar to steps we followed to find confidence intervals for one population mean. To start, we suppose that the sample variances for both groups are known. For the above example, we assume that $\sigma_1^2 = 0.8$ and $\sigma_2^2 = 1$.

Now, we need to find the sampling distribution of $\bar{X}_{12}$. By the Central Limit Theorem, the sampling distributions of $\bar{X}_1$ and $\bar{X}_2$ are approximately normal (exactly normal if the random variable itself is normally distributed) as follows:

$$\bar{X}_1 \sim N\left(\mu_1, \sigma_1^2/n_1\right),$$
$$\bar{X}_2 \sim N\left(\mu_2, \sigma_2^2/n_2\right),$$

where n_1 and n_2 are the number of observations, and σ_1^2 and σ_2^2 are the population variances for body temperature in group 1 and group 2, respectively.

The statistic $\bar{X}_{12}$ is the difference between the two normally distributed variables $\bar{X}_1$ and $\bar{X}_2$. As discussed in Sect. 5.8, if two random variables are normally distributed, their difference is also normally distributed with the mean equal to the difference of the means and variance equal to the *sum* of the variances. Therefore, the sampling distribution of $\bar{X}_{12}$ is also approximately normal as follows:

$$\bar{X}_{12} \sim N\left(\mu_1 - \mu_2, \ \sigma_1^2/n_1 + \sigma_2^2/n_2\right).$$

That is, the sampling distribution of $\bar{X}_{12}$ is normal with mean $\mu_{12} = \mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. We use SD_{12} to denote the standard deviation of the sampling distribution of $\bar{X}_{12}$,

$$SD_{12} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

Therefore, we can write the sampling distribution of $\bar{X}_{12}$ as follows:

$$\bar{X}_{12} \sim N\left(\mu_{12}, \ SD_{12}^2\right).$$

For our example, the variance of the sampling distribution is $0.8/25 + 1/27 = 0.07$, and the standard deviation is $SD_{12} = \sqrt{0.07} = 0.26$.

As before, we can use our point estimate and the corresponding standard deviation to find confidence intervals. In this case, the confidence interval for $\mu_{12} = \mu_1 - \mu_2$ is obtained as follows:

$$[\bar{x}_{12} - z_{\text{crit}} \times SD_{12}, \ \bar{x}_{12} + z_{\text{crit}} \times SD_{12}],$$

where z_{crit} is obtained for a given confidence level c as before. For example, the 95% confidence interval for the difference between the population means of body temperature for women and men is

$$[-0.2 - 2 \times 0.26, -0.2 + 2 \times 0.26] = [-0.72, 0.32].$$

Therefore, at 0.95 confidence level, we believe that the true difference between the two means falls between -0.72 and 0.32.

Note that the above confidence interval shows that the difference could be negative or positive. More specifically, the interval includes 0, which is interpreted as no difference between the two means, i.e., no difference between women and men in terms of mean body temperature. Therefore, even though our point estimate for the difference between the means is negative (lower mean body temperature among women compared to men), our confidence interval shows that the true difference (i.e., between population means) is quite likely to be positive (i.e., higher mean body temperature among women compared to men). As a result, we cannot say with confidence that mean body temperature among women is lower than that of men, even though our point estimate indicates that. In what follows, we discuss this more formally in the context of hypothesis testing.

We now return to our hypothesis that $H_A : \mu_{12} \neq 0$ (i.e., the difference between the two means is not zero) against the null hypothesis that $H_0 : \mu_{12} = 0$. To use $\bar{X}_{12}$ as a test statistic, we need to find its sampling distribution under the null hypothesis (i.e., its null distribution). We found that the sampling distribution of $\bar{X}_{12}$ is

$$\bar{X}_{12} \sim N\left(\mu_{12}, SD_{12}^2\right).$$

If the null hypothesis is true, then $\mu_{12} = 0$. Therefore, the null distribution of $\bar{X}_{12}$ is

$$\bar{X}_{12} \sim N\left(0, SD_{12}^2\right).$$

For the body temperature example, the null distribution of $\bar{X}_{12}$ is $N(0, 0.26^2)$.

Because we can find the distribution of $\bar{X}_{12}$ under the null (even though it is an approximate distribution when the random variable is not normally distributed), we can use it as a test statistics to examine hypotheses regarding the difference between the means of two groups, μ_{12}. As before, however, it is more common to standardize the test statistic by subtracting its mean (under the null) and dividing the result by its standard deviation. In this case, of course, the mean of the $\bar{X}_{12}$ under the null hypothesis is zero. Therefore,

$$Z = \frac{\bar{X}_{12}}{SD_{12}},$$

where Z is called the z-statistic, and it has the standard normal distribution: $Z \sim N(0, 1)$. Similarly, we standardize the observed value of the test statistic, $\bar{x}_{12}$:

$$z = \frac{\bar{x}_{12}}{SD_{12}}.$$

We refer to z as the z-score. For the body temperature example, the z-score is

$$z = \frac{-0.2}{0.26} = -0.76.$$

To test the null hypothesis $H_0 : \mu_{12} = 0$, we determine the z-score. Then, depending on the alternative hypothesis, we can calculate the p-value, which is the observed significance level, as:

$$\begin{array}{ll}
\text{if } H_A : \mu_{12} > 0, & p_{obs} = P(Z \geq z), \\
\text{if } H_A : \mu_{12} < 0, & p_{obs} = P(Z \leq z), \\
\text{if } H_A : \mu_{12} \neq 0, & p_{obs} = 2 \times P(Z \geq |z|).
\end{array}$$

The above tail probabilities are obtained from the standard normal distribution. This hypothesis testing procedure is known as the **two-sample z-test**.

For our example, $H_A : \mu_{12} \neq 0$ and $z = -0.76$. Therefore, $p_{obs} = 2P(Z \geq |-0.76|) = 2 \times 0.22 = 0.44$.

As before, we use the p-value to measure the amount of evidence against the null hypothesis. To decide whether we should reject the null hypothesis, we compare p_{obs} with predefined significance levels (cutoffs) such as 0.01, 0.05 and 0.1. For a given cutoff, we reject the null hypothesis and conclude that the result of our test is statistically significant if p_{obs} is less than the cutoff.

For the body temperature example, $p_{obs} = 0.44$ is greater than the commonly used significance levels (e.g., 0.01, 0.05, and 0.1). Therefore, the test result is not statistically significant, and we cannot reject the null hypothesis (which states that the population means for the two groups are the same) at these levels. That is, any observed difference could be due to chance alone. Recall that when we cannot reject the null hypothesis, our test remains inconclusive since our failure to reject the null could be either due to the fact that the null hypothesis is true, or it could be the case that the null hypothesis is false, but we do not have enough evidence to reject it.

8.2.1 Two-Sample t-tests for Comparing the Means

So far, we have assumed that the population variances σ_1^2 and σ_2^2 for the two groups are known, so we could find the standard deviation, SD_{12}, of the sampling distribution of $\bar{X}_{12}$. In general, this is not a realistic assumption. In this section, we discuss statistical inference regarding population means for two groups where population

variances σ_1^2 and σ_2^2 are unknown. As before, we can use the sample variances S_1^2 and S_2^2 to estimate σ_1^2 and σ_2^2, and take this additional source of uncertainty into account by using t-distributions instead of the standard normal distribution. We use s_1^2 and s_2^2 to denote the specific values of S_1^2 and S_2^2 based on the observed data. We regard s_1^2 and s_2^2 as our point estimates for population variances σ_1^2 and σ_2^2 and use them to estimate the standard deviation, SD_{12}, of the sampling distribution of $\bar{X}_{12}$.

Recall that the standard deviation of $\bar{X}_{12}$ is $SD_{12} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. We refer to our estimate of this standard deviation as the *standard error* of $\bar{X}_{12}$ and denote it as SE_{12},

$$SE_{12} = \sqrt{s_1^2/n_1 + s_2^2/n_2}.$$

For the body temperature example, suppose that the sample variances based on our sample of $n_1 = 25$ women and $n_2 = 27$ men are $s_1^2 = 1.1$ and $s_2^2 = 1.2$, respectively. Then the standard error of $\bar{X}_{12}$ is

$$SE_{12} = \sqrt{1.1/25 + 1.2/27} = 0.30.$$

Using the specific value of $\bar{X}_{12}$, which is denoted $\bar{x}_{12}$, as our point estimate for the difference between the two population means, $\mu_{12} = \mu_1 - \mu_2$, along with the standard error SE_{12} of $\bar{X}_{12}$, we find confidence intervals for μ_{12} as follows:

$$[\bar{x}_{12} - t_{\text{crit}} \times SE_{12}, \ \bar{x}_{12} + t_{\text{crit}} \times SE_{12}],$$

where t_{crit} is the t-critical value from a t-distribution for the desired confidence level c.

Previously, when we discussed statistical inference regarding one population mean with unknown variance, we used a t-distribution, whose degrees of freedom parameter df was set to $n - 1$. When comparing the population means for two groups, the formula for finding the degrees of freedom is as follows:

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{1}{n_1-1}(s_1^2/n_1)^2 + \frac{1}{n_2-1}(s_2^2/n_2)^2}. \tag{8.1}$$

For our example,

$$df = \frac{(1.1/25 + 1.2/27)^2}{\frac{1}{25-1}(1.1/25)^2 + \frac{1}{27-1}(1.2/27)^2} = 49.9.$$

Note that the degrees of freedom is not necessarily a whole number anymore as it is the case for inference regarding one population mean.

To find the corresponding t_{crit}, we follow similar steps as before. Suppose that we are interested in 95% confidence interval for μ_{12}. We find t_{crit} from the t-distribution with $df = 49.9$ degrees of freedom. In R-Commander, click `Distributions` $\rightarrow$ `t distribution` $\rightarrow$ `t quantiles`. Then enter $(1 - 0.95)/2 = 0.025$ for `Probabilities`, 49.9 for `Degrees of freedom`, and check the option `Upper tail`. The corresponding t-critical value is 2.01. Therefore,

$$[-0.2 - 2.01 \times 0.30, -0.2 + 2.01 \times 0.30] = [-0.80, 0.40].$$

Therefore, at 0.95 confidence level, we believe that the true difference between the two means falls between -0.80 and 0.40.

The formula for finding the degrees of freedom is slightly complex. As we will see later, for this type of hypothesis testing, we usually employ statistical software such as R-Commander. Therefore, we rarely need to calculate the degrees of freedom manually. Alternatively, we could use a conservative approach and set df to $\min(n_1 - 1, n_2 - 1)$, i.e., the smaller of $n_1 - 1$ and $n_2 - 1$. This leads to slightly wider confidence intervals since it uses a slightly larger t-critical value. For the above example, we could set $df = \min(25 - 1, 27 - 1) = 24$ to be conservative. The corresponding t_{crit} for 0.95 confidence level is 2.06. This results in the following 95% confidence interval:

$$[-0.2 - 2.06 \times 0.30, -0.2 + 2.06 \times 0.30] = [-0.82, 0.42],$$

which is slightly wider than what we found previously based on a more exact calculation of the degrees of freedom.

For testing a hypothesis regarding $\mu_{12} = \mu_1 - \mu_2$ when the population variances are unknown, we follow similar steps as above, but we use SE_{12} instead of SD_{12} and use the following t-statistic instead of the z-statistic to account for the additional source of uncertainty involved in estimating the population variances:

$$T = \frac{\bar{X}_{12}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}},$$

where $\bar{X}_{12} = \bar{X}_1 - \bar{X}_2$ as before. Using the observed data, we obtain $\bar{x}_{12} = \bar{x}_1 - \bar{x}_2$ as the observed value of $\bar{X}_{12}$. We also use the observed data to obtain s_1 and s_2 as the observed values of sample variances. Then, we calculate the observed value of the test statistic T as follows:

$$t = \frac{\bar{x}_{12}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$
$$= \frac{\bar{x}_{12}}{SE_{12}},$$

which is called the t-score.

Depending on the alternative hypothesis, we calculate p_{obs} as

$$\begin{aligned}
\text{if } H_A : \mu_{12} > 0, \quad & p_{\text{obs}} = P(T \geq t), \\
\text{if } H_A : \mu_{12} < 0, \quad & p_{\text{obs}} = P(T \leq t), \\
\text{if } H_A : \mu_{12} \neq 0, \quad & p_{\text{obs}} = 2 \times P(T \geq |t|),
\end{aligned}$$

where T has a t-distribution with the degrees of freedom obtained as above (Eq. 8.1). The hypothesis testing process is then called the **two-sample t-test**.

For the body temperature example,

$$t = \frac{-0.2}{0.30} = -0.67.$$
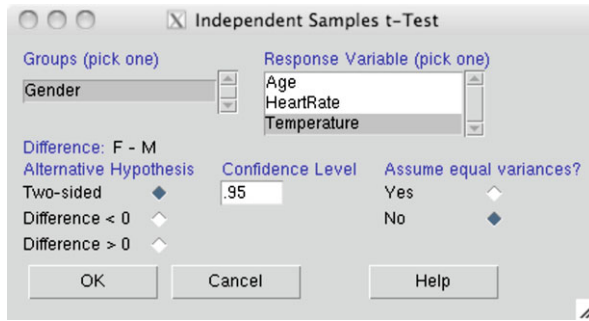
Fig. 8.1 Using R-Commander to perform two-sample t-test for the body temperature example. The binary variable Gender is selected as the factor, and the variable Temperature as the response variable. Notice that when you click on Gender, the Difference changes to F - M indicating that the mean of male group is subtracted from the mean of female group

The alternative hypothesis is $H_A : \mu_{12} \neq 0$. Using the t-distribution with $df = 49.9$ degrees of freedom, the upper tail probability of $|-0.67| = 0.67$ is $P(T > 0.67) = 0.25$. The observed significance level is $p_{obs} = 2 \times 0.25 = 0.50$, which is considered to be large (compared to commonly used significance levels). Therefore, the result is not statistically significant, and we cannot reject the null hypothesis, which indicates that the two populations (men and women) have the same mean body temperature.

For the above examples, we followed several steps to obtain the confidence interval and perform two-sample t-test. Next, we will show how to perform statistical inference regarding the difference between two population means more conveniently in R-Commander.

From the book website (http://extras.springer.com), download the "BodyTemperature.txt" data and upload it into R-Commander. To use R-Commander for two sample t-test, click Statistics → Means → Independent samples t-test (Fig. 8.1). Select Gender as the Groups variable. When Gender is selected, the Difference changes to F - M. This means that R-Commander is considering the null and alternative hypotheses in terms of the population mean in the female group minus the population mean in the male group. Now select Temperature as the Response Variable and Two-sided for the Alternative Hypothesis. Lastly, keep the confidence level at 0.95 and option Assume equal variances? as No.

The resulting t-score, the degrees of freedom df, the 95% confidence interval, and the p-value are all provided in the *Output* window (Fig. 8.2). Based on the observed data, the 95% confidence interval is $[-0.12, 0.64]$, which as before includes negative and positive values. We are 95% confident that the true value of $\mu_{12} = \mu_1 - \mu_2$ is between -0.12 and 0.64. Note that this range includes 0, which is the value of the difference between the two population means according to the null hypothesis.

The t-score is 1.35, the degrees of freedom for the t-distribution (i.e., the null distribution) is 92.26, and the corresponding p-value is 0.18. Because the p-value

```
Output Window                                                          Submit

> t.test(Temperature~Gender, alternative='two.sided', conf.level=.95,
+   var.equal=FALSE, data=bodyTemperature)

        Welch Two Sample t-test

data:  Temperature by Gender
t = 1.3526, df = 92.265, p-value = 0.1795
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1212510  0.6390581
sample estimates:
mean in group F mean in group M
      98.45686        98.19796
```

Fig. 8.2 The results of two-sample t-test for the body temperature example. The t-score is 1.35, the degrees of freedom for the t-distribution (i.e., the null distribution) is 92.26, and the corresponding p-value is 0.18. The 95% confidence interval is $[-0.12, 0.64]$

Fig. 8.3 The `balance` data sets that includes measurements of mean sway range (in millimeters) in the forward/backward plane and side/side plane for two groups of subjects: elderly and young

subject_No	forward_backward	side_side	Age_Group
1	19	14	elderly
2	30	41	elderly
3	20	18	elderly
4	19	11	elderly
5	29	16	elderly
6	25	24	elderly
7	21	18	elderly
8	24	21	elderly
9	50	37	elderly
1	25	17	young
2	21	10	young
3	17	16	young
4	15	22	young
5	14	12	young
6	14	14	young
7	22	12	young
8	17	18	young

is 0.18, we cannot reject the null hypothesis that $\mu_{12} = 0$ at commonly used significance levels (0.01, 0.05, 0.1). We say the result is not statistically significant, and any observed difference could be due to chance alone.

As the second example, we consider an experiment where the amount of mean sway range (in millimeters) in the forward/backward plane and side/side plane were recorded for two groups of subjects, young and elderly, while taking part in a reaction time test [35]. The data set includes $n_1 = 9$ elderly subjects and $n_2 = 8$ young subjects. Each subject was asked to stand barefoot on a "force platform" and maintain a stable upright position. Then, they were supposed to react as quickly as possible to an unpredictable noise by pressing a hand-held button. The noise was produced randomly. The platform automatically measured how much a subject swayed in millimeters in both the forward/backward and the side-to-side directions.

Obtain the data set from http://lib.stat.cmu.edu/DASL/Datafiles/Balance.html, save it as a text file and load it into R-Commander under the name `balance`. The data set is shown in Fig. 8.3.

We are interested in the relationship between the age group and the sway range in the forward/backward plane. Denote the variable for sway range in the for-

```
Output Window                                                       Submit

        Welch Two Sample t-test

data:   forward_backward by Age_Group
t = 2.3035, df = 10.971, p-value = 0.04183
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.3627401 16.0539266
sample estimates:
mean in group elderly    mean in group young
             26.33333                18.12500
```

Fig. 8.4 The results of two-sample t-test for the balance example. The t-score is 2.3, the degrees of freedom for the t-distribution is 10.97, and the corresponding p-value is 0.042. The 95% confidence interval is [0.36, 16.05]

ward/backward plane as X. The population mean and variance of X among the elderly subjects are denoted as μ_1 and σ_1^2. For the young subjects, we denote the population mean and variance of X as μ_2 and σ_2^2. We set $\mu_{12} = \mu_1 - \mu_2$. We hypothesize that the age group and the sway range in the forward/backward plane are related so the population means of X for young and old subjects are different. We specify this hypothesis as $H_A : \mu_{12} \neq 0$. In contrast, we specify the null hypothesis as $H_0 : \mu_{12} = 0$.

We can use R-Commander to estimate confidence intervals and perform hypothesis testing. Click Statistics → Means → Independent samples t-test. Select Age_Group as the Groups variable. When Age_Group is selected, the Difference changes to elderly - young. This means that R-Commander is considering the null and alternative hypotheses in terms of the population mean in the elderly group minus the population mean in the young group. Now select forward_backward as the Response Variable and Two-sided for the Alternative Hypothesis. Lastly, keep the confidence level at 0.95 and option Assume equal variances? as No.

The resulting t-score, the degrees of freedom df, the 95% confidence interval, and the p-value are all given in the *Output* window (Fig. 8.4). The t-score is $t = 2.3$. Based on the $df = 10.97$, the 95% confidence interval is [0.36, 16.05]. Therefore, we are 95% confident that the true value of $\mu_{12} = \mu_1 - \mu_2$ is between 0.36 and 16.05. Note that the values in this range are all positive. More specifically, the value 0 stated by the null hypothesis for μ_{12} is not included in this range. We investigate this more formally through hypothesis testing.

The results in Fig. 8.4 show that the p-value is 0.042. Consequently, at the 0.05 significance level (but not at the 0.01 significance level), the data provide enough evidence to reject the null hypothesis that $\mu_{12} = 0$, i.e., the population means are equal for elderly people and young people. Therefore, at the 0.05 significance level we conclude that the difference between the two groups in terms of the sway range in the forward/backward plane is statistically significant. That is, the observed difference $x_{12} = 8.2$ between the two groups is not likely to be due to chance alone.

8.2.2 Pooled t-test

When we used R-Commander to perform two-sample t-tests, we kept the option `Assume equal variances?` at its default value `No`. This means that we do not assume that the variances for the two groups are the same. Assuming $\sigma_1^2 = \sigma_2^2$ is not reasonable in general and should be avoided. Statisticians used to make this assumption for convenience when computer programs for statistical analysis were not available.

If we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we need to estimate only one (instead of two) variance parameter, σ^2, which is the common variance between the two groups. We estimate σ^2 using the **pooled sample variance**, s_p^2, as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \tag{8.2}$$

where n_1 and n_2 are the sample sizes and s_1 and s_2 are sample variances for the two groups. Note that the pooled sample variance is in fact the weighted average of group-specific sample variances, where the $n_1 - 1$ and $n_2 - 1$ are the weights (i.e., the group with larger sample size is weighted higher).

To obtain the t-score, the t-critical value, and p-value, we follow a similar procedure as the standard two-sample t-test discussed above, but this time, we use s_p^2 instead of s_1^2 and s_2^2, and set the degrees of freedom to $df = n_1 + n_2 - 2$. In this case, the standard error (for $\bar{X}_{12}$) and t-score are calculated as follows:

$$SE_{12} = \sqrt{s_p^2/n_1 + s_p^2/n_2} = s_p\sqrt{1/n_1 + 1/n_2},$$

$$t = \frac{\bar{x}_{12}}{s_p\sqrt{1/n_1 + 1/n_2}},$$

where $\bar{x}_{12}$ is the difference between the observed sample means as before.

Repeat the steps for using R-Commander to perform two sample t-test in order to compare body temperature between male and female groups, but this time set the option `Assume equal variances?` to `Yes`. Compare your results with those from the standard two sample t-test.

8.2.3 Paired t-test

When using two-sample t-test to investigate the relationship between a binary variable that defines the grouping of the individuals (e.g., gender) and the response variable (e.g., body temperature), we hope that the individuals in the two samples are quite comparable except for the characteristic that defines the groups. In our body temperature example, the two groups (female and male) should be similar with respect to other possibly important factors affecting body temperature, such as age and ethnicity, so the only factor that separates the two groups is gender. This way, if the observed difference in mean body temperature between the two groups is significant, it is likely to be related to gender. (Of course, even if we establish that
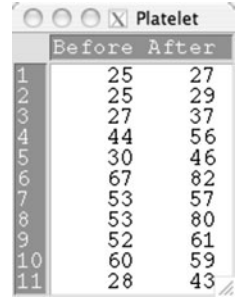
there is a relationship between gender and body temperature, we cannot define it as causation since the data are obtained from an observational study.)

While we hope that the two samples taken from the population are comparable except for the characteristic that defines the grouping, this is not guaranteed in general. For the body temperature example, one group might include relatively older participants. To mitigate the influence of other important factors (e.g., age) that are not the focus of our study, we sometimes **pair** each individual in one group with an individual in the other group so that the paired individuals are very similar to each other except for the characteristic that defines the grouping. For example, when we are investigating the relationship between gender and body temperature and we are concerned that the two samples might not be comparable with respect to factors such as age and ethnicity, we can recruit twins with different genders for our study so that each individual in the female group is paired by her twin in the male group. This way, we make sure that the two samples are exactly the same in terms of age and ethnicity, and they are comparable with respect to other possibly important characteristics such as genetic factors.

Often, not only are the two samples related, they in fact include the same individuals. For example, suppose that we are investigating the effect of a specific diet on blood pressure. We could of course recruit a sample of subjects and ask them to follow that specific diet for six months. For comparison, we recruit another sample of subjects who do not follow our prescribed diet. At the end of the study period, we compare the two groups in terms of their blood pressure using the two-sample t-test described previously. However, it is possible that just by chance the subjects in the diet group tend to have lower blood pressure even before our experiment starts. For example, they might be relatively younger than the control group, or they might exercise more. To avoid such issues, we can design our experiment so that the same individuals participate in both groups. To this end, we can recruit subjects that are not following our prescribed diet, measure their blood pressure, ask them to follow the diet for six months, and measure their blood pressure again at the end of the study. This way, the two groups include the same subjects under different conditions: before the diet and after the diet.

The two-sample t-test we described previously is based on the assumption that the two groups are unrelated (independent). When the individuals in the two groups are paired, we use the **paired t-test** to take the pairing of the observations between the two groups into account. In what follows, we use the study of the effect of tobacco smoke on platelet function by Levine [16] to describe this method. In his study, Levine hypothesized that the higher frequency of arterial thrombosis in cigarette smokers could be partially explained by increased platelet aggregation caused by smoking. To test this hypothesis, Levine conducted an experiment where he selected a group of eleven people and measured platelet aggregation before and after smoking a cigarette for each individual. Therefore, observations in the "Before" sample and "After" sample are from the same subjects. For each subject, an observation in the "Before" sample is paired with an observation in the "After" sample. The data set `Platelet` for this experiment is available from the book website (http://extras.springer.com). Load the `Platelet` data set in R-Commander and try viewing it (Fig. 8.5).

Fig. 8.5 Viewing the `Platelet` data set in R-Commander. For 11 people, there are observations on platelet aggregation before smoking (`Before`) and platelet aggregation after smoking (`After`)

Fig. 8.6 Computing the difference variable D in R-Commander
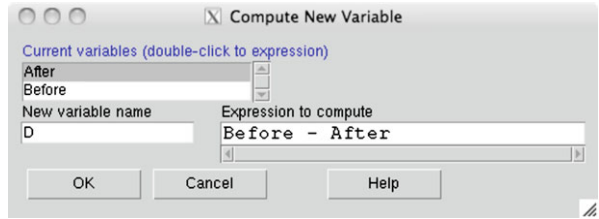
> To account for the dependency between the observations in the two groups, we use the paired t-test instead of the independent two-sample t-test discussed above. Specifically, we compare each observation in the first group to its corresponding observation in the second group. Using the difference between the paired observations, the hypothesis testing problem reduces to a single sample t-test problem (Sect. 7.4).

For the `Platelet` data, we want to compare platelet aggregation measurements for the same person before and after smoking. Let us define a new random variable D, which represents the difference in platelet aggregation from before to after. In R-Commander, we can create the difference variable D and then conduct a single sample t-test. Click `Data → Manage variables in active data set → Compute new variables`. Under `New variable name` enter D and under `Expression to compute` enter `Before - After`, as in Fig. 8.6. Now try viewing the data set. The values of D are the differences in platelet aggregation measurements from `Before` to `After`.

Because we believe that platelet aggregation before smoking tends to be less than after, we expect D to be negative on average. Therefore, we could express the alternative hypothesis as $H_A : \mu < 0$, where μ is the population average of the random variable D. However, to be conservative, we consider the possibility that μ could also be positive and specify the alternative hypothesis as $H_A : \mu \neq 0$. Then the null hypothesis is that the mean of change in platelet aggregation due to smoking is zero, $H_0 : \mu = 0$. We can use the methods discussed in the previous chapter for inference regarding one population mean to find confidence intervals for μ and test the null hypothesis. As before, we use the sample mean, $\bar{D}$, for this purpose.

Using the `Platelet` data, the observed value of $\bar{D}$ is $\bar{d} = -10.27$. Using the sample standard deviation $s = 7.98$ for D and the sample size $n = 11$, the standard error (i.e., estimated standard deviation) of $\bar{D}$ is

$$SE = \frac{7.98}{\sqrt{11}} = 2.41.$$

Suppose that we are interested in 95% confidence interval estimate for μ (i.e., population mean of the difference between before and after smoking). Using the t-distribution with $n - 1 = 10$ degrees of freedom, we obtain $t_{\text{crit}} = 2.23$. Therefore, the 95% confidence interval is

$$[\bar{d} - t_{\text{crit}} \times SE, \bar{d} + t_{\text{crit}} \times SE] = [-10.27 - 2.23 \times 2.41, -10.27 + 2.23 \times 2.41]$$
$$= [-15.64, -4.90].$$

At 0.95 confidence level, we believe that the true mean of the difference in platelet aggregation measurements before and after smoking is between -15.64 and -4.90. Note that this range includes negative values only. More specifically, it does not include the value 0 specified by the null hypothesis.

To perform hypothesis testing, we find the t-score (here, $\mu_0 = 0$ according to the null hypothesis) as follows:

$$t = \frac{-10.27}{2.41} = -4.26.$$

To find the p-value, we find the upper tail probability of $|-4.26| = 4.26$ from the t-distribution with 10 degrees of freedom, and multiply the results by 2 for two-sided hypothesis testing:

$$p_{\text{obs}} = 2P(T > 4.26) = 2 \times 0.0008 = 0.0016.$$

At 0.01 confidence level, we can reject the null hypothesis and conclude that the test result is statistically significant. In this case, we interpret this as a statistically significant relationship between smoking and platelet aggregation.

Now suppose that we had ignored the pairing and treated the two groups `Before` and `After` as unrelated (independent). Then we would erroneously conduct a two-sample t-test. The value of the t-score in this case would be $t = -1.42$, the degrees of freedom would be $df = 19.52$, and the resulting p-value for two-sided hypothesis testing would be $p_{\text{obs}} = 0.17$. Then, we would fail to reject the null hypothesis at commonly used significant levels and conclude that the relationship between smoking and platelet aggregation is not statistically significant.

> When performing t-tests, ignoring the dependence between the two groups is inappropriate and possibly results in the wrong conclusion.

In general, we perform the paired t-test as follows. Suppose that there are n observations in the first sample and n observations in the second sample. Therefore, there are n pairs of observations and $2n$ observations in total. Now consider the ith

pair of observations, x_{i1} and x_{i2}, where x_{i1} is the observation in the first sample, and x_{i2} is the corresponding observation in the second sample. We find the difference $d_i = x_{i1} - x_{i2}$ between the paired observations. We assume that d_i is an observation for the random variable D. We will have n observed values for D, where each value is the difference between a pair of observations from the original data. We now use the single sample t-test (Sect. 7.4) to evaluate the null hypothesis $H_0 : \mu = \mu_0$, where μ is the population mean of D, and μ_0 is usually zero (i.e., the difference between paired observations is zero on average). As before, the alternative is either one sided or two sided.

Using the observed sample mean of D, which we denote as $\bar{d}$, and the observed sample standard deviation s, we find confidence intervals for μ:

$$[\bar{d} - t_{\text{crit}} \times SE, \bar{d} + t_{\text{crit}} \times SE],$$

where t_{crit} is the factor obtained for the desired confidence level c from the t-distribution with $n - 1$ degrees of freedom, and $SE = s/\sqrt{n}$ is the standard error for the statistic $\bar{D}$.

To test the null hypothesis $H_0 : \mu = 0$, we calculate the T statistic,

$$T = \frac{\bar{D}}{S/\sqrt{n}},$$

where $\bar{D}$ is the sample mean of the paired differences, S is the sample standard deviation of D, and n is the number of pairs. If the null hypothesis is true, then the test statistic T has the t-distribution with $n - 1$ degrees of freedom. We calculate the corresponding t-score as follows:

$$t = \frac{\bar{d}}{s/\sqrt{n}}.$$

Then the p-value is the probability of having as extreme or more extreme values than the observed t-score:

$$
\begin{aligned}
&\text{if } H_A : \mu > 0, \quad p_{\text{obs}} = P(T \geq t), \\
&\text{if } H_A : \mu < 0, \quad p_{\text{obs}} = P(T \leq t), \\
&\text{if } H_A : \mu \neq 0, \quad p_{\text{obs}} = 2 \times P(T \geq |t|).
\end{aligned}
$$

Instead of creating the variable D and performing the single-sample t-test, we can use R-Commander to perform a paired t-test directly. Click Statistics $\rightarrow$ Means $\rightarrow$ Paired t-test. Select Before as the First variable and After as the Second variable as in Fig. 8.7. Then select Two-sided as the Alternative Hypothesis. The difference variable D is automatically calculated as the first variable minus the second variable. (If we had specified the alternative hypothesis as $H_A : \mu < 0$, we would have set the option to Difference < 0.) The results shown in the *Output* window (Fig. 8.8) are identical to those found by the single sample t-test for D.

Fig. 8.7 Paired t-test in
R-Commander. We are
testing the null hypothesis
that the mean difference in
platelet aggregation `Before`
and `After` smoking is 0
against the alternative
hypothesis that $H_A : \mu \neq 0$,
where μ is the mean of the
paired differences
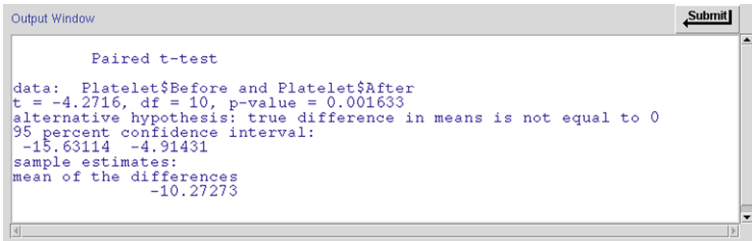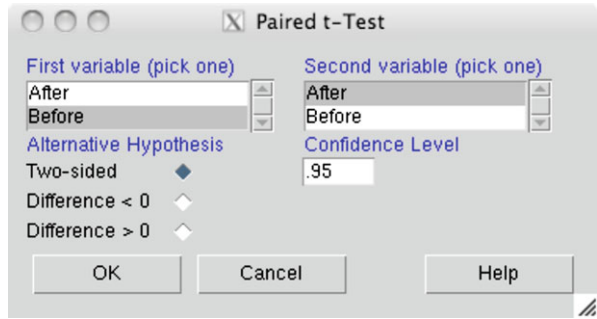
Fig. 8.8 The output of the paired t-test for evaluating the null hypothesis that the mean differ-
ence in platelet aggregation `Before` and `After` smoking is 0 against the alternative hypothesis
that $H_A : \mu \neq 0$. The results are similar to those based on creating the difference variable D and
performing the single sample t-test

8.3 Inference about the Relationship Between Two Binary Variables

In this section, we discuss statistical inference methods for evaluating the relation-
ship between two binary random variables. As an example, suppose that we want
to investigate whether smoking during pregnancy increases the risk of having a low
birthweight baby. We use the `birthwt` data set from the `MASS` package for this
purpose. The random variable of interest (i.e., response variable) is `low`, indicating
whether the baby's birthweight was less than 2.5 kg. The explanatory variable is
`smoke`, indicating the mother's smoking status during pregnancy. Since these vari-
ables are recorded as `0` and `1`, first make sure that they are converted to categorical
variables. (Click `Data → Manage variables in active data set →`
`Convert numeric variables to factors`.)

A common way to analyze the relationship between binary (in general, categor-
ical) variables is to use contingency tables. Contingency tables are a tabular rep-
resentations of the frequencies for all possible combinations of the two variables.
To obtain the contingency table in R-Commander, click `Statistics → Con-`
`tingency tables → Two-way table`. Select `smoke` as the `Row vari-`
`able` (i.e., X) and `low` as the `Column variable` (i.e., Y), as in Fig. 8.9. Check
`Row percentages` and uncheck `Chi-square test of independence`
for now.

Fig. 8.9 Creating the contingency table for the mother's smoking status (the row variable) by the baby's birthweight status (the column variable)
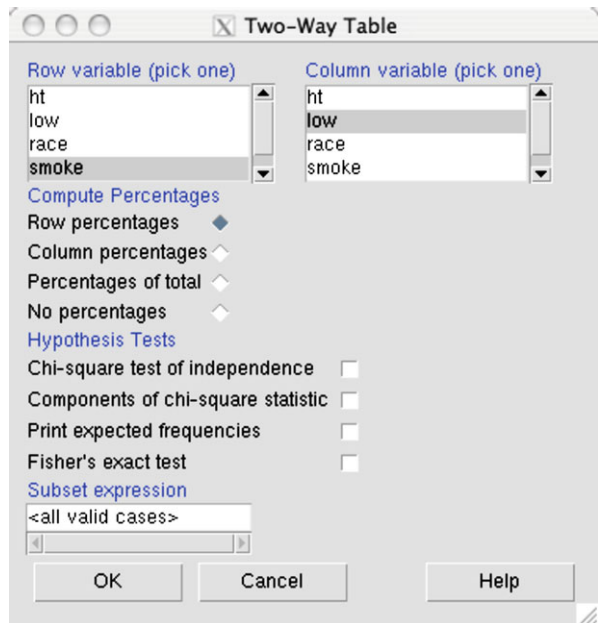
Table 8.1 Contingency table of `low` by `smoke`

Frequency		low		Total
		0	1	
smoke	0	86	29	115
	1	44	30	74

Table 8.2 Sample proportions of babies with normal birthweight (`low=0`) and babies with low birthweight (`low=1`) for each smoking status

Proportion		low		Total
		0	1	
smoke	0	0.75	0.25	1
	1	0.60	0.40	1

The resulting tables shown in the *Output* window (see Tables 8.1 and 8.2). Table 8.1 provides the frequency of each cell. The first row in this table shows non-smoking mothers, and the second row shows smoking mothers. The first column shows the number of babies of normal birthweight, and the second column shows the number of babies of low birthweight. We regard nonsmoking mothers as the first group, with $n_1 = 115$, and smoking mothers as the second group. with $n_2 = 74$.

To perform statistical inference in this section, we rely on the CLT and assume that the distributions of sample proportions (i.e., sample means for n binary random variables) are approximately normal. For this assumption to be reasonable, the frequencies in each cell of the contingency table should be at least 5.

Table 8.2 (row percentages) provides the relative frequencies or sample proportions for each row separately. For example, the proportion of babies with low birthweight among nonsmoking mothers (i.e., first row and second column) is 0.25.

We are interested in the relationship between the two binary variables. If having low-birthweight babies is related to smoking during pregnancy, we expect the population means of low-birthweight babies to be different between smoking and nonsmoking mothers. Of course, for binary random variables, population mean is the same as the population proportion for the outcome of interest (denoted as 1). We denote the population proportion of low-birthweight babies for nonsmoking mothers as μ_1, and the population proportion of low-birthweight babies for smoking mothers as μ_2. We use μ_{12} to denote the difference between these two proportions: $\mu_{12} = \mu_1 - \mu_2$.

If smoking and low birthweight are related, we expect the two population proportions to be different and μ_{12} to be away from zero. Therefore, we can express our hypothesis regarding the relationship between the two variables as $H_A : \mu_{12} \neq 0$. We could of course be more specific and specify our hypothesis as $H_A : \mu_{12} < 0$ if we believe that the population proportion of low-birthweight babies among nonsmoking mothers, μ_1, is less than the population proportion of low-birthweight babies among smoking mothers, μ_2. However, to be conservative, we use the two-sided alternative. The corresponding null hypothesis is then $H_0 : \mu_{12} = 0$.

To find the confidence intervals for μ_{12}, we use the difference between sample proportions, $\bar{X}_{12} = \bar{X}_1 - \bar{X}_2$, as the statistic. According to the CLT, the sampling distribution of $\bar{X}_{12}$ is approximately normal,

$$\bar{X}_{12} \sim N\left(\mu_1 - \mu_2, \ \sigma_1^2/n_1 + \sigma_2^2/n_2\right).$$

As before, we use SD_{12} to denote the standard deviation of the sampling distribution of $\bar{X}_{12}$. Recall that for binary random variables, the population variance is $\sigma^2 = \mu(1 - \mu)$. Therefore, we can write SD_{12} as follows:

$$SD_{12} = \sqrt{\mu_1(1 - \mu_1)/n_1 + \mu_2(1 - \mu_2)/n_2}.$$

Then, we write the sample distribution of $\bar{X}_{12}$ as

$$\bar{X}_{12} \sim N\left(\mu_{12}, SD_{12}^2\right).$$

The observed value of this statistic for the sample data is $\bar{x}_{12} = \bar{x}_1 - \bar{x}_2$. For binary random variables, it is common to use p_1 and p_2 for observed sample proportions. Therefore, we denote the observed difference between sample proportions as p_{12}. For our example, $p_1 = 0.25$ and $p_2 = 0.40$. These are the point estimates for μ_1 and μ_2, respectively. As a result, the point estimate of μ_{12} is $p_{12} = 0.25 - 0.40 = -0.15$.

We also use p_1 and p_2 to estimate the standard deviation SD_{12}. We refer to our estimate of SD_{12} as the standard error and denote it as SE_{12}:

$$SE_{12} = \sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}.$$

For the above example,

$$SE_{12} = \sqrt{0.25(1 - 0.25)/115 + 0.40(1 - 0.40)/74} = 0.07.$$

Using the point estimate p_{12} along with the standard error SE_{12}, we can find confidence intervals for μ_{12} as follows:

$$[p_{12} - z_{\mathrm{crit}} \times SE_{12}, p_{12} + z_{\mathrm{crit}} \times SE_{12}],$$

where z_{crit} is obtained for a given confidence level c as before. Note that we use z_{crit} even though the population variances were unknown. This is because we did not use the data to estimate them separately; rather, we used our point estimates for the population proportions.

For the birthweight example, the 95% confidence interval of μ_{12} is

$$[-0.15 - 2 \times 0.07, -0.15 + 2 \times 0.07] = [-0.29, -0.01].$$

Therefore, we are 95% confident that the difference between the two population proportions falls between -0.29 and -0.01. Note that all the values in this range are negative. The value of μ_{12} is negative when μ_1 (i.e., population proportion of low-birthweight babies among nonsmoking mothers) is less than μ_2 (i.e., population proportion of low-birthweight babies among smoking mothers). More specifically, the interval does not include 0, which is the value specified by the null hypothesis.

More formally, we test the null hypothesis, $H_0 : \mu_{12} = 0$, using the two-sample z-test as follows. First, we obtain the z-score,

$$z = \frac{\bar{p}_{12}}{SE_{12}}.$$

Then, we calculate the p-value, which is the observed significance level:

$$\begin{aligned} &\text{if } H_A : \mu_{12} > 0, & p_{\mathrm{obs}} &= P(Z \geq z), \\ &\text{if } H_A : \mu_{12} < 0, & p_{\mathrm{obs}} &= P(Z \leq z), \\ &\text{if } H_A : \mu_{12} \neq 0, & p_{\mathrm{obs}} &= 2 \times P(Z \geq |z|). \end{aligned}$$

The above tail probabilities are obtained from the standard normal distribution.

For our example,

$$z = \frac{-0.15}{0.07} = -2.14.$$

Because we specified the alternative hypothesis as $H_{12} : \mu_{12} \neq 0$, the p-value is two times the upper tail probability of $|-2.14| = 2.14$ from the standard normal distribution, that is, $p_{\mathrm{obs}} = 2 \times 0.016 = 0.032$.

At 0.05 level (but not at 0.01 level), we can reject the null hypothesis and conclude that the observed difference in the proportion of low-birthweight babies is statistically significant and is probably not due to the chance alone. Therefore, at 0.05 level, we can conclude that the two variables, smoking during pregnancy and having low-birthweight babies, are related.

8.4 Inference Regarding the Linear Relationship Between Two Numerical Variables

In this section, we discuss statistical inference methods for investigating possible linear relationship between two numerical variables. As an example, suppose that

we believe that percent body fat is related to the abdomen circumference measurement among men. Let us denote abdomen circumference as X and percent body fat as Y.

A simple approach to quantify the strength and direction of a linear relationship between two random variables is **Pearson's correlation coefficient**, also known as **Pearson's product-moment correlation**. For a population, we denote this measure as ρ (Greek letter "rho") and calculate it as

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}.$$

Here, μ_x and μ_y are the population means of X and Y, σ_x and σ_y are the population standard deviations, and N is the population size.

Therefore, ρ is the average of the product of deviations (each observation from its population mean) scaled by the standard deviations. It is a number between -1 and 1, and as the linear relationship becomes stronger, ρ moves away from zero and approaches 1 for positive relationships and -1 for negative relationships.

As usual, we cannot measure ρ directly, because we do not have access to all members of the population. Therefore, we need to estimate ρ by obtaining a sample of size n from the population.

The usual estimator for the population correlation coefficient ρ is the sample correlation coefficient R. Given n pairs of values, $(X_1 Y_1), (X_2, Y_2), \ldots,$ (X_n, Y_n), randomly sampled from the population, we obtain R as

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}.$$
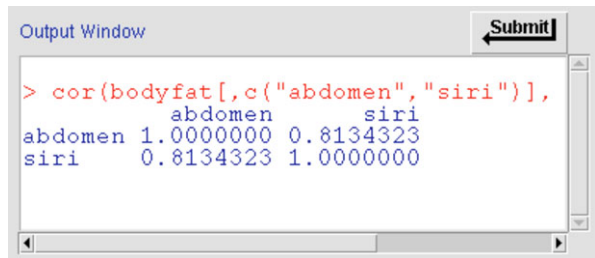
Here, $\bar{X}$ and $\bar{Y}$ are the sample means, and S_x and S_y are the sample standard deviations for X and Y, respectively. Note that similar to the sample variance, we use $n-1$ instead of n in the denominator. We denote the specific value of R based on our sample as r,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}.$$

Here, $(x_n, y_n), (x_n, y_n), \ldots, (x_n, y_n)$ are n pairs of observed values in our data.

To examine the relationship between percent body fat and abdomen circumference, we use the `bodyfat` data set (discussed in Chap. 3) in order to calculate r as

Fig. 8.10 Viewing the correlation matrix for `abdomen` and `siri`. Each element is the sample correlation coefficient between the row variable and column variable


```
Output Window                                        Submit

> cor(bodyfat[,c("abdomen","siri")],
             abdomen      siri
abdomen 1.0000000 0.8134323
siri    0.8134323 1.0000000
```

our point estimate for ρ. (Follow the steps discussed in Chap. 3 to load the data in R-Commander.) In this data set, `siri` shows the percent body fat for each person, and `abdomen` shows the measurements for abdomen circumference in centimeters. To calculate r, click `Statistics → Summaries → Correlation matrix`. Select `abdomen` and `siri` (hold down the control key) as the `Variables` and `Pearson product-moment` for `Type of Correlations`.

The result, shown in Fig. 8.10, has a matrix format. Each element of the matrix provides the sample correlation coefficient between the corresponding row and column variables. The top right element, $r = 0.81$, is the sample correlation coefficient between `abdomen` and `siri` based on a sample of $n = 252$ men. This is the same as the bottom left element, which shows the correlation coefficient between `siri` and `abdomen`, since the order of the two random variables does not affect their correlation. The sample correlation coefficient in this case is away from zero and close to 1. This indicates that there is a strong positive linear relationship between the two variables: as one increases, the other one also tends to increase.

Now repeat the above steps to find the sample correlation coefficient between `height` and `siri`. This time, $r = -0.09$. While the estimate of correlation coefficient is negative, it does not indicate a strong *linear* relationship between the two variables because it is very close to zero. In what follows, we discuss a simple statistical method for evaluating the strength of the linear relationship captured by the correlation coefficient.

We can express our hypothesis about the linear relationship between two random variables in terms of their correlation coefficient. For example, we might believe that as abdomen circumference increases, percent body fat also increases (i.e., there is a positive relationship between the variables). Then the alternative hypothesis can be formalized as $H_A : \rho > 0$. Likewise, we might believe that as the height increases, percent body fat decreases: $H_A : \rho < 0$. On the other hand, we might believe that percent body fat is related to height, but we are unsure of the direction: $H_A : \rho \neq 0$. In all cases, the null hypothesis is that these variables are not linearly related, $H_0 : \rho = 0$.

Note that we emphasize the word "linear" since the correlation coefficient captures the linear relationship between two variables; when it is close to zero, it means that either the two variables are not related, or they are related, but the relationship is not linear. Therefore, we should be cautious about interpreting a correlation coefficient close to zero as no relationship between the random variables.

To evaluate the null hypothesis that the two variables are not linearly related ($H_0 : \rho = 0$), we could use a **correlation test** based on the following test statistic:

$$T = \frac{R}{\sqrt{(1 - R^2)/(n - 2)}},$$

where R is the sample correlation coefficient, and n is the sample size. If the null hypothesis is true, then the distribution of T (i.e., its null distribution) is the t-distribution with $n - 2$ degrees of freedom.

The observed value of the test statistic is denoted t and calculated as

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}},$$

where r is the observed correlation coefficient based on our sample. Then we determine the amount of support against the null hypothesis as:

$$\text{if } H_A : \rho > 0, \quad p_{obs} = P(T \geq t),$$
$$\text{if } H_A : \rho < 0, \quad p_{obs} = P(T \leq t),$$
$$\text{if } H_A : \rho \neq 0, \quad p_{obs} = 2 \times P(T \geq |t|).$$

As an example, suppose that we want to examine the linear relationship between `height` and `siri`. We hypothesize that the two variables are related, but we are reluctant to specify the direction of the relationship. Therefore, we want to test $H_0 : \rho = 0$ versus $H_A : \rho \neq 0$.

Previously, we found that the sample correlation coefficient between these two variables is $r = -0.09$ based on a sample of size $n = 252$ men. Therefore, the observed value of the test statistic is

$$t = \frac{-0.09}{\sqrt{(1 - (-0.09)^2)/(252 - 2)}} = -1.42.$$

Since the alternative hypothesis is $H_A : \rho \neq 0$, the p-value is obtained by calculating the upper tail probability of $|-1.42| = 1.42$ based on a t-distribution with $252 - 2 = 250$ degrees of freedom and multiplying the results by 2. Using R-Commander, the observed significance level is $p_{obs} = 2P(T \geq 1.42) = 0.16$. At commonly used significance levels (0.01, 0.05, and 0.1), this is not a statistically significant result, and we cannot reject the null hypothesis. That is, we cannot reject the possibility that the observed negative correlation coefficient could have been due to the chance alone, and we cannot conclude that the two variables are linearly related.

Alternatively, in R-Commander, we can directly test our hypotheses regarding the linear relationship between two numerical variables. For the height and percent body fat example, the null and alternative hypotheses were $H_0 : \rho = 0$ and $H_A : \rho \neq 0$. To evaluate the null hypothesis, click `Statistics` → `Summaries` → `Correlation test`. Again, select `height` and `siri` as the `Variables` and

Fig. 8.11 Correlation test in
R-Commander. Here, we are
testing the null hypothesis
$H_0 : \rho = 0$ against the
alternative $H_A : \rho \neq 0$ for the
relationship between
`height` and `siri` in the
`bodyfat` data set
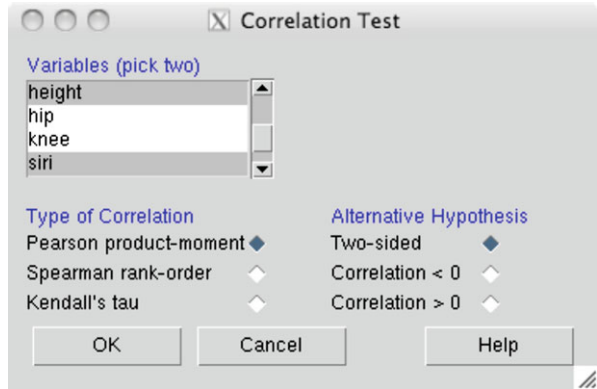
Fig. 8.12 Output for the hypothesis test regarding the linear relationship between height and percent body fat. Based on a sample of $n = 252$ men, the observed test statistic is $t = -1.42$, and p-value is $p_{obs} = 2P(T > |-1.42|) = 0.16$ based on the t-distribution with 250 degrees of freedom

`Pearson product-moment` for the `Type of Correlation`, as shown in Fig. 8.11. Then choose the `Two-sided` as the `Alternative Hypothesis`.

The results are given in the *Output* window (Fig. 8.12). The value of the observed test statistic shown as `t = -1.42`, and the degrees of freedom are 250. The observed significance level is $p_{obs} = 2P(T > 1.42) = 0.16$.

As we can see in Fig. 8.12, R-Commander also provides the 95% confidence interval for the population correlation coefficient ρ. For this example, the 95% confidence interval is $[-0.21, 0.03]$. Therefore, we are 95% confident that the true value of the correlation coefficient is between -0.21 and 0.03. Note that the interval includes negative and positive values. More specifically, it includes 0, which is the value of ρ according to the null. This is consistent with the result of our hypothesis testing, where we failed to reject the null hypothesis.

8.5 Advanced

In this section, we review some useful R functions for testing hypotheses related to the relationship between two variables.

8.5.1 Two-Sample t-test Using R

For two-sample t-test, we use the function `t.test()`. For example, using the `birthwt` data set, we can examine whether smoking during pregnancy and birthweight are related:

```
> t.test(bwt ~ smoke, mu = 0, alternative = "two.sided",
+       data = birthwt)

        Welch Two Sample t-test

data:   bwt by smoke
t = 2.7299, df = 170.1, p-value =
0.007003
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
  78.57486 488.97860
sample estimates:
mean in group 0 mean in group 1
      3055.696        2771.919
```

The first argument to the `t.test()` function is the "formula" specifying the response variable and the factor (explanatory) variable in the form of `response ~ factor`. In this case, the response variable is `bwt`, and the factor is `smoke`. We are using the `data=birthwt` option to avoid having to write `birthwt$bwt ~ birthwt$smoke`. The `mu` option is used to specify the difference in the population means according to the null hypothesis.

When the observations in the two groups are related (paired), we need use the paired t-test. For example, suppose our alternative hypothesis is that platelet aggregation is lower before smoking than after, $H_A : \mu < 0$ versus $H_0 : \mu = 0$. In R, we still use the function `t.test()` to examine the support for these hypotheses, but this time, we set the argument `paired` to `TRUE`:

```
> t.test(Platelet$Before, Platelet$After,
+       alternative = "less", paired = TRUE)

        Paired t-test

data:   Platelet$Before and Platelet$After
t = -4.2716, df = 10, p-value =
0.0008164
alternative hypothesis: true difference in means is
less than 0
95 percent confidence interval:
      -Inf -5.913967
sample estimates:
```

```
mean of the differences
             -10.27273
```

The first argument to the `t.test()` function provides the first group of observations, and the second argument provides the second group of observations.

Removing the `paired=TRUE` option would ignore the dependence between the observations in the two groups (in other words, we would use the independent two-sample *t*-test):

```
> t.test(Platelet$Before, Platelet$After,
+     alternative = "less")

        Welch Two Sample t-test

data:  Platelet$Before and Platelet$After
t = -1.4164, df = 19.516, p-value =
0.08621
alternative hypothesis: true difference in means is
less than 0
95 percent confidence interval:
     -Inf 2.251395
sample estimates:
mean of x mean of y
 42.18182  52.45455
```

The results are very different. Ignoring the dependence between observations is inappropriate and might result in wrong conclusions.

8.5.2 Correlation Test Using R

To test hypotheses about a linear relationship between two numeric variables, we use Pearson's correlation coefficient and the `cor.test()` function in R. The following code examines whether percent body fat and abdomen circumference from the `bodyfat` data set are positively correlated, $H_A : \rho > 0$ versus $H_0 : \rho = 0$:

```
> cor.test(bodyfat$siri, bodyfat$abdomen,
+     alternative = "greater")

        Pearson's product-moment correlation

data:  bodyfat$siri and bodyfat$abdomen
t = 22.1117, df = 250, p-value <
2.2e-16
alternative hypothesis: true correlation is greater
than 0
```

Table 8.3 Contingency table of heart attack by the type of treatment

	Heart attack	No heart attack
Placebo	189	10845
Aspirin	104	10933

```
95 percent confidence interval:
 0.77505 1.00000
sample estimates:
      cor
0.8134323
```

The arguments to the `cor.test()` function are the two random variables, and the `alternative="greater"` option specifies the $H_A : \rho > 0$. As before, the other options are "two.sided" and "less".

8.6 Exercises

1. Use the `Pima.tr` to find the difference between the sample means of diastolic blood pressure for diabetic and nondiabetic Pima Indian women. Is the difference between the means of diastolic blood pressure statistically significant at 0.01 level?
2. Answer the above question for the number of pregnancies and BMI.
3. In Sect. 3.4, we discussed a study comparing ascorbic acid (one form of vitamin C) content between two different cultivars, c39 and c52, of cabbage. The data set `cabbages` is available from the `MASS` package. Use an appropriate hypothesis testing procedure to examine the relationship between the vitamin C content and cultivars.
4. Charles Darwin (1809–1882), the author of The Origin of Species (1859) investigated the effect of cross-fertilization on the size of plants. The Data and Story library (http://lib.stat.cmu.edu/DASL/Stories/student.html) has the results of one of his experiments (given by R.A. Fisher). In this experiment, pairs of plants, one cross- and one self-fertilized, were planted and grown in the same plot. The following table gives the difference in height (eighths inches) for 15 pairs of plants (cross-fertilized minus self-fertilized Zea mays) raised by Charles Darwin. Use this data to evaluate the null hypothesis that the two methods are not different.

   Difference: 49, −67, 8, 16, 6, 23, 28, 41, 14, 29, 56, 24, 75, 60, −48

5. Consider the contingency table (Table 8.3) based on the study conducted to investigate whether taking aspirin reduces the risk of heart attack. Evaluate the null hypothesis that there is no relationship between taking aspiring and the risk of heat attack.

6. Use the `birthwt` data set to examine the relationship between hypertension history (`ht`) and the risk of having low-birthweight baby (`low`).

7. In Sect. 3.6, we used the `GBSG` (German Breast Cancer Study Group) data set from the `mfp` package to create a new variable called `rfs` (recurrence-free survival) such that `rfs="No"` if the patient had at least one recurrence or died (i.e., `cenc=1`) and `rfs="Yes"` otherwise. Use the data to investigate whether recurrence-free survival is related to hormonal therapy. (In `GBSG`, the variable `htreat` indicates whether a patient has received hormonal therapy or not.)

8. For the Pima Indian women population, find the sample correlation coefficient between BMI and diastolic blood pressure. Is the correlation between these two variables statistically significant at 0.01 level?

9. Use the "BodyTemperature.txt" to estimate the correlation coefficient between normal body temperature and heart rate. Is the correlation between these two variables statistically significant at 0.01 level? How about the correlation between age and normal body temperature?

10. Read the article "Caloric restriction improves memory in elderly humans" by Witte et al. [39]. (This paper is available online at http://www.pnas.org/content/106/4/1255.full.) What was their estimate of the correlation coefficient between memory score and insulin level? Was the correlation statistically significant at 0.1 level?

11. Read the paper "A Critical Appraisal of 98.6°F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich" by Mackowiak et al. [19]. (The paper is available online at http://jama.ama-assn.org/cgi/reprint/268/12/1578.) What method they used to evaluate the relationship between gender and body temperature? What did they find? What was their conclusion about the relationship between race and body temperature?

12. Read the paper by Kettunen et al. [14] on the effect of arthroscopy in patients with chronic patellofemoral pain syndrome. (This paper is available online at http://www.biomedcentral.com/1741-7015/5/38.)

    (a) What is the point estimate and 95% confidence interval for the mean improvement in the Kujala score for each treatment group.

    (b) Was the difference between the two group in terms of mean improvement in the Kujala score statistically significant?

    (c) Based on the results published in this paper, create a contingency table, where the row variable is the treatment group, and the column variable is an indicator that is equal to 1 if the patient reports at least moderate improvement at the end of follow-up period and 0 otherwise. Investigate the relationship between the type of treatment and reporting at least moderate improvement.

Chapter 9
Analysis of Variance (ANOVA)

9.1 Introduction

In Chap. 8, we discussed how two-sample t-tests can be used to evaluate hypotheses regarding the difference between the means of two groups. We mentioned that we typically use this approach to investigate the relationship between a binary categorical (factor) variable, which specifies the two groups, and a numerical variable, which is regarded as the response variable. In this chapter, we discuss Analysis of Variance (ANOVA) models that generalize the t-test and are used to compare the means of multiple groups identified by a categorical variable with more than two possible categories. As before, the categorical variable is called the **factor** and is typically considered as the explanatory variable. In contrast, the numerical variable, whose means across different groups are compared, is regarded as the response variable.

In this chapter, we mainly focus on ANOVA models with only one factor. These models are known as **one-way ANOVA**. In Advanced section, we briefly discuss **two-way ANOVA** models that include two factors (i.e., two categorical explanatory variables) in the analysis.

9.2 Analysis of Variance (ANOVA)

As an example, we analyze the `Cushings` data set [2], which is available from the `MASS` package. Cushing's syndrome is a hormone disorder associated with high level of cortisol secreted by the adrenal gland. The `Cushings` data set includes 27 observations ($n = 27$). For each individual in the sample, the urinary excretion rates of two steroid metabolites are recorded. These are urinary excretion rate (mg/24 hr) of Tetrahydrocortisone and urinary excretion rate (mg/24 hr) of Pregnanetriol. The `Type` variable in the data set shows the underlying type of syndrome, which can be one of four categories: adenoma (a), bilateral hyperplasia (b), carcinoma (c), and unknown (u).

Fig. 9.1 Viewing the
`Cushings` data set in
R-Commander. For 27
individuals, the urinary
excretion rates of two steroid
metabolites and the
underlying type of syndrome
are recorded

To load the data in R-Commander, click `Data` → `Data in packages` → `Read data set from an attached package`. Select `MASS` under `Package` and `Cushings` under `Data set`. Figure 9.1 shows a small part of the data. The highlighted observation in row "c4" is an outlier with `Tetrahydrocortisone = 53.8`, which is much higher than typical values observed for the variable `Tetrahydrocortisone`. We should further investigate this observation and remove it only if we are convinced that it was recorded by mistake and we cannot recover the correct values. In what follows, we assume that this is a legitimate observation, so we include it in our analysis. (If we decide to remove an outlier that is recorded by mistake, we can do so by clicking `Data` → `Active data set` → `Remove row(s) from active data set` and entering the row name, e.g., 'c4' (with quotations), under `Indices or quoted names of row(s) to remove`.)

Our objective is to find whether the four groups are different with respect to urinary excretion rate of Tetrahydrocortisone. We denote by Y the urinary excretion rate of Tetrahydrocortisone and by X the `Type` variable, where $X = 1$ for `Type=a`, $X = 2$ for `Type=b`, $X = 3$ for `Type=c`, and $X = 4$ for `Type=u`. Then, our objective could be defined as investigating whether the *mean* of the response variable Y differs for different values (levels) of the factor X.

Denote the individual observations as y_{ij}: the urinary excretion rate of Tetrahydrocortisone of the jth individual in group i. The total number of observations is $n = 27$, and the number of observations in each group is $n_1 = 6$, $n_2 = 10$, $n_3 = 5$, and $n_4 = 6$. The overall (for all groups) observed sample mean for the response variable is $\bar{y} = 10.46$. We also find the group specific means, which are $\bar{y}_1 = 3.0$, $\bar{y}_2 = 8.2$, $\bar{y}_3 = 19.7$, and $\bar{y}_4 = 14.0$. You can find the group specific means by clicking `Statistics` → `Summaries` → `Numerical summaries`. Then select `Tetrahydrocortisone` under `Variables`, and select `Type` by clicking `Summarize by groups`.

Now, consider the dot plot of Y by X in Fig. 9.2. Here, each observation is represented by a point, and the overall average $\bar{y}$ of the response variable is represented by the dashed horizontal line at 10.46. Likewise, the sample average $(\bar{y}_1, \ldots, \bar{y}_4)$ for each group is shown as a small horizontal line.

Across the four groups, there appears to be considerable variation in the group means (i.e., deviations of the small solid lines from the dashed line). Likewise, within groups, there are different degrees of variation of the observations from their specific mean (i.e., variation of points around the corresponding small horizontal

Fig. 9.2 Strip chart of
`Tetrahydrocortisone`
by syndrome type. The
overall sample mean $\bar{y}$ of
`Tetrahydrocortisone`
for all groups is shown as the
dashed horizontal line, and
the sample average $\bar{y}_i$ for
each group is shown as the
small horizontal line
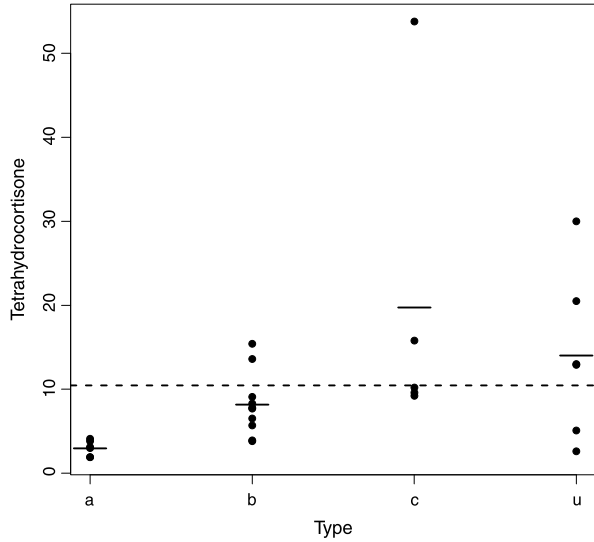
line). Both sources of variation contribute to the total variation of the observations around the overall mean (dashed line).

> In general, the **between-groups variation** is denoted as SS_B and calculated by
>
> $$SS_B = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2, \tag{1}$$
>
> where k is the number of groups (here, 4).

To find SS_B, we first find the squared difference between each group mean (i.e., the solid short lines) and the overall mean (i.e, the dashed line). In order to account for varying sample sizes, the squared distance is then multiplied by the number of observations in that group, n_i. (Therefore, groups with more observations are weighted more heavily.) The sum of these squared and weighted differences is the between-groups variation.

> The **within-groups variation** is denoted as SS_W and calculated by
>
> $$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

To find SS_W, we first calculate the sum of squared deviations of each observation (i.e., the point) from the group mean (i.e., the short horizontal line) for each group separately. Then we add the results over all groups.

We measure the **total variation** in Y by

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 .$$

To find SS, we find the sum of squared distances of each observation to the overall average (i.e., the dashed line). It seems intuitive and can be shown that the total variation SS is equal to the sum of the between-groups variation SS_B and the within-groups variation SS_W,

$$SS = SS_B + SS_W .$$

In other words, the total variation can be attributed partly to the variation within groups and partly to the variation between groups. SS_B is interpreted as the part of total variation SS that is associated with (and can be explained by) the factor variable X (e.g., syndrome type). In contrast, SS_W is regarded as the unexplained part of total variation and is regarded as random.

In our example, if `Tetrahydrocortisone` does not depend on the type of syndrome, we expect the group-specific averages to be the same. That is, we expect the solid lines to lie on the dashed line and any observed variation of solid lines around the dashed line to be due to chance alone. On the other hand, if there is a substantial difference in `Tetrahydrocortisone` depending on the type of syndrome, then we would expect the variation between groups to be large. We examine the amount of between-groups variation relative to the variation within groups (which occurs randomly).

Let us denote the overall population mean of Y as μ and group-specific population means as $\mu_1, \ldots, \mu_4$. Then we can express the null hypothesis of no difference in means between the groups as

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu .$$

That is, if we had data for the whole population, the small solid lines would lie on the dashed line. The alternative hypothesis H_A is that at least one of the group means μ_i is different from the overall mean μ.

Fig. 9.3 Comparing the plots of the probability density function for an F-distribution with various degrees of freedom. The *solid line* represents the pdf of $F(1, 1)$, the *dashed line* represents the pdf of $F(2, 5)$, and the *dotted line* represents the pdf of $F(10, 20)$
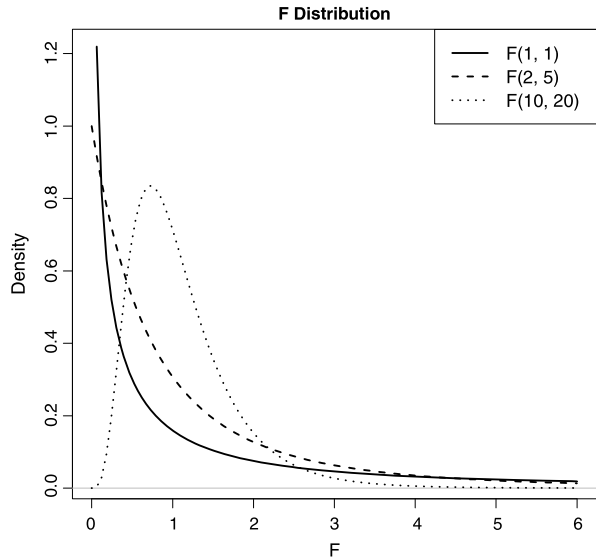
The process of evaluating hypotheses regarding the group means of multiple populations is called the **Analysis of Variance (ANOVA)**. Since we are only considering one factor only, this method is specifically called **one-way ANOVA**. The test statistic for examining the null hypothesis is called F-**statistic** (more specifically, ANOVA F-statistic) and is defined as

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)},$$

where n is the total sample size, and k is the number of groups. The numerator is called the **mean square for groups**, and the denominator is called the **mean square error** (MSE).

Note that the above test statistic is based on comparing the variation between groups (which is explained by the factor) and the variation within groups (which is unexplained and random). When the group means are substantially different, and their variation is relatively large compared to the random variations within groups, the value of the F statistic becomes large.

We denote the observed value of the F-statistic as f. If the null hypothesis is true, then the test statistic F has an F-distribution.

The F-distribution, which is a continuous probability distribution, is very important for hypothesis testing. It is specified by two parameters, df_1 and df_2, and is denoted as $F(df_1, df_2)$. We refer to df_1 and df_2 as the *numerator degrees of freedom* and *denominator degrees of freedom*, respectively. Both parameters must be positive. Figure 9.3 shows the pdf of F-distribution for different values of df_1 and df_2.

For the one-way ANOVA, the F-statistic has $F(df_1 = k - 1, \; df_2 = n - k)$ distribution under the null hypothesis (i.e., assuming that the null hypothesis is true). Here, $df_1 = k - 1$, which is the number of groups minus 1, is called the numerator degrees of freedom, and $df_2 = n - k$, which is the sample size minus the number of groups, is called the denominator degrees of freedom. The underlying assumption here is that the observations in each group are IID (e.g., obtained through SRS) and have a normal distribution. The results are not sensitive to the normality assumption as long as the sample sizes are large enough for the CLT to hold.

Additionally, the underlying assumption of the ANOVA method discussed here is that all groups have the same population variance, σ^2, which is unknown.

For the above example, the degrees of freedom parameters are $df_1 = 4 - 1 = 3$ and $df_2 = 27 - 4 = 23$. Try plotting the $F(3, 23)$ distribution using R-Commander. Click `Distribution` $\rightarrow$ `Continuous distributions` $\rightarrow$ `F distribution Plot F distribution`. Set the `Numerator degrees of freedom` to 3 and the `Denominator degrees of freedom` to 23.

When the group means are very different from each other, the between-groups variation SS_B is high. As a result, the F-statistic is large. Therefore, large values of the F-statistic are considered as extreme if the null hypothesis is true. Therefore, large values of F provide strong evidence against the null hypothesis. To find the observed significance level p_{obs}, we find the probability of values as or more extreme than the observed value of the test statistic, f. For this, we calculate the upper tail probability of f based on an $F(df_1 = k - 1, df_2 = n - k)$ distribution. For the above example, the observed value of the test statistic is $f = 3.2$. Therefore, the p-value is $p_{obs} = P(F \geq 3.2)$, which is shown as the shaded area in Fig. 9.4.

To calculate the p-value in R-Commander, click `Distributions` $\rightarrow$ `Continuous distributions` $\rightarrow$ `F distribution` $\rightarrow$ `F probabilities`. Then enter 3.2 for `Variable value`, 3 for `Numerator degrees of freedom`, and 23 for `Denominator degrees of freedom`. Also make sure the `Upper tail` option is selected. The result, given in the *Output* window, suggests that there is moderate evidence against the null hypothesis: $p_{obs} = P(F \geq 3.2) = 0.04$. Therefore, we can reject H_0 at 0.05 significance level (but not at 0.01) and conclude that the differences among group means for urinary excretion rate of Tetrahydrocortisone are statistically significant (at 0.05 level).

Using R-Commander, we can directly perform the Analysis of Variance. Click `Statistics` $\rightarrow$ `Means` $\rightarrow$ `One-way ANOVA`. Select `Tetrahydrocorti-sone` as the `Response Variable`. (Notice how R-Commander correctly identifies `Type` as the `Factor`.) The results of the analysis of variance are presented as a table called the *ANOVA table* (Fig. 9.5). The first row of this table is for the group variable (`Type`) and shows the explained part of the total variation (i.e., between groups). The last row (`Residuals`) shows the unexplained part (i.e., ran-

Fig. 9.4 The density plot of $F(3, 23)$-distribution. This is the distribution of F-statistic for the `Cushings` data assuming that the null hypothesis is true. The observed value of the test statistic is $f = 3.2$, and the corresponding p-value is shown as the *shaded area* above 3.2
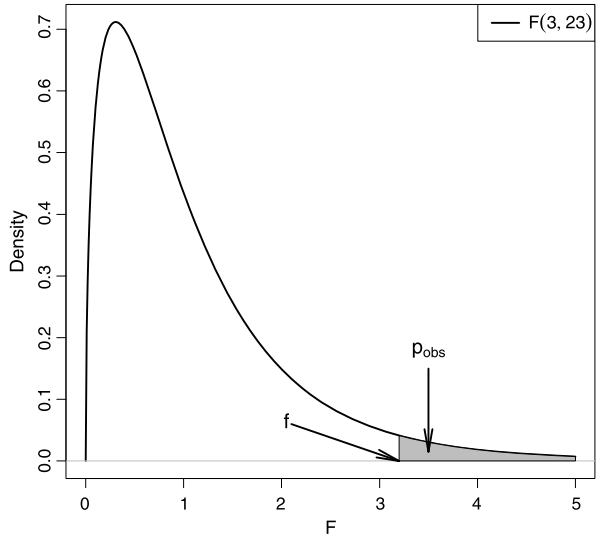

```
Output Window                                                Submit

> AnovaModel.1 <- aov(Tetrahydrocortisone ~ Type, data=Cushings)

> summary(AnovaModel.1)
            Df  Sum Sq Mean Sq F value  Pr(>F)
Type         3  893.52 297.840  3.2257 0.04122 *
Residuals   23 2123.65  92.332
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> numSummary(Cushings$Tetrahydrocortisone , groups=Cushings$Type,
+    statistics=c("mean", "sd"))
      mean        sd  n
a  2.966667  0.9244818  6
b  8.180000  3.7891072 10
c 19.720000 19.2388149  5
u 14.016667 10.0958242  6
```

Fig. 9.5 The ANOVA table resulting from the hypothesis test regarding the mean `Tetrahydrocortisone` of various syndrome types. Specifically, the null hypothesis is that there is no difference in the group means. The *first row* corresponds to the factor, and the *second* to the residuals

dom variations within groups) of the total variation in the data . The first column shows the degrees of freedom (`Df`), which are $k - 1 = 3$ and $n - k = 23$, respectively. The values of the second column, labeled `Sum Sq`, are the between-groups and within-groups variations: $SS_B = 893.5$ and $SS_W = 2123.6$. The observed value of F-statistic is $f = 3.2$ given under the column labeled `F value`. The resulting p-value is then 0.04. Below the ANOVA table, R-Commander provides the group-specific means, the group-specific standard deviations, and the number of observations in each group, n_i.

In Fig. 9.7, we showed the dot plot of `Tetrahydrocortisone` by syndrome type. Alternatively, we can use R-Commander to create the plot of means as described in Chap. 3. For this, click `Graphs` → `Plot of means` and select `Type` as the `Factors` and `Tetrahydrocortisone` as the `Response Variable`

Fig. 9.6 Creating a plot of means for the `Tetrahydrocortisone` by syndrome type from the `Cushings` data set. Here, we choose `Confidence intervals` at 0.95 level of confidence for `Error Bars`. This way, R-Commander shows the 95% confidence interval around the sample mean for each group
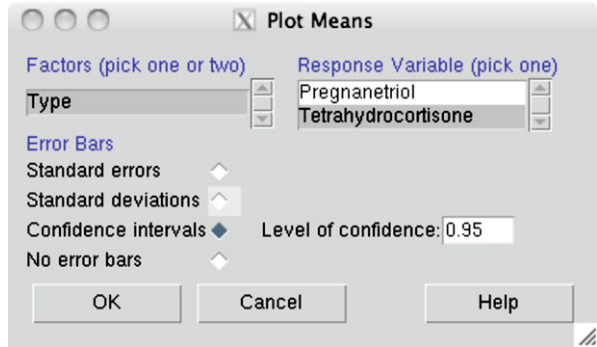
Fig. 9.7 Plot of means for `Tetrahydrocortisone` by syndrome type. The *points* show the location of the sample mean for the corresponding syndrome type. The *bars* show the 95% confidence intervals around the sample means
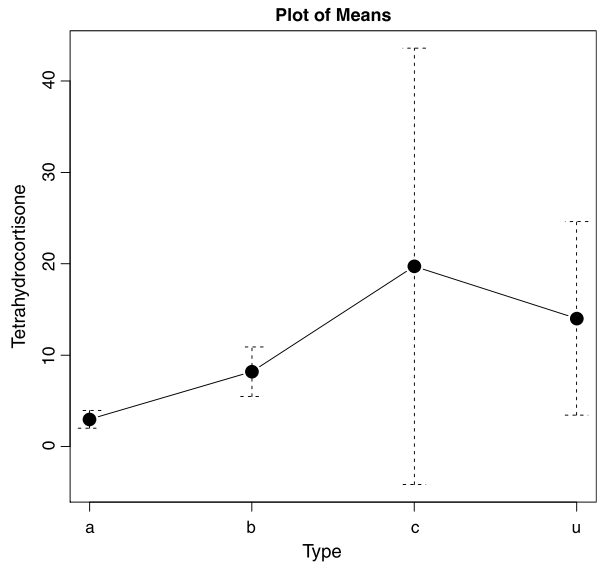
(Fig. 9.6). In Sect. 3.4, we chose `No error bars`. Now that we learned about confidence intervals, we can choose the option `Confidence intervals` instead. This way, R-Commander uses dashed horizontal lines to show the corresponding confidence interval around the sample mean of each group as shown in Fig. 9.7. Based on this graph, the type "a" syndrome has the lowest sample mean. The sample means increase from type "a" to type "c", and then it slightly drops for type "u".

9.3 The Assumptions of ANOVA

To use ANOVA models, we assume that the samples are selected randomly from the population and independently from each other (e.g., by using simple random

Fig. 9.8 Plot of means for the log of `Tetrahydrocortisone` by syndrome type
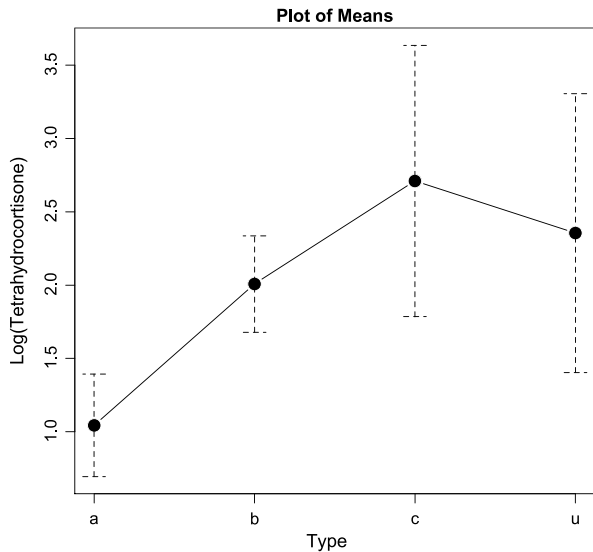
sampling). Further, we assume that the response variable in each group has a normal distribution. While the means of these normal distributions can change from one group to another, we assume that they all have the same variance. This is the same assumption we used for pooled t-tests.

Violation of these assumption could lead to wrong inference. The independence assumption is violated, for example, if we obtain multiple observations (e.g., over a period of time) for each subject in our sample. In this case, we need to use **repeated-measures ANOVA** (not discussed in this book), which can be regarded as a generalization of paired t-tests. The consequence of violating the normality assumption is not very severe as long as the sample sizes for all the groups are large enough so the distributions of the sample means are approximately normal due to the central limit theorem. Finally, the assumption of equal variance among all groups is usually unrealistic and is often violated in practice. This is clearly the case for the example discussed in this chapter (Fig. 9.2). In practice, you are likely to see problems more similar this example (maybe not ad sever) that problems where the assumption of equal variance holds. Similar to the normality assumption, the results of ANOVA are not severely affected if the group variances moderately differ from each other. Alternatively, we could use ANOVA models without the equal variance assumption. See Weerahandi (2003) [38] for example.

Sometimes, we can stabilize the variance (i.e., making it approximately constant) by using simple data transformations such as log or square root. For the example discussed above, using the log-transformation of Tetrahydrocortisone (instead of the original variable) makes the equal variance assumption more reasonable (Fig. 9.8). In R-Commander, create a new variable by taking the log of Tetrahydrocortisone. Then, repeat the steps discussed above to perform analysis of variance for this newly created variable. In this case, the observed value of F-statistic is $f = 7.6$, and the corresponding p-value is 0.001.

9.4 Advanced

In this section, we briefly discuss ANOVA with two factors. We also provide some useful R functions to perform ANOVA.

9.4.1 Two-Way ANOVA

Consider the study by Bailey (1953) to investigate the inheritance of maternal influences on the growth of the rat [29]. In this study, rat litters were separated from their natural mothers, and they were nurtured by foster mothers. Mothers and litters can have four different genotypes: A, B, I, and J. In R-Commander, load the `geno-type` data set from the `MASS` library. Suppose that we want to investigate whether weight gain (`Wt`) of the litter (in grams) at age 28 days is related to foster mother's genotype (`Mother`). (Rat litters were separated from their natural mothers at birth and given to foster mothers.) For this, we could use the one-way ANOVA procedure to compare the means of weight gain across different groups (genotypes). That is, we regard `Wt` as the response variable and `Mother` as the factor. For this example, however, we might want to take into account the genotype of rat litters (`Litter`) as well. The litter's genotype is itself a factor, and even though it is not the main factor in this study, it should be included in the analysis since we believe that it could influence the relationship between the main factor, mother's genotype, and the response variable, weight gain.

An ANOVA with two factors is called a **two-way ANOVA**. (In general, we can have a multi-way ANOVA by including multiple factors.) In many two-way ANOVA procedures, one of the two factors is the main explanatory variable of interest. The other factor is included since it is believed to be important in the study of the relationship between the main factor and the response variable. This is the case for the "rat genotype" example. In this example, we are mainly interested in the variation of weight gain across different genotypes of mothers. However, we need to account for possible weight gain variation due to the genotype of the litters. By including both factors `Mother` and `Litter`, we are dividing the total variation, SS, into three sources: (1) variation explained by the mother's genotype, SS_M, (2) variation explained by the litter's genotype, SS_L, and (3) the random variation, SS_E, of weight gain not explained by either mother's genotype or litter's genotype. (Note that we have switched our notation from SS_W to SS_E.) So,

$$SS = SS_M + SS_L + SS_E.$$

This type of two-way ANOVA is commonly used for experiments with a randomized block design discussed in Sect. 1.7. For these experiments, the treatment variable is the factor whose effect on the response variable is of main interest. The categorical variable used for blocking is the factor which is believed to be important, but its relationship with the response variable is not the focus of the experiment.

In the next subsection, we show how two-way ANOVA models can be implemented in R. Here, we discuss how R-Commander can be used for two-way ANOVA

```
Output Window                                                    Submit

Anova Table (Type II tests)

Response: Wt
              Sum Sq Df F value   Pr(>F)
Litter         63.63  3  0.3911 0.760004
Mother        775.08  3  4.7632 0.005736 **
Litter:Mother 824.07  9  1.6881 0.120053
Residuals    2440.82 45
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
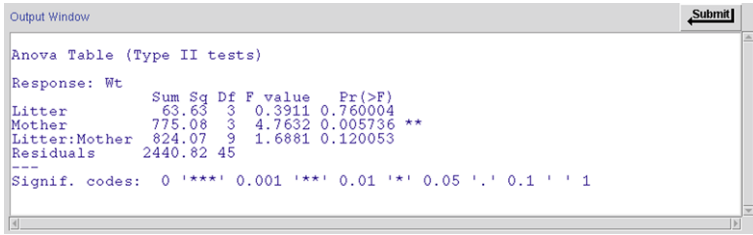
Fig. 9.9 The output of two-way ANOVA for "rat genotype". Here, the weight gain is the response variable; the two factors are mother's genotype and litter's genotype

problems, where both factors are considered to be important, and we are interested in learning how the relationship between one factor with the response variable changes depending on the value of the other factor. For example, we might be interested in the effect of three different diets, A, B, and C, on blood pressure, but we believe that the diet effect varies between male and female groups. In this case, we say that there is an **interaction** between the two factors diet and gender.

For the "rat genotype" example, suppose we believe that the relationship between mother's genotypes and weight gain changes depending on litter's genotype. Therefore, we need to consider possible interaction between Mother and Litter. We use $M \times L$ to denote this interaction. Including the interaction between the two factors in a two-way ANOVA means that we believe a part of the total variation SS is explained by the combination of the two factors. That is, we can write the total variation as follows:

$$SS = SS_M + SS_L + SS_{M \times L} + SS_E.$$

The variation in the response variable due to specific combinations of the two factors is usually referred to as the **interaction effect**. In contrast, the variation in the response variable due to one of the factors alone (i.e., regardless of the values of the other factor) is called the **main effect**.

Now we use R-Commander to perform two-way ANOVA for the "rat genotype" example. Make sure genotype (available from the MASS package) is the active data set and then click Click Statistics → Means → Multi-way ANOVA. Under Factors, select both Mother and Litter. The response variable Wt is automatically selected since it is the only numerical variable in this data set. The ANOVA table appears in the Output window and is shown in Fig. 9.9.

By default, R-Commander includes the interaction between the two factors into the ANOVA model. You can use R directly to perform two-way ANOVA without including interaction. This is discussed in the next subsection.

The interpretation of sum of squares, degrees of freedom, F-statistic, and p-value is similar to one-way ANOVA. In this example, $SS_M = 775.08$, $SS_L = 63.63$, and $SS_{M \times L} = 824.07$. Based on these results, only the relationship between mother's genotype and weigh gain is statistically significant at 0.05 level ($p_{\text{obs}} = 0.006$). The interaction effect (shown as Litter:Mother) and the main effect of litter's genotype are not statistically significant at 0.05 level.

9.4.2 ANOVA Using R

After we obtain f, the observed value of the F-statistic, we can use the F-distribution to obtain the corresponding p-value by calculating the upper tail probability of f. The functions $df()$, $pf()$, and $qf()$ provide the density, tail probability, and quantiles from the F-distribution with given degrees of freedom. For the Cushings examples, $f = 3.2$, and the degrees of freedom are $df_1 = 3$ and $df_2 = 23$. The upper tail probability of 3.2 is obtained as follows:

```
> pf(3.2, df1 = 3, df2 = 23, lower.tail = FALSE)

[1] 0.04226148
```

Note that we need to set the lower.tail to "FALSE" to obtain the upper tail probability. (The default is `lower.tail=TRUE`.)

Alternatively, we can use the function `aov()` to perform ANOVA directly. For this, we specify the response and factor variables using the same formula notation we used for the t-test: `response ~ factor`. Here, the response variable is `Tetrahydrocortisone`, and the factor is `Type`:

```
> library(MASS)
> data(Cushings)
> aov1.out <- aov(Tetrahydrocortisone ~ Type,
+       data = Cushings)
```

The output of ANOVA is assigned to the object `aov.out`. We can create the ANOVA table by applying the `summary()` function to this object:

```
> summary(aov1.out)

            Df  Sum Sq Mean Sq F value Pr(>F)
Type         3  893.52 297.840  3.2257  0.04122 *
Residuals   23 2123.65  92.332

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

For two-way ANOVA, we use also use `aov()`, but the right side of the formula includes both factors. We separate the two factors by the "+" sign if we do not want to include their interaction. For the "rat genotype" example discussed in Sect. 9.4.1, we use the formula `Wt ~ Mother + Litter`:

```
> library(MASS)
> data(genotype)
> aov2.out <- aov(Wt ~ Mother + Litter,
+       data = genotype)
```

```
> summary(aov2.out)

          Df Sum Sq Mean Sq F value Pr(>F)
Mother     3  771.6 257.202  4.2540  0.009055 **
Litter     3   63.6  21.211  0.3508  0.788698
Residuals 54 3264.9  60.461

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

To include the interaction of the two factors in the ANOVA model, we use "*" instead of "+".

```
> library(MASS)
> data(genotype)
> aov2.int.out <- aov(Wt ~ Mother * Litter,
+      data = genotype)
> summary(aov2.int.out)

              Df  Sum Sq Mean Sq F value Pr(>F)
Mother         3  771.61 257.202  4.7419  0.005869 **
Litter         3   63.63  21.211  0.3911  0.760004
Mother:Litter  9  824.07  91.564  1.6881  0.120053
Residuals     45 2440.82  54.240

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

9.5 Exercises

1. We would like to investigate the effectiveness of various feed supplements (feed) on the growth rate (weight) of chickens. In R-Commander, load the chickwts data set from the datasets package. (Click Data → Data in packages → Read data set from an attached package.) Use boxplots and a plot of means to visualize the difference between feed types. Use ANOVA to examine the effectiveness of feed supplements. Comment on your findings and appropriateness of your assumptions.
2. We believe that mean urinary excretion rate of Pregnanetriol changes based on the underlying type of Cushing's syndrome. In R-Commander, load the Cushings data set from the MASS library and investigate whether there is statistically significant mean difference for this steroid metabolites.
3. For the "rat genotype" data discussed in Sect. 9.4.1, use one-way ANOVA to investigate whether weight gain (Wt) of the litter (in grams) at age 28 days is

related to mother's genotype (`Mother`). Repeat the analysis for the relationship between weight gain (`Wt`) and genotype of the litter (`Litter`). Compare the plot of means for the first analysis to that of the second one.

4. Load the `anorexia` data set from the `MASS` package. This data set was collected to investigate the effectiveness of different treatments (`Treat`) on increasing weight for young female anorexia patients. Create a new variable called `Difference` by subtracting the weight of patient before study period (`Prewt`) from her weight after the study period (`Postwt`): `Difference = Postwt - Prewt`. Use a plot of means to visualize how this variable changes depending on the type of treatment. Use ANOVA to investigate whether the type of treatment makes a difference in the amount of weight gain.

5. In Sect. 3.4, we discussed a study comparing ascorbic acid (one form of vitamin C) content between two different cultivars (c39 and c52) of cabbage. The data set `cabbages` for this example is available from the `MASS` package. In this data set, the two different cultivars were planted on three different dates, denoted as d16, d20, or d21. The variable `Data` is a factor that specifies the planting date for each cabbage. Use two-way ANOVA to evaluate the relationship between the vitamin C content and cultivars while controlling for the effect of planting dates.

6. Obtain the "Stepping and Heart Rates" data set from the Data and Story Library (http://lib.stat.cmu.edu/DASL/Datafiles/Stepping.html). The data are obtained based on an experiment conducted by students at the Ohio State University to investigate possible relationship between a person's heart rate and the frequency at which that person stepped up and down on steps of various heights. Create a new variable called `diffHR` whose values are the increase in heart rate of the subjects after a trial compared to their resting heart rate before a trial. Use two-way ANOVA to evaluate effect of the rate of stepping (`Frequency`) and the height of the steps (`Height`) on `diffHR`.

Chapter 10
Analysis of Categorical Variables

10.1 Introduction

In Chap. 7, we talked about hypothesis testing regarding population proportions. There, we used the central limit theorem (for large enough sample sizes) to obtain an approximate normal distribution of the sample proportion, which we used as the test statistic. We followed a similar approach in Chap. 8 in order to test hypotheses regarding the relationship between two binary random variables.

In this chapter, we discuss **Pearson's χ^2 (chi-squared) test** for testing hypotheses regarding the distribution of a categorical variable or the relationship between two categorical variables. Pearson's test evaluates whether the probabilities for different categories are equal to the values specified by the null hypothesis. Although it is not necessary, we can think of the probability of each category as its population proportion. This makes the discussion easier to follow. For example, when we talk about the probability of heart attack survival being 0.7, we can interpret this as 70% of heart attack patients (i.e., 70% of the entire population of people suffered from heart attack) survive. As before, we use the sample proportion of each category as a point estimate for its probability (i.e., its population proportion).

Pearson's χ^2 test uses a test statistic, which we denote as Q, to measure the discrepancy between the observed data and what we expect to observe under the null hypothesis (i.e., assuming the null hypothesis is true). Higher levels of discrepancy between data and H_0 results in higher values of Q. We use q to denote the observed value of Q based on a specific sample of observed data. As usual, we need to find the null distribution of Q (i.e., its sampling distribution assuming that H_0 is true) and measure the observed significance level p_{obs} by calculating the probability of values as or more extreme than the observed value q.

10.2 Pearson's χ^2 Test for One Categorical Variable

10.2.1 Binary Variables

We start our discussion of Pearson's method by focusing on binary random variables first. Then, we show how we can extend this approach for situations where categorical variables have more than two possible values.

Let us denote the binary variable of interest as X, based on which we can divide the population into two groups depending on whether $X = 1$ or $X = 0$. Further, suppose that the null hypothesis H_0 states that the probability of group 1 (i.e., the probability that an individual belongs to the group 1) is μ_{01} and the probability of group 2 is μ_{02}. Of course, because the sum of probabilities adds up to one, $\mu_{02} = 1 - \mu_{01}$. As a running example, we use the heart attack survival rate (i.e., the probability of survival after heart attack) within one year after hospitalization. Suppose that H_0 specifies that the probability of surviving is $\mu_{01} = 0.70$ and the probability of not surviving is $\mu_{02} = 0.30$.

If we take a random sample of size $n = 40$ from the population (people who suffer from heart attack), we expect that 70% of them survive and 30% of them die within one year from the time of hospitalization if in fact the null hypothesis is true. That is, we expect that $0.70 \times 40 = 28$ of subjects belong to the first group (survived) and $0.30 \times 40 = 12$ of subjects belong to the second group (nonsurvived).

> If the null hypothesis is true, we expect that, out of n randomly selected individuals, $E_1 = n\mu_{01}$ belong to the first group, and $E_2 = n(1 - \mu_0)$ belong to the second group. We refer to E_1 and E_2 as the **expected frequencies** under the null.

In our example, $E_1 = 28$ and $E_2 = 12$.

Now suppose that we randomly select 40 people who have suffered from heart attack. After one year from the time of hospitalization, we find that 24 of them have survived and 16 of them did not survive. We refer to the observed number of people in each group as the **observed frequencies** and denote them O_1 and O_2 for group 1 and group 2, respectively. In our example, $O_1 = 24$ and $O_2 = 16$.

> Pearson's χ^2 test measures the discrepancy between the observed data and the null hypothesis based on the difference between the observed and expected frequencies as follows:
>
> $$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}.$$
>
> We use q to denote the observed value of the test statistic Q.

Fig. 10.1 The plot of the pdf for a χ^2 distribution with various degrees of freedom
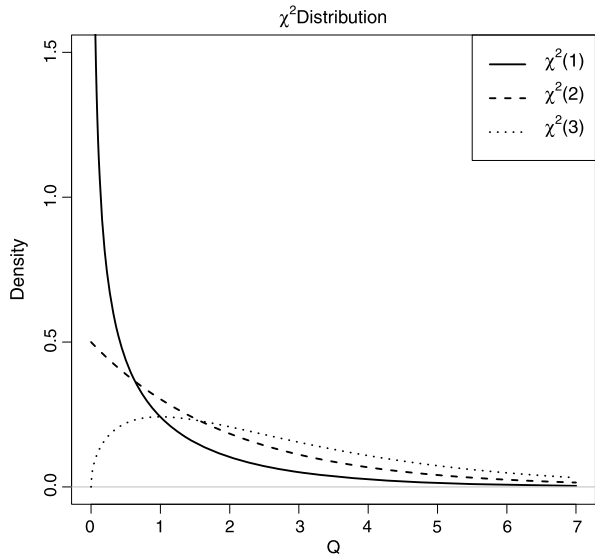
For the heart attack survival example, the observed value of the test statistic is

$$q = \frac{(24-28)^2}{28} + \frac{(16-12)^2}{12} = 1.90.$$

The value of Q will be zero only when the observed data matches our expectation under the null exactly. When there is some discrepancy between the data and the null hypothesis, Q becomes greater than zero. The higher discrepancy between our data and what is expected under H_0, the larger Q and therefore the stronger the evidence against H_0.

To evaluate the null hypothesis, we need to find the p-value, which is, as usual, the probability of observing as or more extreme values compared to the observed value of the test statistic. For this, we first need to find the sampling distribution of the test statistic Q under the null and calculate the probability of observing as large or larger values than q.

If the null hypothesis is true, then the approximate distribution of Q is χ^2. Like the t-distribution, the χ^2-distribution is commonly used for hypothesis testing. Also, similar to the t distribution, the χ^2 distribution is defined by its degrees of freedom df (which is a positive number) and is denoted $\chi^2(df)$. The pdf of the χ^2 distribution for various degrees of freedom is given in Fig. 10.1.

For binary random variables (i.e., when there are two possible groups), the approximate distribution of Q is $\chi^2(1)$ distribution. Try plotting this distribution using R-Commander. Click `Distribution` → `Continuous distributions` → `Chi-squared distribution Plot chi-squared distribution`. Set the `Degrees of freedom` to 1. The resulting distribution is shown in Fig. 10.2.

To evaluate the null hypothesis regarding the probabilities of two groups, we determine the observed significance level p_{obs} by calculating the probability of Q

Fig. 10.2 The sampling
distribution for Q under the
null hypothesis: $Q \sim \chi^2(1)$.
The p-value is the upper tail
probability of observing
values as extreme or more
extreme than $q = 1.90$
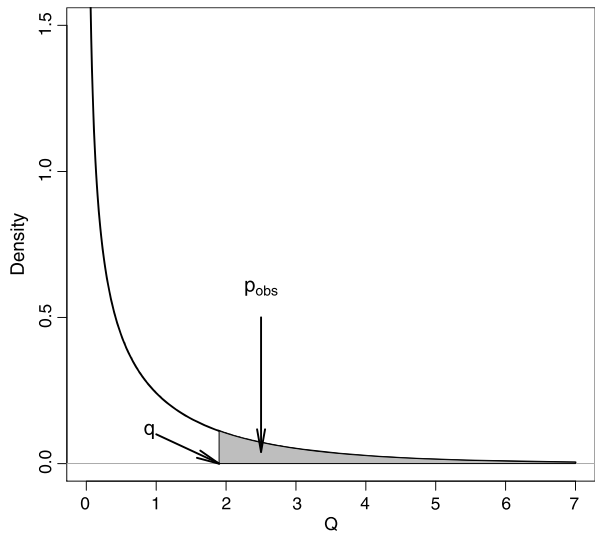
Fig. 10.3 Obtaining the
p-value from a χ^2
distribution with 1 degree of
freedom. In this example,
$p_{obs} = P(Q \geq 1.90) = 0.17$
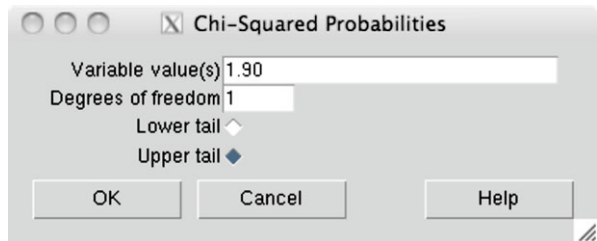
values as or more extreme than the observed value q using the χ^2 distribution with
1 degree of freedom. This corresponds to the upper tail probability of q from the
$\chi^2(1)$ distribution. For the heart attack survival example, where $q = 1.90$, this prob-
ability is shown as the shaded area in Fig. 10.2.

To obtain this probability in R-Commander, click `Distributions` →
`Continuous Distributions` → `Chi-squared distribution` →
`Chi-squared probabilities`. Then enter 1.90 for `Variable value` and
1 for `Degrees of freedom`, and select `Upper tail` (as in Fig. 10.3). The re-
sult, shown in the *Output* window, is the probability of observing values as extreme
or more extreme than 1.90 based on a χ^2 distribution with 1 degree of freedom. This
probability is $p_{obs} = 0.17$. Therefore, the results are not statistically significant, and
we cannot reject the null hypothesis at commonly used significance levels (e.g.,
0.01, 0.05, and 0.1). In this case, we believe that the difference between observed
and expected frequencies could be due to chance alone.

10.2.2 Categorical Variables with Multiple Categories

Pearson's χ^2 test can be generalized to situations where the categorical random variable can take more than two values. Let us reconsider the heart attack example. This time, suppose that we monitor heart attack patients for one year and divide them into three groups:

1. patients who did not have another heart attack and survived,
2. patients who had another heart attack and survived,
3. and finally patients who did not survive.

Now suppose that the probabilities of these three groups according to the null is $\mu_{01} = 0.5$, $\mu_{02} = 0.2$, and $\mu_{03} = 0.3$. That is, among 70% of patients who survive, 20% of them have another heart attack within a year from their first hospitalization. As before, we can find the expected frequencies of each category for a sample of $n = 40$ patients assuming that the null hypothesis is true:

$$E_1 = 0.5 \times 40 = 20, \qquad E_2 = 0.2 \times 40 = 8, \qquad E_3 = 0.3 \times 40 = 12.$$

This time, suppose that the actual observed frequencies based on a sample of size $n = 40$ for the three groups are

$$O_1 = 13, \qquad O_2 = 11, \qquad O_3 = 16.$$

Again, we measure the amount of discrepancy between the observed data and the null hypothesis based on the difference between the observed and expected frequencies:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3}.$$

For the heart attack survival example, the observed value of this test statistic is

$$q = \frac{(13 - 20)^2}{20} + \frac{(11 - 8)^2}{8} + \frac{(16 - 12)^2}{12} = 4.91.$$

In general, for a categorical random variable with I possible categories, we calculate the test statistic Q as

$$Q = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{E_i}.$$

The approximate distribution of Q is χ^2 with the degrees of freedom equal to the number of categories minus 1: $df = I - 1$. Therefore, to find p_{obs}, we calculate the upper tail probability of q (the observed value of Q) from the $\chi^2(I - 1)$ distribution.

For the above example, the p-value is the upper tail probability of 4.91 for a $\chi^2(2)$ distribution. Using R-Commander, we find $p_{obs} = P(Q \geq 8.67) = 0.086$ using the χ^2 distribution with 2 degrees of freedom. Therefore, we can reject the null

hypothesis at 0.1 level but not at 0.05 level. At the 0.1 significance level, we can conclude that the difference between observed and expected frequencies is statistically significant, and it is probably not due to chance alone.

10.3 Pearson's χ^2 Test of Independence

We now discuss the application of Pearson's χ^2 test for evaluating a hypothesis regarding possible relationship between two categorical variables. As before, we measure the discrepancy between the observed data and the null hypothesis. More specifically, we measure the difference between the observed frequencies and expected frequencies under the null. The null hypothesis in this case states that the two categorical random variables are independent. Recall that two random variables are independent if they do not affect each other's probabilities. For two independent random variables, the joint probability is equal to the product of their individual probabilities. In what follows, we use this rule to find the expected frequencies. As a running example, we investigate the relationship between smoking and low birthweight.

When investigating the relationship between two categorical variables, we typically start by creating a contingency table to summarize the data. In Chap. 3, we showed how to use R-Commander to create contingency tables. In R-Commander, load the `birthwt` data set and make sure that the variables `low` and `smoke` are converted to factors (categorical) variables. Then, create a contingency table by clicking `Statistics → Contingency tables → Two-way table`. Select `smoke` for the `Row variable` and `low` for the `Column variable`. For now, uncheck `Chi-square test of independence` under `Hypothesis Tests` but make sure that `Percentages of total` and `Print expected frequencies` are selected (Fig. 10.4). R-Commander provides Table 10.1, which shows the observed frequency of each cell (i.e., each combination of mother's smoking status and baby's birthweight status). We denote the observed frequency in row i and column j as O_{ij}.

Because we checked the options `Percentages of total` and `Print expected frequencies`, R-Commander also provides Table 10.2, which shows the proportion of observations, out of the total number of observations, that fall within each cell (i.e., each possible combination of smoking status and birthweight status). Table 10.3 shows the expected frequency of each cell if the null hypothesis was true and the two random variables were independent. We denote the expected frequency in row i and column j as E_{ij}.

Table 10.1 Contingency table of `low` by `smoke`

Observed frequency		low	
		0	1
smoke	0	86	29
	1	44	30

Fig. 10.4 Creating the contingency table along with the expected frequencies for the mother's smoking status (the row variable) by the baby's birthweight status (the column variable)
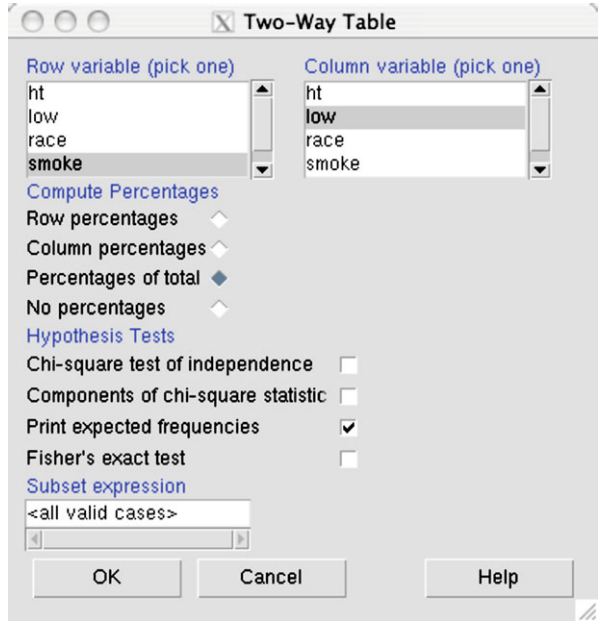
Table 10.2 Sample proportions for each combination of `low` and `smoke`

Proportion		low		Total
		0	1	
smoke	0	0.455	0.153	0.608
	1	0.233	0.159	0.392
	Total	0.688	0.312	1

Table 10.3 Expected frequencies for different combinations of `low` and `smoke` assuming the null hypothesis is true

Expected frequency		low	
		0	1
smoke	0	79.1	35.9
	1	50.9	23.1

To see how the expected frequencies are calculated, recall that for two independent variables, the probability of the intersection of events is the product of their individual probabilities. Therefore, for example, the probability that the mother is smoker (i.e., `smoke=1`) and the baby has low birthweight (i.e., `low=1`) is the product of smoker and low-birthweight probabilities. We use sample proportions to estimate these probabilities. The proportion of observations with `smoke=1` according to Table 10.2 is 0.392, and the proportion of observations with `low=1` is 0.312. Therefore, our estimate of the joint probability (under the null) is 0.392 $\times$

Table 10.4 Comparing the observed and expected (under the null hypothesis) frequencies for different combinations of birthweight status and smoking status

	Observed			Expected	
	Normal	Low		Normal	Low
Nonsmoking	86	29	Nonsmoking	79.1	35.9
Smoking	44	30	Smoking	50.9	23.1

$0.312 = 0.122$. Consequently, out of 189 babies, we expect $0.122 \times 189 = 23.1$ babies to have smoker mother and have low birthweight if the null hypothesis is true and the two variables are in fact independent. This value is shown in the second row and second column of the expected frequency table (Table 10.3). We find the expected frequency of other cells under the null hypothesis similarly.

If the null hypothesis is true, the observed frequencies would be close to the expected frequencies under the null. We therefore use the difference between the observed and expected frequencies as a measure of disagreement between the observed data and what we expected under the null. This would be interpreted as evidence against the null hypothesis. For this, we use the following general form of Pearson's χ^2 test, which summarizes the differences between the expected frequencies (under the null hypothesis) and the observed frequencies over all cells of the contingency table:

$$Q = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} and E_{ij} are the observed and expected values in the ith row and jth column of the contingency table. The double sum simply means that we add the individual measures of discrepancies for cells by going through all cells in the contingency table.

As before, higher values of Q provide stronger evidence against H_0. For $I \times J$ contingency tables (i.e., I rows and J columns), the Q statistic has approximately the χ^2 distribution with $(I - 1) \times (J - 1)$ degrees of freedom under the null. Therefore, we can calculate the observed significance level by finding the upper tail probability of the observed value for Q, which we denote as q, based on the χ^2 distribution with $(I - 1) \times (J - 1)$ degrees of freedom.

For the baby weight example, we can summarize the observed and expected frequencies in the contingency tables (Table 10.4).

Then Pearson's test statistic is

$$Q = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(E_{22} - E_{22})^2}{E_{22}},$$

$$q = \frac{(86 - 79.1)^2}{79.1} + \frac{(29 - 35.9)^2}{35.9} + \frac{(44 - 50.9)^2}{50.9} + \frac{(30 - 23.1)^2}{23.1} = 4.9.$$

Because the table has $I = 2$ rows and $J = 2$ columns, the approximate null distribution of Q is χ^2 with $(2 - 1) \times (2 - 1) = 1$ degrees of freedom. Consequently, the observed p-value is the upper tail probability of 4.9 using the $\chi^2(1)$ distribution.

Fig. 10.5 Results from χ^2 test of independence regarding the association between the mother's smoking status and the baby's birthweight. The observed test statistic Q (called `X-squared`) is 4.9, and the *p*-value is 0.026

In R-Commander, click `Distributions` → `Continuous distribu-` `tions` → `Chi-squared distribution` → `Chi-squared probabilities`. Enter 4.9 as the `Variable value` and 1 as the `Degrees of freedom`, and select `Upper tail`. We find $p_{\text{obs}} = P(Q \geq 4.9) = 0.026$. Therefore, at the 0.05 significance level (but not at 0.01 level), we can reject the null hypothesis that the mother's smoking status and the baby's birthweight status are independent.

So far, we showed how to test a hypothesis regarding the relationship between two categorical variables by summarizing the observed and expected frequencies in contingency tables, calculating the value of Pearson's test statistic, and then finding p_{obs} from χ^2 distribution. Alternatively, we could use R-Commander to perform Pearson's χ^2 test of independence directly. In R-Commander, click `Statistics` → `Contingency tables` → `Two-way table`. As before, choose `smoke` as the `Row variable` and `low` as the `Column variable`. Now under `Hypoth-` `esis Tests`, select `Chi-square test of independence`.

In the *Output* window, R-Commander provides the results of Pearson's χ^2 test (Fig. 10.5). As calculated manually, the observed test statistic Q (which is called `X-squared`) is 4.9, and the *p*-value is 0.026.

As the second example, suppose that we would like to investigate whether the race of mothers is related to the risk of having babies with low birthweight. In R-Commander, make sure `birthwt` is the active data set and variables `low` and `race` are converted to factors (i.e., categorical variables).

The `race` variable can take three values: 1 for white, 2 for African-American, and 3 for others. As before, the `low` variable can take 2 possible values: 1 for babies with birthweight less than 2.5 kg and 0 for other babies. Therefore, all possible combinations of `race` and `low` can be presented by a 3×2 contingency table. (Each observation falls into one cell of this table.) In R-Commander, click `Statistics` → `Contingency tables` → `Two-way table`. Choose `race` as the `Row variable` and `low` as the `Column variable`, check the options `Percentage of total`, `Chi-square test of independence`, and `Print expected frequencies` as shown in Fig. 10.6. The contingency tables (Tables 10.5 and 10.6) appear in the *Output* window.

Table 10.5 provides the observed frequency of each cell, Table 10.6 provides the expected frequency of each cell if the null hypothesis is true. For example, there are 73 babies in the first row and first column. This is the number of babies in

Fig. 10.6 Obtaining the contingency table of `race` by `low` from the `birthwt` data set
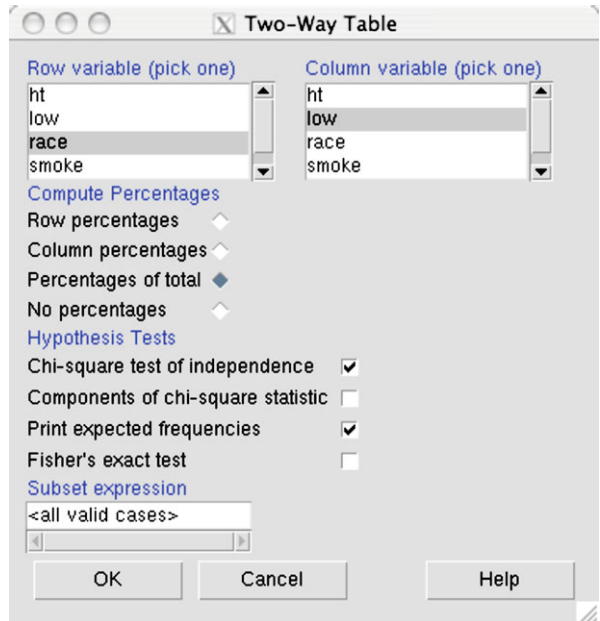
Table 10.5 Contingency table of birthweight status by race
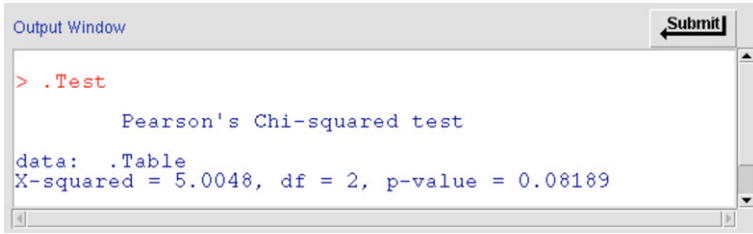
Observed frequency			low	
			0	1
race	1		73	23
	2		15	11
	3		42	25

Table 10.6 Expected frequencies for different combinations of `low` and `race` assuming the null hypothesis is true

Expected frequency			low	
			0	1
race	1		66	30
	2		18	8
	3		46	21

the intersection of `race = 1` (mother is white) and `low=0` (having a baby with normal birthweight). If the null hypothesis is true, the expected number of babies in this cell would have been 66.

For this example, the observed value of the test statistic Q is $q = 5.0$, and the distribution of Q under the null hypothesis is (approximately) χ^2 with $(3-1) \times (2-1) = 2$ degrees of freedom. (Use R-Commander to plot the pdf of this distribution.) To find the corresponding p-value, we need to find the probability of

Fig. 10.7 Results from the χ^2 test of independence for the relationship between `race` and `low`. The test statistic is $q = 5.0$, and using the χ^2 distribution with 2 degrees of freedom, the resulting p-value is 0.08

observing values as or more extreme (i.e., greater) than 5.0. This is the upper-tail probability of 5 from the $\chi^2(2)$ distribution: $p_{obs} = P(Q \geq 5)$.

We can use R-Commander to obtain the p-value. Click `Distributions` $\rightarrow$ `Continuous distributions` $\rightarrow$ `Chi-squared distribution` $\rightarrow$ `Chi-squared probabilities`. Then enter 5 for `Variable value` and 2 for `Degrees of freedom`. Also make sure the `Upper tail` option is selected. The value of p_{obs}, which is 0.08, appears in the *Output* window. We can reject the null hypothesis at 0.1 level but not at 0.05 level. At 0.05 level, the relationship between the two variables (i.e., race of mothers and birthweight status) is not statistically significant. This means that either the null hypothesis is in fact true (the two variables are independent), or it is false, but we do not have enough empirical evidence to reject it at 0.05 level. In this case, we believe that the difference between observed and expected frequency could be due to chance alone.

While it is a good exercise to follow the above steps in order to find Q and its corresponding p-value, R-Commander has provided these values when we selected the `Chi-square test of independence` option above. Figure 10.7 shows the R-Commander output.

10.4 Entering Contingency Tables into R-Commander

In R-Commander, we can enter and analyze contingency tables without importing individual observations. For example, let us consider the study for investigating whether regular aspirin intake reduces the mortality from cardiovascular disease [36]. In this case, the null hypothesis is that the heart attack is independent of aspirin intake. Based on the contingency table (Table 10.7) of the observed frequencies, we can evaluate the strength of relationship between these two binary variables.

Table 10.7 Contingency table of heart attack status by the type of treatment

	Heart attack	No heart attack
Placebo	189	10845
Aspirin	104	10933

Fig. 10.8 Entering and
analyzing a contingency table
in R-Commander. The default
table size is 2×2 and can be
changed with the `Number`
`of Rows` and `Number of`
`Columns` buttons. Here, we
are testing the null hypothesis
of no relationship between
aspirin intake and the
probability of a heart attack
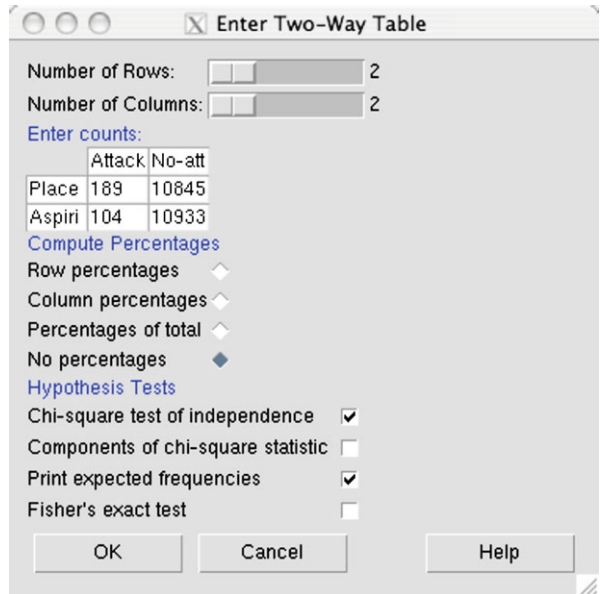
To enter this contingency table in R-Commander, click `Statistics` →
`Contingency tables` → `Enter and analyze two-way table`.
R-Commander then opens a blank table. Enter the row and column labels and fre-
quencies for each cell as shown in Fig. 10.8. Then select `Chi-square test of`
`independence` and `Print expected frequencies` under `Hypothesis`
`Tests`.

In the *Output* window, R-Commander provides contingency tables for the ob-
served and expected frequencies. The results of Pearson's χ^2 test of independence
are also given. In this case, the observed value of the test statistic Q is $q = 25.01$,
and the *p*-value is $p_{obs} = 5.69 \times 10^{-7}$, which is quite small. Consequently, at any
reasonable level of significance, we can reject the null hypothesis of independence
between the variables and conclude that the results are statistically significant, and
so the observed departure from the null hypothesis is quite unlikely to be due to
chance alone.

10.5 Advanced

In this section, we discuss Fisher's exact test for analyzing contingency tables from
small data sets. We also provide some commonly used R functions for analyzing
categorical data.

10.5.1 Fisher's Exact Test

For Person's χ^2 test to be valid, the expected frequencies (E_{ij}) under the null should
be at least 5, so we can assume that the distribution of Q is approximately χ^2 under

Table 10.8 Contingency table of diabetes status by weight status

Frequency	type		Total
	No	Yes	
weight.status			
Underweight	2	0	2
Normal	21	2	23
Overweight	35	8	43
Obese	74	58	132
Total	132	68	200

```
Output Window                                                    Submit

        Pearson's Chi-squared test

data:   .Table
X-squared = 17.9462, df = 3, p-value = 0.0004512

Messages
[21] WARNING: 1 expected frequencies are less than 1
2 expected frequencies are less than 5
```

Fig. 10.9 When examining the relationship between weight.status and type, R-Commander gave the warning message: "2 expected frequencies are less than 5." Therefore, Fischer's exact test is more appropriate for analyzing the contingency table

the null. Occasionally, this requirement is violated (especially when the sample size is small, or the number of categories is large, or some of the categories are rare), and some of the expected frequencies become small (less than 5).

Recall that in Sect. 2.5, we created a categorical variable called weight. status based on BMI values in the Pima.tr data set. This variable had four categories: "Underweight", "Normal", "Overweight", and "Obese". After you create this new variable, follow the above steps to perform Pearson's χ^2 test in order to investigate the relationship between weight.status and type (disease status). You should obtain a 4×2 contingency table, similar to Table 10.8, by selecting weight.status as the row variable and type as the column variable. Note that there are only two underweight women in this sample. (This seems to be a rare event in the population.)

Based on the above table, the expected frequencies $E_{1,1} = 1.32$ and $E_{12} = 0.68$. (The remaining expected frequencies are above 5.) Therefore, when we use the χ^2 test, R-Commander gives a warning message indicating that "2 expected frequencies are less than 5" (Fig. 10.9). In this case, instead of using Pearson's χ^2 test (which assumes that Q statistic has an approximate χ^2 distribution), we should use Fisher's exact test (which is based on the exact distribution of a test statistic that captures the deviation from the null).

In R-Commander, follow the above steps to create the contingency table for weight.status and type, but this time, select Fisher's exact test, instead of the default option Chi-square test of independence, under

the `Hypothesis tests`. The resulting p-value is 0.0002, which is slightly lower than the p-value of 0.0004 based on χ^2 test. At 0.01 level, we can reject the null hypothesis, which indicates that the disease status is independent from the weight status, and conclude that the relationship between the two variables is statistically significant.

10.5.2 Pearson's χ^2 Test Using R

To test a hypothesis regarding the probabilities (population proportions) of different categories for a categorical variable, we can use the `chisq.test()` function to perform Pearson's χ^2 test in R:

```
> chisq.test(x = c(24, 16), p = c(0.7, 0.3))

        Chi-squared test for given
        probabilities

data:  c(24, 16)
X-squared = 1.9048, df = 1, p-value =
0.1675
```

The first argument to the `chisq.test()` function provides the observed frequencies for each possible category. (Here, there are two categories.) The second argument, `p`, specifies the corresponding probabilities under the null hypothesis. In the output, `X-squared` provides the observed value of the test statistics (which we denoted Q).

We can also use the `chisq.test()` function for categorical variables with multiple categories. For the heart attack example discussed previously, the null hypothesis was $H_0 : \mu_{01} = 0.5$, $\mu_{02} = 0.2$, $\mu_{03} = 0.3$. The observed frequencies were $O_1 = 13$, $O_2 = 11$, and $O_3 = 16$. Therefore, we can perform Pearson's χ^2 test as follows:

```
> chisq.test(x = c(13, 11, 16), p = c(0.5, 0.2, 0.3))

        Chi-squared test for given
        probabilities

data:  c(13, 11, 16)
X-squared = 4.9083, df = 2, p-value =
0.08593
```

As before, `x` provides the number of observations in each group, and `p` provides the corresponding probability of each group under the null hypothesis.

To test the relationship between two binary random variables, we use the χ^2 test to compare the observed frequencies to the expected frequencies based on the null

hypothesis. To use `chisq.test()` for this purpose, we first create the contingency table using the `table` function and then pass the resulting contingency table to `chisq.test()`.

For example, the following code creates the contingency table for `smoke` by `low` from the `birthwt` data set and then performs the χ^2 test to examine their relationship:

```
> birthwt.tab <- table(birthwt$smoke, birthwt$low)
> birthwt.tab

     0  1
  0 86 29
  1 44 30

> chisq.test(birthwt.tab, correct = FALSE)

        Pearson's Chi-squared test

data:  birthwt.tab
X-squared = 4.9237, df = 1, p-value =
0.02649
```

Note that we have set the option `correct` to `FALSE` to obtain the same results as we obtained in earlier sections. By default, the function `chisq.test()` performs continuity correction (not discussed in this chapter) for analyzing 2×2 contingency tables.

If we only have the summary of the data in the form of a contingency table as oppose to individual observations, we can enter the contingency table in R and perform the χ^2 test as before. For example, consider the study investigating the relationship between aspirin intake and the risk of a heart attack [36]. We can enter the given the contingency table directly in R.

```
> contTable <- matrix(c(189, 10845, 104, 10933),
+     nrow = 2, ncol = 2, byrow = TRUE)
> rownames(contTable) <- c("Placebo", "Aspirin")
> colnames(contTable) <- c("No heart attack",
+     "Heart attack")
> contTable

         No heart attack Heart attack
Placebo              189        10845
Aspirin              104        10933
```

Here, the first parameter to the `matrix()` function is a vector of values. We also specify the number of rows (`nrow`) and the number of columns (`ncol`). The `byrow` option tells R to fill the matrix by rows. We then use the `rownames()` and `colnames()` to add names to the rows and columns, respectively.

To examine the relationship between heart attack and aspirin intake, we use the `chisq.test()` function as before:

```
> output <- chisq.test(contTable, correct = FALSE)
> output

        Pearson's Chi-squared test

data:   contTable
X-squared = 25.0139, df = 1, p-value =
5.692e-07
```

The argument to the `chisq.test()` function is the contingency table of observed values. We have assigned the output of the function to a new object called `output`. From this object, we can obtain the observed and expected frequencies with the $ operator:

```
> output$observed

          No heart attack Heart attack
Placebo               189        10845
Aspirin               104        10933

> output$expected

          No heart attack Heart attack
Placebo          146.4801     10887.52
Aspirin          146.5199     10890.48
```

10.6 Exercises

1. Consider the problem of estimating the proportion of people who smoke. Suppose that we interview 150 people and find that 27 of them smoke. Use Pearson's χ^2 test to evaluate the null hypothesis stating that the probability of smoking is 0.2.
2. Use the `birthwt` data set to evaluate the relationship between having low-birthweight babies and mother's hypertension history. (In `birthwt` data set, the variable `ht` shows the hypertension history: `ht=1` when women have history of hypertension, `ht=0` otherwise.)
3. In Sect. 3.6, we used the `GBSG` (German Breast Cancer Study Group) data set from the `mfp` package to create a new variable called `rfs` (recurrence-free survival) such that `rfs="No"` if the patient had at least one recurrence or died (i.e., `cenc=1`) and `rfs="Yes"` otherwise. Use the χ^2 test to investigate the relationship between this newly created variable and the tumor grade (`tumgrad`). Make sure you convert `tumgrad` to a categorical (factor) variable first.

Table 10.9 Frequencies of people with heart disease for different levels of snoring based on a sample of 2484 people

Snoring Severity	Heart Disease	Total
Never	24	1379
Occasionally	35	638
Nearly every night	21	213
Every night	30	254

4. Consider the data (see Table 10.9) collected to investigate snoring as a risk factor for heart disease [24]. Use Pearson's χ^2 test to examine whether the relationship between snoring severity and the risk of heart disease is statistically significant.

5. In R-Commander, click `Data` → `Data in pacakges` → `Read data set from an attached package`, then select the `HairEyeColor` data from the `datasets` package. The data include hair and eye color and sex for 592 statistics students at the University of Delaware reported by Snee (1974). The first column shows different hair colors (Black, Brown, Red, Blond), the second column shows different eye colors (Brown, Blue, Hazel, Green), and the third column shows genders (Male, Female) of students. For each row, the last column shows the number of students with a specific hair color, eye color, and gender.

   (a) Use Pearson's χ^2 test to evaluate the null hypothesis that different hair colors have equal probabilities. Use Pearson's χ^2 test to evaluate the null hypothesis that different eye colors have equal probabilities.

   (b) Create a 4×4 contingency table where the rows represent different hair colors and the columns represent different eye colors. Is there a relationship between hair color and eye color?

   (c) Enter the contingency table in R-Commander and use `Chi-square test of independence` to evaluate the relationship between hair color and eye color.

Chapter 11
Regression Analysis

11.1 Introduction

In Chap. 8, we discussed testing a hypothesis regarding the relationship between a binary explanatory variable (referred to as the factor) and a numerical response variable using two-sample t-tests. We also discussed situations where we investigate the linear relationship between two numerical variables (e.g., percent body fat and abdomen circumference). In this case, we typically consider one of the two numerical variables (e.g., percent body fat) as the response (or target) variable and the other one (e.g., abdomen circumference) as the explanatory variable. The common theme for both methods is that we investigate the relationship between an explanatory variable (either categorical or numerical) and a numerical random variable. In this chapter, we discuss an alternative approach for analyzing such problems. This approach use **linear regression models** for either testing a hypothesis regarding the relationship between one or more *explanatory variables* and a response variable, or **predicting** unknown values of the response variable using one or more *predictors*. Note that we use the terms "explanatory variables" and "predictors" to distinguish the role of variables (other than the response variable) in the model.

> Occasionally, we might want to use linear regression models for both hypothesis testing and prediction. However, in most cases, our objective is either examining the relationship between the response variable and a set of explanatory variables, or predicting the unknown values of the response variable using a set of predictors. It is very important to specify our objective clearly before starting the analysis. As we discuss later in this chapter, our strategy to build a linear regression model depends on our objective.

Throughout this chapter, we use X to denote explanatory variables and Y to denote response variables. We start by focusing on problems where the explanatory variable is binary. As before, the binary variable X can be either 0 or 1. We then

Fig. 11.1 The dot plot for
systolic blood pressure for 25
elderly people, where 15
people follow a low sodium
chloride diet ($X = 0$), and 10
people follow a high sodium
chloride diet ($X = 1$)
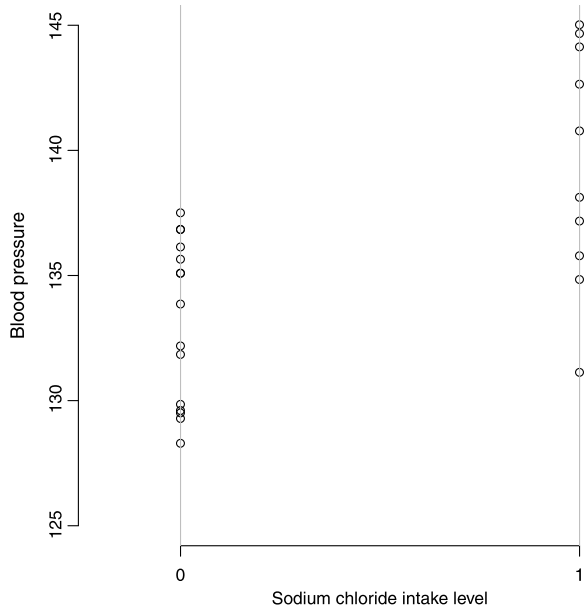
continue our discussion for situations where the explanatory variable is numerical.
Finally, we discuss linear regression problems where there are more than one ex-
planatory variable (possibly, a combination of binary and numerical variables).

11.2   Linear Regression Models with One Binary Explanatory Variable

Suppose that we want to investigate the relationship between sodium chloride (salt)
consumption and blood pressure among elderly people (e.g., above 65 years old).
We take a random sample of 25 people from this population and find that 15 of
them keep a low sodium chloride diet (less than 6 grams of salt per day) and 10
of them keep a high sodium chloride diet (more than 6 grams of salt per day). We
use the variable X for the sodium chloride intake level, where $X = 0$ means low
sodium chloride intake (group 1), and $X = 1$ means high sodium chloride intake
(group 2). For people in our sample, we measure systolic blood pressure, which
we denote as Y. Therefore, for each individual i in our sample, we have a pair of
observations (x_i, y_i), where the first element shows the group (low or high sodium
chloride diet), and the second element shows the blood pressure measure. You can
find the data for this example from the book website (http://extras.springer.com).
(These are simulated data for illustrative purposes.)

Figure 11.1 shows the dot plot for the observed data. As we can see, blood
pressure tends to be higher among people with high sodium chloride diet. Fig-
ure 11.2 shows the dot plot along with sample means, shown as black circles,

Fig. 11.2 The dot plot for systolic blood pressure for 25 elderly people. Here, the sample means among the low and high sodium chloride diet groups are shown as *black circles*. A *straight line* connects the sample means. The line intercepts the vertical axis at $a = 133.17$ and has slope $b = 6.25$
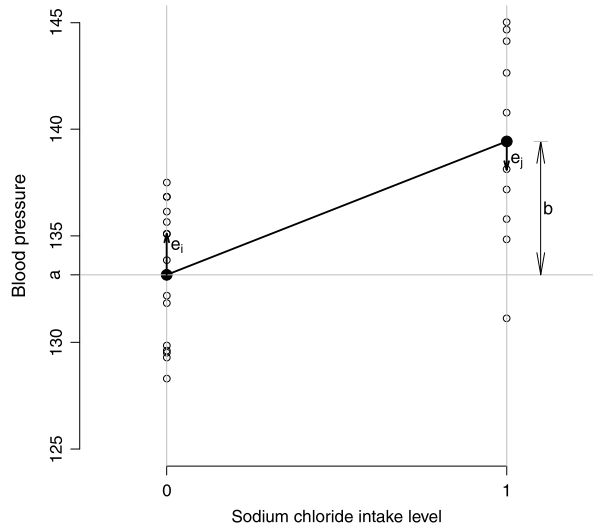
for each group. In this graph, the difference between the two sample means is denoted as b. Recall that the difference between the sample means was what we used to perform the two-sample t-test. By connecting the two sample means, we can show the overall pattern for how blood pressure changes from one group to another.

The sample mean among the first group (the black point in the left) is regarded as our point estimate for population mean of systolic blood pressure among people with low sodium chloride diet ($X = 0$). If we do not know someone's blood pressure measure, y, but we know that she belongs to the first group (i.e., $x = 0$), the sample mean provides a reasonable point estimate of her blood pressure. We show this as $\hat{y}_{x=0}$ and interpret it as the estimate of the response variable among individuals whose value of the explanatory variable is $x = 0$. Similarly, we can use the sample mean for the second group (the black point in the right) as our point estimate of the response variable (i.e., blood pressure) among people whose value of the explanatory variable (sodium chloride intake level) is $x = 1$. We denote this estimate as $\hat{y}_{x=1}$.

For the above example, the sample means of blood pressure for the first and second groups are 133.17 and 139.42, respectively. Therefore, $\hat{y}_{x=0} = 133.17$ and $\hat{y}_{x=1} = 139.42$.

Now consider the straight line that connects the two point estimates $\hat{y}_{x=0}$ and $\hat{y}_{x=1}$. This line contains both point estimates (black points); hence, for any value of x (i.e., whether $x = 0$ or $x = 1$), the line can be used to find the estimate of the response variable (blood pressure). We denote this estimate as $\hat{y}$. Because the horizontal difference between the two points is 1 unit, and the vertical difference between the two points is $b = \hat{y}_{x=1} - \hat{y}_{x=0}$, the slope of this line is $b/1 = b$. The line intercepts the vertical axis at a. In this case, $a = \hat{y}_{x=0}$. For the above example, $a = 133.17$ and $b = 139.42 - 133.17 = 6.25$.

Using the intercept a and slope b, we can write the equation for the straight line (Fig. 11.2) that connects the estimates of the response variable for different values of X as follows:

$$\hat{y} = a + bx.$$

The above equation specifies a straight line called the **regression line**. The regression line captures the linear relationship between the response variable (here, blood pressure) and the explanatory variable (here, "low" versus "high" sodium chloride diet).

The slope of the regression line has a central role in capturing the linear relationship between the response variable and the explanatory variable: the slope b is interpreted as our estimate of the expected (average) change in the response variable associated with one unit increase in the value of the explanatory variable. Note that in general, we *cannot* interpret this as the amount of increase in the response variable *caused* by one unit increase in the explanatory variable unless the data are obtained through a randomized experiment, where the value of the explanatory variable is changed by intervention.

For the blood pressure example, the regression line is

$$\hat{y} = 133.17 + 6.25x.$$

Based on the slope of $b = 6.25$, we expect that on average the blood pressure increases by 6.25 units for one unit increase in the value of the explanatory variable. In this case, one unit increase in X from 0 to 1 means moving from low sodium chloride diet group to high sodium chloride diet group.

We can use the regression line to estimate the blood pressure of a subject given his or her sodium chloride intake level. For an individual with $x = 0$ (i.e., low sodium chloride diet), the estimate according to the above regression line is

$$\hat{y} = a + b \times 0 = a$$
$$= \hat{y}_{x=0},$$

which is the sample mean for the first group. For an individual with $x = 1$ (i.e., high sodium chloride diet), the estimate according to the above regression line is

$$\hat{y} = a + b \times 1 = a + b$$
$$= \hat{y}_{x=0} + \hat{y}_{x=1} - \hat{y}_{x=0}$$
$$= \hat{y}_{x=1}.$$

Note that in this case, where the explanatory variable is binary with 0–1 values, we can evaluate the equation for the regression line at $x = 0$ and $x = 1$ only.

Now consider the observed data for our sample of 25 people. For each individual in this sample, there is a difference between the actual observed blood pressure and

the estimate we obtain from the regression line. For individual i, whose values of the explanatory variable and the response variable are x_i and y_i, respectively, the estimated value of the response variable, denoted as $\hat{y}_i$, is

$$\hat{y}_i = a + bx_i.$$

We refer to the difference between the observed and estimated values of the response variable as the **residual**. For individual i, we denote the residual e_i and calculate it as follows:

$$e_i = y_i - \hat{y}_i.$$

By rearranging the terms in the above equation, we can write the observed value y_i in terms of the estimate obtained from the regression line and the corresponding residual,

$$y_i = \hat{y}_i + e_i$$
$$= a + bx_i + e_i.$$

For the blood pressure example, $\hat{y} = a$ for all individuals in the first group, and $\hat{y} = a + b$ for all individuals in the second group. For instance, if an individual i belongs to the first group, $x_i = 0$, her estimated blood pressure is $\hat{y}_i = a = 133.17$. Now if the observed value of her blood pressure is $y_i = 135.08$, then the residual is

$$e_i = 135.08 - 133.17 = 1.91.$$

In Fig. 11.2, the residuals e_i and e_j are shown as vertical arrows for two individuals (one from each group). For individual i (in the first group), the residual is positive since y_i is greater than $\hat{y}_i = a$. For individual j (in the second group), the residual is negative since y_j is less than $\hat{y}_j = a + b$. The directions of the arrows show the sign of the residuals: upward for positive residuals and downward for negative residuals.

When the explanatory variable is binary, the residuals are in fact deviations from the corresponding group sample means. As mentioned in Chap. 2, the sum of the deviations from the sample mean over all observed values is always zero. Therefore, the sum (hence, the mean) of all the residuals is zero. The sum of the squares of the residuals, however, is not zero in general (it is a positive value) and is commonly used as a measure of overall discrepancy between observed values and estimates from the regression line. This is analogous to the within-group sum of squares measure, SS_W, we used for ANOVA.

As a measure of discrepancy between the observed values and those estimated by the line, we calculate the **Residual Sum of Squares** (RSS):

$$RSS = \sum_{i}^{n} e_i^2. \tag{11.1}$$

Here, e_i is the residual of the ith observation, and n is the sample size. The square of each residual is used so that its sign (i.e., the direction of the arrows in Fig. 11.2) becomes irrelevant.

To capture the overall change in blood pressure from one group to another, we decided to draw a line by connecting the sample means. We could have of course chosen different lines between the two groups; for example, we could have connected the sample medians. For any possible line, we can define the residuals as before (i.e., the vertical difference between the observed values and the line) and calculate RSS. It turns out that among all possible straight lines we could have drawn, the linear regression line discussed above provides the smallest value of RSS. Therefore, the above approach for finding the regression line is called the **least-squares** method, and the resulting line is called the **least-squares regression line**.

Using R-Commander for Finding Regression Lines   For the blood pressure example, we can simply find the regression line by calculating *a* and *b* using the sample means of blood pressure for the two groups separately. Download the "saltBP.txt" data from the book website (http://extras.springer.com) and load it into R-Commander. In this data set, BP shows the observed systolic blood pressure, salt shows the amount of sodium chloride intake per day, and saltLevel is a binary variable indicating whether sodium chloride intake per day is less than 6 grams (saltLevel = 0) or above 6 grams (saltLevel = 1).

To find the sample means for each group, convert saltLevel to factor, then click Statistics → Summaries → Numerical summaries. Select BP under Variables. Next, click Summarize by group and select saltLevel. You will find the sample mean for the first group (saltLevel = 0) to be 133.17 and for the second group (saltLevel = 1) to be 139.42. Using these sample means, you can find *a* and *b* for the regression line as described above.

R-Commander can be used to find *a* and *b* directly without calculating the sample means. For this, click Statistics → Fit models → Linear regression. Then under Model Formula, enter BP ∼ saltLevel, as in Fig. 11.3. The variable on the left of the ∼ sign is always the response variable, and the variable (or variables as we will see later) on the right side of the ∼ sign is the explanatory variable. Instead of typing the name of the response variable and explanatory variable, we can first double click on BP and then double click on saltLevel.

Fig. 11.3 Using
R-Commander to find the
least-squares regression line
with `BP` as the response
variable and `saltLevel` as
the explanatory variable
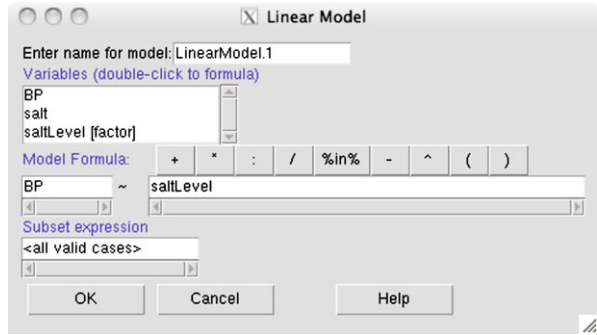
Fig. 11.4 Output of R-Commander for finding the regression line for the blood pressure example.
The first column of the `Coefficients` table provides the intercept and the slope

R-Commander automatically fills the space under `Model Formula`. Also, notice
that R-Commander has assigned a name, here `LinearModel.1`, to the regression
line. You can specify a different name if you want. When you press OK, `Linear-
Model.1` (or any other name you specify) appears under the menu bar in front of
`Model`.

When we press OK, R-Commander provides a table with the title `Coeffi-
cients`, where the first column, called `Estimate`, includes the intercept 133.17
and the slope 6.25 (Fig. 11.4). The output, of course, includes more information re-
garding statistical inference using regression analysis, which is the focus of the next
section.

11.3 Statistical Inference Using Simple Linear Regression Models

In the previous section, we discussed finding regression lines with binary explana-
tory variable. We showed how we can find the intercept and slope of the regression
line and discussed how we can use regression lines to estimate the values of the
response variable. As usual, we would like to extend our findings to the entire pop-
ulation. This is the topic of this section.

Using the regression line, we can estimate the unknown value of the response
variable for members of the population who did not participate in our study. In
this case, we refer to our estimates as **predictions**. For example, we can use the

linear regression model we built in the previous section to predict the value of blood pressure for a person with high sodium chloride diet (i.e., $x = 1$),

$$\hat{y} = 133.17 + 6.25x$$
$$= 133.17 + 6.25 \times 1$$
$$= 139.42.$$

Next, we want to use the linear regression model to comment on the type and strength of the relationship between Y and X. Using the observed data, the regression line captures the linear relationship between the response variable and the explanatory variable as follows:

$$\hat{y} = a + bx.$$

Based on this line, we can write the value of the response variable for individual i in terms of the above regression line and the residual:

$$y_i = a + bx_i + e_i.$$

The linear relationship between Y and X in the entire population can be presented in a similar form,

$$Y = \alpha + \beta X + \varepsilon, \tag{11.2}$$

where α is the intercept, and β is the slope of the regression line if we had used the entire population to find the regression line. Here, ε is called the **error term**, representing the difference between the estimated (based on the regression line for the entire population) and the actual values of Y in the population. We refer to the above equation as the **linear regression model**. More specifically, we call it the **simple linear regression model** since there is only one explanatory variable. We refer to α and β as the **regression parameters**. More specifically, β is called the **regression coefficient** for the explanatory variable. The process of finding the regression parameters is called **fitting** a regression model to the data.

As usual, the regression parameters for the population remain unknown. We estimate these parameters using a random sample from the population. Suppose that we have a sample of size n: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. In this case, we have a pair of values (one for the explanatory variable and one for the response variable) for each individual in our sample. Using this sample, we can estimate the regression parameters by fitting a linear regression model to the observed data as described above:

$$\hat{Y} = A + BX.$$

Here, A and B are statistics (i.e., calculated based on the observed data only), which are used as estimators. We used capital letters since A and B themselves are random

variables and their values can change every time we take a new sample of size n from the population.

Of course, as before, we only have one such sample, $(x_1, y_1), (x_2, y_2)$, $\ldots, (x_n, y_n)$. Using the observed data, we find the point estimates a and b (i.e., the specific values of estimators A and B) for α and β, respectively, following the steps discussed in the previous chapter. For blood pressure example, the point estimates for α and β were $a = 133.17$ and $b = 6.25$. These estimates are provided in the first column of the Coefficients table under Estimate in Fig. 11.4.

As discussed in earlier chapters, the point estimates do not reflect our uncertainty; we always remain uncertain about the actual values of α and β, and our estimates for these parameters can change when we take a different sample from the population. Therefore, we use the point estimates a and b obtained from the observed data along with the sampling distribution of the estimators A and B to find confidence interval estimates for α and β. This is similar to what we had for the population mean. Because the slope parameter β has a central role in capturing the linear relationship between Y and X, we focus on finding its confidence intervals. (Similar approach can be used for the intercept α.)

11.3.1 Confidence Interval for Regression Coefficients

Finding confidence intervals for β is quite similar to the approach we used to find confidence intervals for the population mean. First, we need to find the standard error (i.e., estimated standard deviation of the sampling distribution of B) of our estimate. We denote this as SE_b. Next, we need to specify the required confidence level, c, and find its corresponding factor, t_{crit}. Then, we can find the confidence interval for β as follows:

$$[b - t_{\text{crit}} \times SE_b, b + t_{\text{crit}} \times SE_b].$$

For simple linear regression models, SE_b is obtained as follows:

$$SE_b = \frac{\sqrt{RSS/(n-2)}}{\sqrt{\sum_i (x_i - \bar{x})^2}}, \tag{11.3}$$

where x_i are the observed values of the explanatory variable, which takes either 0 or 1, and $\bar{x}$ is the sample mean (which is the same as the sample proportion for a binary variable). When we fit a linear regression model using R-Commander, it provides the standard error of the regression coefficient. From the Coefficients table in Fig. 11.4, the Std. Error column provides the standard error for intercept as $SE_a = 1.01$ and the standard error for the regression coefficient as $SE_b = 1.59$.

The steps to find t_{crit} are the same as the steps discussed in Chap. 6 for the population mean. Here, however, t_{crit} is obtained from the t-distribution with $n - 2$ degrees of freedom. In our example, the sample size is $n = 25$. Therefore, we use the t-distribution with $25 - 2 = 23$ degrees of freedom. If we set the confidence level to 0.95, then $t_{\text{crit}} = 2.07$, which is obtained from the t-distribution with 23 degrees

Fig. 11.5 Least-squares
regression line (*solid line*)
and its 95% confidence
interval (*dashed curves*) for
the relationship between
blood pressure and sodium
chloride intake level
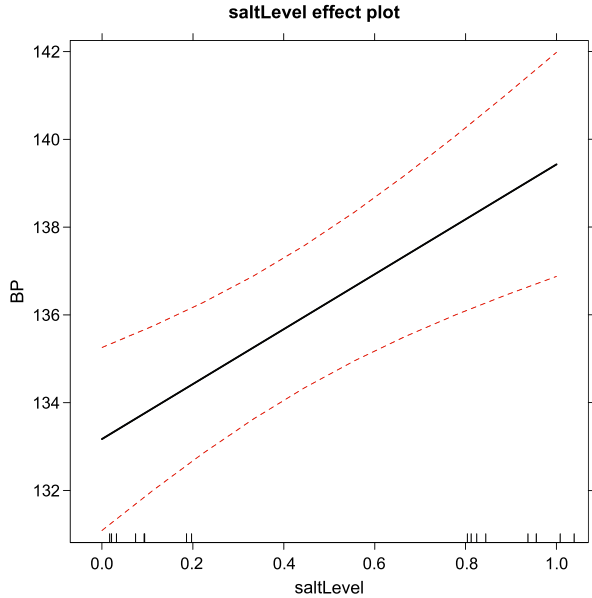
of freedom by setting the upper tail probability to $(1 - 0.95)/2 = 0.025$. Therefore, the 95% confidence interval for β is

$$[6.25 - 2.07 \times 1.59, 6.25 + 2.07 \times 1.59] = [2.96, 9.55].$$

At 0.95 confidence level, the slope of the regression line is between 2.96 and 9.55. In other words, by moving from the low sodium chloride diet to high sodium chloride diet, the expected (average) amount of increase in blood pressure is estimated to be somewhere between 2.96 and 9.55 units.

Alternatively, we can use R-Commander to obtain the confidence intervals. Make sure that the current model shown under the menu bar is LinearModel.1 (or any other name you gave to the regression model). Then click Models → Confidence intervals and set the confidence level. (The default is 0.95.) R-Commander provides the point estimates along with the confidence intervals for α and β in the output window.

As our estimates for α and β change, the least-squares regression line changes. Therefore, we can obtain confidence intervals for the regression line and predictions we obtain based on this line. To obtain the 95% confidence interval of the regression line for the blood pressure example, make sure that LinearModel.1 (or any other name you specified) is the active model in R-Commander and then click Models → Graphs → Effect plots. The resulting plot is shown in Fig. 11.5. In this plot, the solid straight line is the least-squares regression line for the observed data, and the dashed curves show the 95% confidence interval for the regression line. The curves show how much the regression line can change as we take different samples from the population.

11.3.2 Hypothesis Testing with Simple Linear Regression Models

Linear regression models can be used for testing hypotheses regarding possible linear relationship between the response variable and the explanatory variable. For this, the regression coefficient β, its estimator B, and its point estimate b play a central role. For linear regression models with a binary explanatory variable, we found the regression line by connecting the sample means of the two groups. In these models, the slope b is the difference between the two sample means. Similarly, the regression coefficient β captures the difference between the population means for the two groups. If the response variable is not related to the binary explanatory variable, the two population means are the same, and the slope of the regression line for the whole population will be zero. That is, the null hypothesis stating no relationship between the two variable can be written as $H_0 : \beta = 0$.

This is analogous to the application of the two-sample t-test for hypothesis testing regarding the relationship between a numerical variable and a binary variable. Similar to the two-sample t-test, we need to obtain the t-score as the observed value of the test statistic. Recall that we obtained the t-score by dividing the observed difference between the sample means by its standard error. Similarly, we find the t-score for linear regression models as follows:

$$t = \frac{b}{SE_b}.$$

Then, we find the p-value (i.e., the observed significance level) by calculating the probability of as or more extreme values than t-score under the null hypothesis.

> To assess the null hypothesis $H_0 : \beta = 0$, which is interpreted as no linear relationship between the response variable and the explanatory variable, we first calculate the $t = b/SE_b$ and find the corresponding p-value as follows:
>
> $$\text{if } H_A : \beta < 0, \quad p_{\text{obs}} = P(T \leq t),$$
> $$\text{if } H_A : \beta > 0, \quad p_{\text{obs}} = P(T \geq t),$$
> $$\text{if } H_A : \beta \neq 0, \quad p_{\text{obs}} = 2 \times P(T \geq |t|),$$
>
> where T has the t-distribution with $n - 2$ degrees of freedom.

In the blood pressure example, the estimate of the regression coefficient was $b = 6.25$, and the standard error was $SE_b = 1.59$. Therefore,

$$t = \frac{6.25}{1.59} = 3.93.$$

If $H_A : \beta \neq 0$ (which is the common form of the alternative hypothesis), we find the p-value by calculating the upper tail probability of $|3.93| = 3.93$ from the t-distribution with $25 - 2 = 23$ degrees of freedom and multiplying the result by 2. For this example, $p_{\text{obs}} = 2 \times 0.00033 = 0.00066$.

Because p_{obs} for this example is quite small and below any commonly used confidence level (e.g., 0.01, 0.05, 0.1), we can reject the null hypothesis and conclude that blood pressure is related to sodium chloride diet level.

When we use R-Commander to fit a linear regression model, the output provides the t-score and its corresponding p-value. In Fig. 11.4, the column with the title t value provides the t-scores for α and β, and the last column, Pr(>|t|), provides the corresponding p-values.

11.4 Linear Regression Models with One Numerical Explanatory Variable

In the previous section, to study the relationship between blood pressure and daily salt intake, we used a binary variable that indicates whether the amount of daily sodium chloride intake for each individual is above a certain cutoff (6 grams per day). The data would be more informative of course if we use the actual amount of sodium chloride intake per day. By doing so, the explanatory variable becomes numerical (quantitative) as opposed to binary.

In this section, we discuss simple linear regression models (i.e., linear regression with only one explanatory variable), where the explanatory variable is numerical. For the most part, we use the same concepts for explaining the model, follow the same steps to fit the model, and interpret the output of the model same way. Here, of course, the explanatory variable can take more than two values. In fact, in many cases (e.g., sodium chloride intake per day), it can take an uncountable number of possible values.

We start our analysis by creating the scatter plot of the response variable and the explanatory variable. As before, upload the "saltBP.txt" data set (available from the book website) into R-Commander. Then, click Graphs → Scatterplot. Choose salt, which is the actual amount of sodium chloride intake per day, under x-variable and choose BP under y-variable. Under Options, make sure that all the options are unchecked. When you press OK, R-Commander creates a scatter plot for blood pressure vs. sodium chloride intake similar to the one shown in Fig. 11.6.

As we can see, there is a clear upward trend indicating that increase in sodium chloride intake tends to coincide with increase in blood pressure. Moreover, the trend seems to be linear, so a straight line can capture the overall pattern. Indeed, the process of fitting a linear regression model to the data involves finding a straight line that can be considered as the best representation of the overall relationship between blood pressure and sodium chloride intake. The left panel of Fig. 11.7 shows the scatterplot of blood pressure by daily sodium chloride intake along with some candidate lines for capturing the overall relationship between the two variables.

To choose a line, we need to explain what we mean by the "best representation" of the data. Similar to the approach we used earlier in this chapter, we measure the discrepancy between each line and the observed data in terms of the residual sum of

Fig. 11.6 The scatterplot of blood pressure by sodium chloride intake
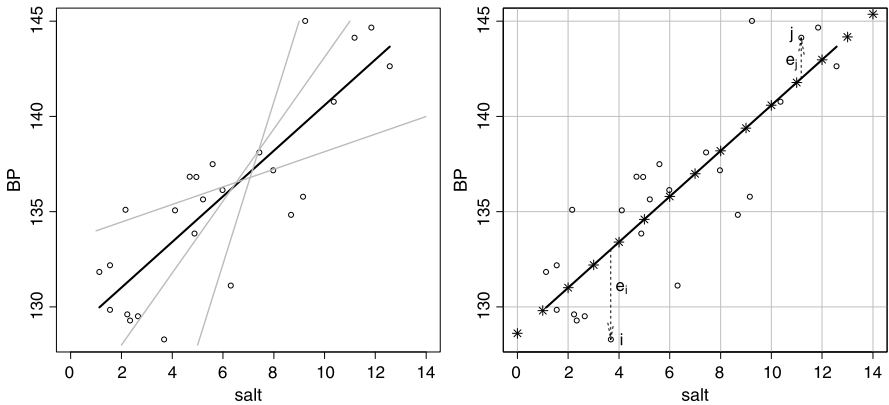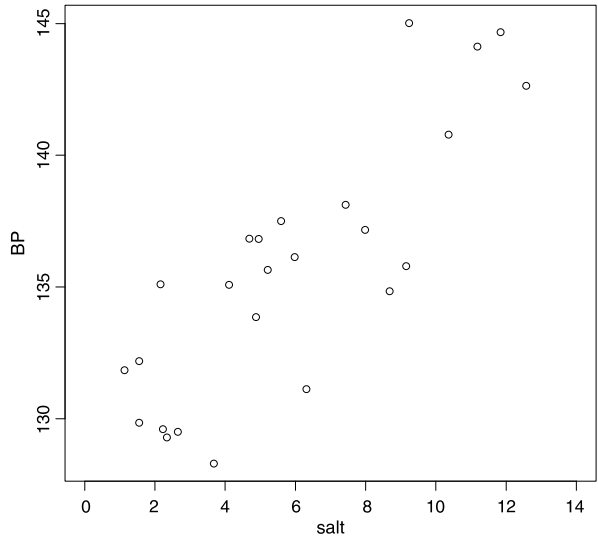
Fig. 11.7 *Left panel*: Scatterplot of blood pressure by daily sodium chloride intake along with some candidate lines for capturing the overall relationship between the two variables. The *black line* is the least-squares regression line. *Right panel*: The least-squares regression line for the relationship between blood pressure and sodium chloride intake. The *vertical arrows* show the residuals for two observations. The *stars* are the estimated blood pressure for daily sodium chloride intakes from 0 to 14 grams

squares (RSS), and choose the line with the smallest value of RSS. As before, we refer to the resulting model as the least-squares linear regression model and to the corresponding line as the least-squares regression line. In the left panel of Fig. 11.7, the black line is the least-squares regression line. As we can see, this line follows the overall pattern more closely compared to any other line. You can obtain this line by following the above steps to create the scatter plot, but this time check the option `Least-squares line` under `Options`.

As before, the regression line is specified by its intercept a and its slope b and can be used to estimate the value of the response variable. Given a and b, we can obtain the point estimate $\hat{y}_i$ for the value of the response variable for individual i, whose value of the explanatory variable is x_i,

$$\hat{y}_i = a + bx_i.$$

In general, this estimated value is different from the observed value y_i of the response variable for the individual i. We use e_i to denote the residual, which is the difference between the estimated and observed value of the response variable,

$$e_i = y_i - \hat{y}_i.$$

The right panel of Fig. 11.7 shows the residuals (vertical arrows) for two observations in the blood pressure data. For a sample of n individuals, the discrepancy between the linear regression model and the observed data is measured using the residual sum of squares,

$$RSS = \sum_{i}^{n} e_i^2. \tag{11.4}$$

As mentioned above, the least-squares regression line provides the smallest possible value of RSS among all candidate straight lines.

For each individual in our sample, we can write the observed value of the response variable in terms of the estimated value according the linear regression model and the corresponding residual as follows:

$$y_i = \hat{y}_i + e_i$$
$$= a + bx_i + e_i.$$

Here, a and b are the point estimates for α and β, which are the parameters of the linear regression model for the entire population,

$$Y = \alpha + \beta X + \varepsilon.$$

For simple linear regression model with binary explanatory variables, we found a and b quite simply by using the sample means for the two groups. For simple linear regression models with numerical explanatory variables, we find a and b as follows.

First, we find the slope of regression line using the sample correlation coefficient r between the response variable Y and the explanatory variable X,

$$b = r\frac{s_y}{s_x}.$$

Here, s_y is the sample standard deviation of Y, and s_x is the sample standard deviation of X. Note that since s_x and s_y are always positive, the sign of b is the same as the sign of the correlation coefficient: $b > 0$ for positively correlated random variables, and $b < 0$ for negatively correlated variables. When $r = 0$ (i.e., the two variables are not linearly related), then $b = 0$. After finding the slope, we find the intercept as follows:

$$a = \bar{y} - b\bar{x},$$

where $\bar{y}$ and $\bar{x}$ are the sample means for Y and X, respectively. Then the least-squares regression line with intercept a and slope b can be expressed as

$$\hat{y} = a + bx.$$

For the blood pressure example, the sample correlation coefficient is $r = 0.84$; the sample standard deviation of blood pressure is $s_y = 4.94$, and the sample standard deviation of sodium chloride intake is $s_x = 3.46$. Therefore,

$$b = 0.84 \times \frac{4.94}{3.46} = 1.20.$$

For the observed data, the sample means are $\bar{y} = 135.68$ and $\bar{x} = 5.90$. Therefore,

$$a = 135.68 - 1.20 \times 5.90 = 128.60.$$

The linear regression model can be written as

$$\hat{y} = 128.60 + 1.20x.$$

We can now use this model to estimate the value of the response variable. For the individual i in the right panel of Fig. 11.7, the amount of daily sodium chloride intake is $x_i = 3.68$. The estimated value of the blood pressure for this person is

$$\hat{y}_i = 128.60 + 1.20 \times 3.68 = 133.02.$$

The actual blood pressure for this individual is $y_i = 128.3$. The residual therefore is

$$e_i = y_i - \hat{y}_i = 128.3 - 133.02 = -4.72.$$

In the right panel of Fig. 11.7, the arrow that represents the residual for observation i starts from 133.02 (i.e., estimated value according to the regression model) and ends at 128.3 (i.e., the observed value of the blood pressure variable), and its length is 4.72; the negative sign corresponds to a downward arrow.

We can also use our model for *predicting* the unknown values of the response variable (i.e., blood pressure) for all individuals in the target population. For example, if we know the amount of daily sodium chloride intake is $x = 7.81$ for an individual, we can predict her blood pressure as follows:

$$\hat{y} = 128.60 + 1.20 \times 7.81 = 137.97.$$

Of course, the actual value of the blood pressure for this individual would be different from the predicted value. The difference between the actual and predicted values

of the response variable is called the model **error** and is denoted as ε. In fact, the residuals are the observed values of ε for the individuals in our sample.

In the right panel of Fig. 11.7, stars show the predicted values of the response variable for values of the explanatory variable from 0 to 14. These are the expected blood pressure values for people in the population of interest with $0, 1, 2, \ldots, 14$ grams of daily sodium chloride intake. Note that in general, we should be cautious about using a regression model for prediction outside the population from which the sample is obtained. For this example, we obtained the data from the population of the elderly people (above 65 years old). Using our model for predicting blood pressure of young people based on their daily amount of sodium chloride consumption would not be appropriate since the relationship between blood pressure and sodium chloride intake might not be the same among the young population. For example, the relationship might be much weaker.

We should also be cautious about using the regression line for prediction outside the range of observed values of the explanatory variables. For the above example, the observed values of X in our sample are between 1 and 13. Using the regression line for prediction outside of this range is called **extrapolation**. As we move away from this range, the overall relationship between the response variable and the explanatory variable may change. In general, extrapolation far beyond the range of observed values for the explanatory variable is not recommended.

The predicted value for an individual with $x = 0$ (i.e., zero gram sodium chloride intake) is equal to the intercept:

$$\hat{y}_i = 128.60 + 1.20 \times 0 = 128.60 = a.$$

This is the point where the regression line intercepts the vertical axis. Therefore, the intercept is interpreted as the expected value of the blood pressure among people with zero sodium chloride diet. Of course, setting the value of the explanatory variable to zero does not always make sense. For example, if we use the weight of individuals as the explanatory variable for blood pressure, we cannot interpret the intercept as the expected value of the blood pressure among people whose weight is zero.

The interpretation of slope b for the above model is similar to the interpretation of slope for simple linear regression models with a binary explanatory variable: the slope represents the expected (average) amount of change in the response variable for one unit increase in the value of the explanatory variable. For binary variables, one unit increase in X meant moving from the group with $X = 0$ to the group with $X = 1$. For a numerical variable such as daily sodium chloride intake, one unit increase could mean increasing the daily amount of sodium chloride intake from 6 to 7 or from 10 to 11. For the above model, people whose daily sodium chloride intake is 7 grams are expected (i.e., on average) to have $b = 1.20$ units higher blood pressure compared to those whose daily intake is 6 grams. The same comment can be made for comparing people with 11 grams of daily intake to those with 10 grams of daily intake.

Finding confidence intervals for regression parameters α and β also remains as before. More specifically, the confidence interval for regression coefficient is ob-

Fig. 11.8 Using
R-Commander to fit a linear
regression for investigating
the relationship between
blood pressure (BP) and
sodium chloride intake
(salt)
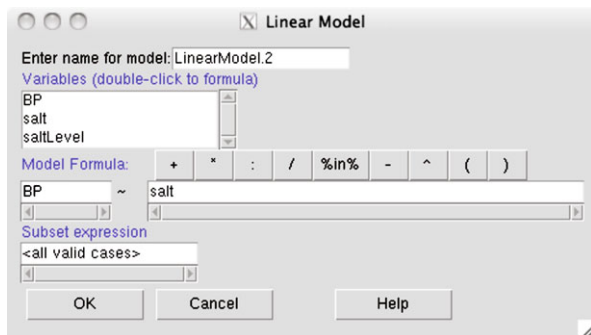
tained as follows:

$$[b - t_{\text{crit}} \times SE_b, b + t_{\text{crit}} \times SE_b].$$

Here, SE_b is the standard error of the regression coefficient and is calculated according to Eq. 11.3. We obtain t_{crit} from the t-distribution with $n - 2$ degrees of freedom for the given confidence level.

The steps for performing hypothesis testing regarding the linear relationship between the response and explanatory variables also remain the same. The null hypothesis is $H_0 : \beta = 0$, which indicates that the two variables are not linearly related. This corresponds to a horizontal regression line, whose slope is zero. To evaluate this hypothesis, we need to find the t-score first,

$$t = \frac{b}{SE_b}.$$

Then, we find the p-value (i.e., the observed significance level) by calculating the probability of as or more extreme values than the t-score under the null hypothesis, in the direction of the alternative hypothesis. To this end, we use the t-distribution with $n - 2$ degrees of freedom to find the lower tail probability of the t-score if $H_A : \beta < 0$, or its upper tail probability if $H_A : \beta > 0$, or two times the upper tail probability of its absolute value if $H_A : \beta \neq 0$.

The steps for using R-Commander to find the points estimates and confidence intervals of regression parameters, and performing hypothesis testing based on linear regression models are the same as what we discussed for simple linear regression models with binary explanatory variables. For the blood pressure example, click Statistics → Fit models → Linear regression. Then under Model Formula, enter BP ~ salt, as in Fig. 11.8, and press OK. Note that the name of the linear regression model is Linear-Model.2.

Figure 11.9 shows the output provided by R-Commander for LinearModel.2. In the Coefficients table, the first column (Estimate) provides the point estimates for the regression parameters: $a = 128.6$ and $b = 1.2$. The next column shows the corresponding standard errors. Dividing $b = 1.2$ by its standard error $SE_b = 0.16$ gives the t-score,

Fig. 11.9 Output of fitting a
linear regression using
R-Commander for
investigating the relationship
between blood pressure (BP)
and sodium chloride intake
(salt)

$$t = \frac{1.2}{0.162} = 7.4,$$

which is the same as the value given in the third column under `t value`. Finally,
the observed significance level p_{obs} is given in the last column. Here, p_{obs} is equal to
the upper tail probability of $|7.4|$ multiplied by 2. Because the p-value is extremely
small, we conclude that the observed association between the two variables, blood
pressure and daily amount of sodium chloride intake, is statistically significant. As a
result, we would feel comfortable to reject the null hypothesis, which indicates that
the two variables are not linearly related.

11.5 Goodness of Fit

When we fit a least-squares regression model to the observed data, R-Commander
provides some other important information about our model besides the
`Coefficients` table. Here, we focus on **R-squared**, which measures how well
the regression model fits the observed data. We denote this measure as R^2.
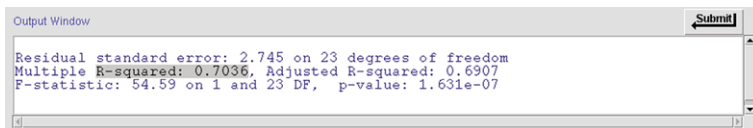
Recall that the residual sums of squares (RSS) quantifies the discrepancy between
the observed data and the regression line used to represent the data: the higher RSS,
the higher discrepancy. Therefore, we can interpret RSS as the **unexplained varia-
tion** in the response variable using the regression model.

Alternatively, we can interpret RSS as the **lack of fit** of the linear regression
model. In contrast, R^2 is a measure of goodness of fit; that is, how well our model
represents the observed data and explains the variation in the response variable. The
value of R^2 is between 0 and 1, and the better the model fits the data, the higher its
R^2 is.

The total variation in the response variable before we fit the regression line is
called the **Total Sum of Squares** (TSS) and is calculated as the squared deviations
of each observed value of the response variable, y_i, from its sample mean $\bar{y}$:

$$TSS = \sum_i^n (y_i - \bar{y})^2.$$

The fraction RSS/TSS can be interpreted as the percent of total variation that was
not explained by the regression model.

Fig. 11.10 Measures of goodness of fit provided by R-Commander for linear regression models. Here, our model uses `salt` as the explanatory variable for `BP`. For this model, $R^2 = 0.70$

In contrast, $1 - RSS/TSS$ is fraction of total variation explained by the model. This fraction is R^2, which measures the goodness of fit for the regression model,

$$R^2 = 1 - \frac{RSS}{TSS}.$$

In Fig. 11.10, the value of R^2 provided by R-Commander is highlighted for the linear regression model with blood pressure as the response variable and daily sodium chloride intake as the explanatory variable. For this model, $R^2 = 0.70$. Therefore, 70% of the total variation in blood pressure can be explained by the daily amount of sodium chloride a person consumes. The remaining 30% of the total variation cannot be explained by this model and is regarded as random.

For simple linear regression models with one numerical explanatory variable, R^2 is equal to the square of the correlation coefficient r.

For the above blood pressure example, the sample correlation coefficient between blood pressure and daily sodium chloride intake is $r = 0.84$. Therefore, the R^2 is

$$R^2 = 0.84^2 = 0.70.$$

This is the same value provided by R-Commander in Fig. 11.10.

While R^2 provides useful information about the fit of the regression model, one should be cautious about overstating its importance. Having a large R^2 only means that the model provides estimates close to the observed values for individuals in our sample. This may or may not translate to better predictions for other individuals that are not included in our sample. Moreover, even when the value of R^2 is small, the model could still be useful for predicting unknown values of the response variable, especially when we consider the alternative option of not using the model (and the explanatory variable) for prediction.

11.6 Model Assumptions and Diagnostics

Statistical inference using linear regression models is based on several assumptions. Violating these assumptions could lead to wrong conclusions.

Linearity   The most important assumption of linear regression model is that the relationship between the explanatory variable X and the response variable Y is **linear**. For simple problems discussed in this book, we can visually evaluate the appropriateness of this assumption using the scatterplot of Y versus X such as the one shown in Fig. 11.6.

When the linearity assumption does not hold, we might still be able to use linear regression models after transforming the original variables. Common transformations are logarithm (usually for the response variable), square root, and square (usually for predictors). For example, we can create a new variable $X^2 = X_1^2$ and include it in the regression model to account for possible nonlinear (in this case, parabolic) relationship between the response variable Y and the predictor X_1,

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon.$$

We should be cautious about interpreting the model parameters. Note that the role of the quadratic term X_1^2 is to capture possible nonlinearity in the relationship between Y and X_1. In this case, the null hypothesis $H_0 : \beta_2 = 0$ indicates that the relationship is not quadratic. Fitting such models is discussed in the next section.

Finding the right transformation is not trivial. Additionally, variable transformations make the interpretation of the results difficult. Therefore, these techniques are usually more appropriate when our objective is predicting the unknown values of the response variables as opposed to explaining its relationship with a set of explanatory variables. For more discussion, refer to Harrell (2001) [10].

Independence   Another important assumption is that the observations are **independent**, which is a reasonable assumption if we use simple random sampling to select individuals that are not related to each other and if we do not obtain multiple observations from the same individual. However, we sometimes sample related subjects (e.g., siblings) in groups. Also, we sometimes select unrelated subjects, but obtain multiple measurements (e.g., over a period of time) of the response variable for each subject. For example, when evaluating the effect of different diets on blood pressure, we might obtain multiple measurements of blood pressure for each person repeatedly over six months. For such data, we need to use regression models that take the dependencies among observations into account. When multiple observations are obtained over time, we typically use a class of statistical models called **longitudinal models**.

Constant Variance and Normality   Using linear regression models also involves some assumptions regarding the probability distribution of the response variable Y, which is the main random variable in regression analysis. However, because of the connection between the response variable and the error term, ε, according to the

linear regression model (Eq. 11.2), it is common to treat ε as a random variable (its values change from one individual to another) and specify these assumptions in terms of the probability distribution of ε.

To make the statistical inference methods we discussed earlier in this chapter valid, it is common to assume that the error term is normally distributed with mean 0,

$$\varepsilon \sim N\left(0, \sigma^2\right).$$

Minor deviations from normality will not have a substantial impact on the results as long as the sample size is relatively large. The important aspects of this assumption are its specifications of the population mean and variance of ε. The population mean is assumed to be zero, so we expect the errors based on the regression line for the whole population to be centered on zero. The population variance of ε is σ^2, which is of course unknown. What we actually mean by this assumption is that whatever the value of this parameter, it does not change for different values of the explanatory variable, e.g., σ^2 remains the same for $x = 5$ and $x = 10$. Informally, this means that we expect that the variation of the actual values of the response variable around the regression line remains the same regardless of the value of the explanatory variable. This is called the **constant variance** assumption, which is also known as the **homoscedasticity** assumption. (Heteroscedasticity refers to situations where the constant variance assumption is violated.) To check the validity of these assumptions, we examine the residuals e that are observed values of the random variable ε.

To illustrate how we check the assumptions regarding ε, we use the blood pressure example with daily amount of sodium chloride intake as the explanatory variable. In R-Commander, make sure `RegModel.2` is the active model, then click `Models → Graphs Basic diagnostic plots`. This creates several model diagnostic plots.

The residual plot in the left panel of Fig. 11.11 shows the residuals e versus the estimated (fitted) values of the response variable $\hat{y}$. The horizontal line represents the regression line. The plot shows that the residuals are scattered randomly around the horizontal dashed line at zero without any detectable pattern. In this figure, the solid line shows the overall pattern of the residuals. This line should remain close to the horizontal dashed line. Moreover, it is important that we do not see any non-random pattern in the residual plot. For example, we should not see small variations around the horizontal line in one region and high variations in another region. For illustration purposes, the right panel of Fig. 11.11 shows a residual plot where the variability of the residuals around the horizontal line changes from one region to another. More specifically, for this example, residuals become more dispersed around the horizontal line as we move from small to large fitted values. In this case, the constant variance assumption is violated.

When the constant variance assumption does not hold, we can sometimes *stabilize* the variance using simple transformations of the response variable so the variance becomes approximately constant. For example, instead of Y, we could use $\sqrt{Y}$ (usually when Y is a count variable), or $\log(Y)$ in the regression model. Another strategy is to use *weighted least squares* (not discussed in this book) instead of the standard least squares approach.
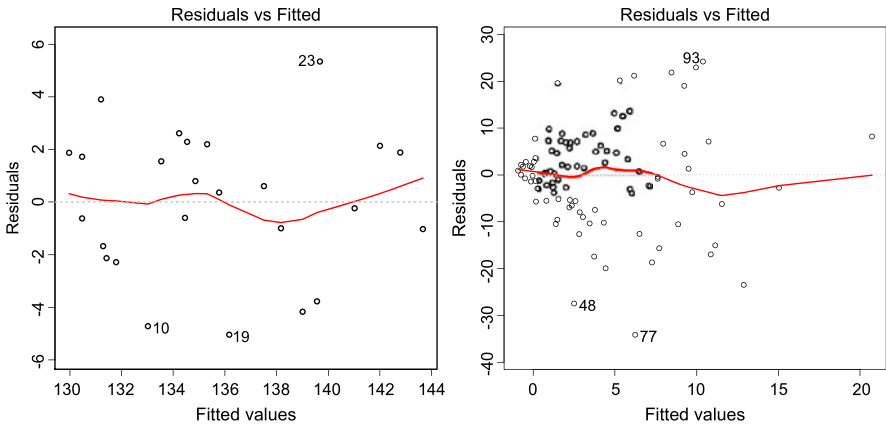
Fig. 11.11 *Left panel:* The residual plot for the blood pressure example (with daily amount of sodium chloride intake as the explanatory variable) to assess the assumptions of linear regression models related to the error term. Here, the scatter plot of residuals vs. the estimated (fitted) values is shown. The *horizontal axis* represents the regression line. The *solid line* on the plot shows the overall pattern of the residuals. The plot shows that the residuals are scattered randomly around the *horizontal dashed line* at zero without any detectable pattern. *Right panel*: An illustrative example, where the constant variance assumption is violated. Here, the residuals become more dispersed around the *horizontal line* as we move from small to large fitted values

In Fig. 11.11, the observations with large residuals (in absolute value) are identified by their row numbers. For these observations, the relationship between the response variable and the explanatory variable does not follow the overall pattern closely. We usually regard such observations as outliers and investigate them further to make sure that they are measured and recorded correctly. Again, we should not remove outliers from the data unless we are absolutely sure that they are recorded by mistake. (Occasionally, we remove outliers temporarily and refit the model to examine the extent of their influence on the regression model.)

We use the residuals as the observed values of the error terms, ε, to estimate its unknown population standard deviation σ as follows:

$$SE_e = \sqrt{RSS/(n-2)}.$$

We refer to SE_e as the **regression standard error**. This is in fact the numerator in Eq. 11.3 for the standard error of the regression coefficient. Therefore, we can rewrite Eq. 11.3 as follows:

$$SE_b = \frac{SE_e}{\sqrt{\sum_i (x_i - \bar{x})^2}}. \tag{11.5}$$

The regression standard error for the blood pressure example with the daily sodium chloride intake as the explanatory variable is $SE_e = 2.745$. This value is shown in Fig. 11.10. Divide this value by $\sqrt{\sum_i (x_i - \bar{x})^2}$ to see that the result is equal to 0.162, the standard error of the regression coefficient.

11.7 Multiple Linear Regression

So far, we have focused on linear regression models with only one explanatory variable. In most cases, however, we are interested in the relationship between the response variable and multiple explanatory variables. Even if we are interested in the relationship between the response variable and only one explanatory variable, very often we need to account for the effect of other important variables that might influence our inference. If our objective is to predict unknown values of the response variable, we might be able to improve prediction accuracy by including multiple predictors in the linear regression model. Models with multiple explanatory variables or predictors are called **multiple linear regression** models.

As an example, suppose that we want to examine the relationship between the birthweight of babies and the smoking status of their mothers during pregnancy. We might however believe that mother's age at the time of pregnancy is an important factor that should be taken into account. To this end, we need to evaluate the relationship between birthweight and smoking status among mothers with similar age.

Alternatively, suppose that our objective is to predict birthweight given age and smoking status of mothers; that is, we are interested in predicting birthweight if, for example, we are told that the mother has been smoking during pregnancy and she is 30 years old. Again, we need to include both age and smoking status in our model.

> In practice, we need to specify our objective for building linear regression models clearly. If our objective is to examine possible relationships between the response variable and one or more explanatory variables, we should specify our hypothesis prior to our analysis. In this case, our decision to include an explanatory variable in the model must be hypothesis-driven and based on our domain knowledge. When testing a hypothesis, we should avoid finding our model through exploring all possible combinations of variables available to us. On the other hand, if our objective is to predict the unknown values of the response variable, we could use our domain knowledge to identify a set of promising predictors and then explore all possible combinations of these variables until we find a model that provide the best predictions.

Note that our criterion for finding the best regression model for prediction should be based on how the model predicts unknown values of the response variable (e.g.,

for the part of the population not included in our sample). To this end, a common criterion is the **mean squared error** (MSE), which measures how close the predicted values are to the actual values,

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2.$$

When used for evaluating predictions, MSE is sometimes called the Expected Prediction Error (EPE) [11] or the Mean Squared Error of Prediction (MSEP) [3]. Note that the sum is over all unknown values (i.e., the whole population). In practice, we cannot calculate MSE directly since the actual values of the response variable are unknown. We can however estimate it using a subset of our sample as the **test set**. These are observations we remove from our sample before fitting the regression model and treat them as future observations. That is, we pretend that we do not know the value of the response variable for these observations. We then use the remaining observations, called the **training set**, in our sample to fit the regression model and use the resulting model to predict the response variable for the test set. Suppose we have m observations in the test set, we estimate MSE as follows:

$$\widehat{\text{MSE}} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2.$$

Here, $\hat{y}_i$ is our prediction for the ith observation in the test set, and y_i is the actual value of the response variable for this observation. When we use regression models for prediction, we prefer models with small $\widehat{\text{MSE}}$.

While fitting a multiple linear regression model to the data follows the same principle (namely, minimizing the residual sum of squares) as what we used for simple linear regression model, estimating regression parameters and performing statistical inference using multiple linear regression models is slightly more complex than what we discussed for simple linear regression models. Here, we focus on using R-Commander to find parameter estimates and interpret the results.

A multiple linear regression model with p explanatory variables can be presented as follows:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon.$$

Here, Y is the numerical response variable, $X_1, \ldots, X_p$ are the explanatory variables, α is the intercept, $\beta_1, \ldots, \beta_p$ are the corresponding regression coefficients, and ε is the error term.

Using the observed data, we estimate the regression parameters $\alpha, \beta_1, \ldots, \beta_p$. To this end, we use the least-squares method as before. We denote the point estimates for these parameters $a, b_1, \ldots, b_p$, respectively. Using the point estimates, we can predict the value of the response variable for an individual whose measurements for the explanatory variables are $x_1, \ldots, x_p$:

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p.$$

For the ith observed data, the difference between the actual value of the response variable, y_i, and its estimate according to the above linear regression model, $\hat{y}_i$, is the residual e_i,

$$e_i = y_i - \hat{y}_i.$$

As before, we measure the discrepancy between the model and the observed data using RSS, which is the sum of the square of the residuals. As before, R^2 measures the goodness of fit for the regression model,

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Using RSS, we can find the regression standard error (i.e., the estimate of σ) as follows:

$$SE_e = \sqrt{RSS/(n - p - 1)},$$

where n is the sample size, and p is the number of explanatory variables in the model.

For each regression coefficient β_j (i.e., the coefficient of the explanatory variable X_j), we find the standard error SE_{b_j} along with its point estimate b_j. Similar to simple linear regression models, confidence intervals for β_j are obtained as follows:

$$[b_j - t_{\text{crit}} \times SE_{b_j}, b_j + t_{\text{crit}} \times SE_{b_j}].$$

Here, however, we obtain t_{crit} from the t-distribution with $n - p - 1$ degrees of freedom for the given confidence level c.

The steps for performing hypothesis testing regarding the β_j is also similar to what we discussed for simple linear relationship models; the null hypothesis is $H_{0j} : \beta_j = 0$; we evaluate the null hypothesis by finding the t-score,

$$t_j = \frac{b_j}{SE_{b_j}}.$$

However, for multiple linear regression models, we obtain p-values (i.e., the observed significance level) using the t-distribution with $n - p - 1$ degrees of freedom.

As mentioned above, we use R-Commander to estimate the parameters of multiple linear regression models and use them to perform statistical inference. For the birthweight example, make sure `birthwt` is the active data set, then click `Statistics` → `Fit models` → `Linear model`. Then under `Model Formula`, enter bwt $\sim$ age + smoke, as in Fig. 11.12. (You can either type the model formula or double click on `bwt`, `age`, and `smoke` in that order; R-Commander then enters the first variable on the left-hand side of the $\sim$ symbol and the rest on its right-hand side.) By doing so, we specify the following multiple linear regression model:

$$\text{bwt} = \alpha + \beta_1 \text{age} + \beta_2 \text{smoke} + \varepsilon.$$

Notice that the intercept is added to the model automatically. If, for some reason, you want to suppress the automatic addition of the intercept in the regression model

Fig. 11.12 Fitting a multiple linear regression model in R-Commander. Here, birthweight (bwt) of babies is the response variable. Mother's age and smoking status (smoke) are used as explanatory variables

Fig. 11.13 Output of R-Commander for fitting a multiple linear regression model to the birthwt data to examine the linear relationship between birthweight (bwt) of babies and two characteristics of their mother, age and smoking status (smoke)

(this is not recommended in general), you can enter the model as bwt ∼ 0 + age + smoke.

The results of the multiple linear regression model are given in the Output window and in Fig. 11.13. The Coefficients table provides the estimates (Estimate), standard errors (Std. Error), t-scores (t value), and p-values (Pr(>|t|)) for the intercept and the regression coefficients of mother's age and her smoking status (smoke). Using the point estimates for regression parameters, we can predict birthweight for a baby by knowing her mother's age and smoking status as follows:

$$\widehat{\text{bwt}} = 2791 + 11 \times \text{age} - 278 \times \text{smoke}.$$

Therefore, if the mother is 30 years old (age=30) and she has been smoking during the pregnancy (smoke=1), our estimate (prediction) for the birthweight of her baby is

$$\widehat{\text{bwt}} = 2791 + 11 \times 30 - 278 \times 1$$
$$= 2843.$$

This can be interpreted as our estimate of the expected birthweight for babies whose mothers are 30 years old and smoke during pregnancy. That is, for these mothers, the expected (average) birthweight of babies is 2843 grams.

The intercept in multiple linear regression model is the expected (average) value of the response variable when all the explanatory variables in the model are set to zero simultaneously. In the above example, the intercept is $a = 2791$, which is obtained by setting age and smoking to zero. We might be tempted to interpret this as the average birthweight of babies for nonsmoking mothers (smoke=0) with age equal to zero. In this case, however, this is not a reasonable interpretation since mother's age cannot be zero.

For multiple linear regression models, we use b_j to denote the point estimate of the regression coefficient β_j. We interpret b_j as our estimate of the expected (average) change in the response variable associated with a unit increase in the corresponding explanatory variable X_j while all other explanatory variables in the model remain fixed.

For the above birthweight example, the point estimate of the regression coefficient for age is $b_1 = 11$ (Fig. 11.13). Therefore, we expect that the birthweight of babies increase by 11 grams as the mother's age increases by one year among mothers with the same smoking status. If we select two groups of mothers with the same smoking status from the population where the first group includes mothers who are, for example, 27 years old, and the second group includes mothers who are 28 years old, the average birthweight for the second group is 11 grams higher than the average birthweight in the first group according to our model.

For this model, the estimate of the regression coefficient for smoke is $b_2 = -278$. Therefore, the expected birthweight decreases by -278 grams associated with one unit increase in the value of the variable smoke among mothers with the same age. In this case, because smoke is a binary variable, one unit increase in its value means changing the smoking status from nonsmoking (smoke=0) to smoking (smoke=1). Note that the age of mothers should remain fixed to make this interpretation valid. For example, if we divide mothers who are 32 years old into two groups according to their smoking status, the expected weight of birthweight for smoking mothers is 278 grams less than that of nonsmoking mothers.

Figure 11.14 shows the above multiple linear regression model, where the linear relationship between birthweight and age is captured by a separate regression line for each smoking status. In this plot, nonsmoking mothers are shown as circles, while smoking mothers are shown as squares. The dashed line shows the regression line among nonsmoking mothers, and the solid line shows the regression line among the smoking mothers. Note that the slopes of the two lines are the same and are equal to 11. Therefore, the expected change in birthweight associated with one unit increase in age remains the same regardless of smoking status. On the other hand, the vertical distance between the two lines remains equal to 278 over all possible values of age. Therefore, for any given age, we expect the same difference in birthweight between children of smoking and nonsmoking mothers.

Fig. 11.14 Presenting the multiple linear regression model fitted to the `birthwt` data. Here, nonsmoking mothers are shown as *circles*, while smoking mothers are shown as *squares*. The *dashed line* shows the regression line among nonsmoking mothers, and the *solid line* shows the regression line among the smoking mothers
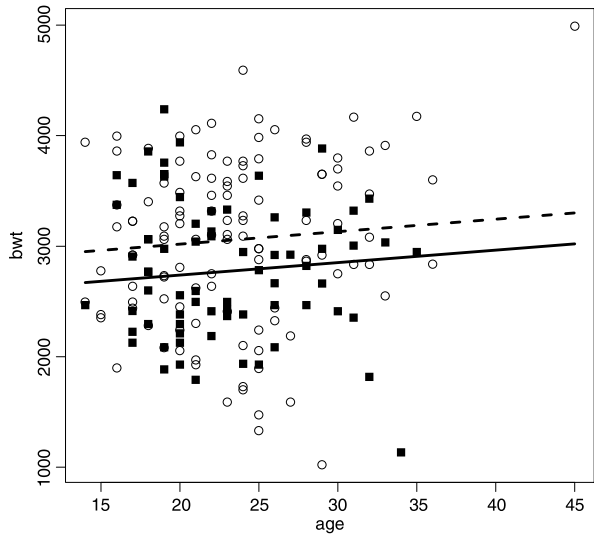
In multiple linear regression models, we usually assume that the effects of explanatory variables on the response variable are **additive**. This means that the expected change in the response variable corresponding to one unit increase in one of the explanatory variables remains the same regardless of the values of other explanatory variables in the model. As a result, when two (or more) explanatory variables change simultaneously, their overall effect on the response variable is the sum of their individual effects. For example, if we change mother's age from 27 to 28 and change smoking status from 0 to 1, the expected value of birthweight (in grams) changes by $11 + (-278) = -267$.

The coefficient table in Fig. 11.13 provides the standard errors along with the point estimate of the each regression coefficient. For this example, the standard errors are roughly $SE_{b_1} = 10$ and $SE_{b_2} = 107$. Suppose that we are interested in the 95% confidence intervals for β_1 and β_2. To this end, we first need to find t_{crit} for 0.95 confidence level from the t-distribution with $df = 189 - 2 - 1 = 186$. (Note that the sample size is $n = 189$ and the number of explanatory variables is $p = 2$.) Using R-Commander, we find $t_{crit} = 1.97$ for 0.95 confidence level. (The process of finding t_{crit} is the same as what we discussed for the population mean in Chap. 6.) Therefor, the 95% confidence interval for regression coefficient of age is

$$[11 - 1.97 \times 10, 11 + 1.97 \times 10] = [-8, 31].$$

Likewise, the 95% confidence interval for the regression coefficient of `smoke` is

$$[-278 - 1.97 \times 107, -278 + 1.97 \times 107] = [-489, -67].$$

As before, we could obtain confidence intervals for regression parameters in R-Commander by clicking `Models → Confidence intervals` and setting the confidence level. (The default is 0.95.)

According to the above results, at 0.95 confidence level, our expected change in birthweight associated with one year increase in mother's age among mothers with the same smoking status is somewhere between -8 and 31 grams. Note that this range includes zero, which is the value specified by the null hypothesis. More formally, considering the p-value of 0.25 provided in Fig. 11.13 for the regression coefficient for age, we cannot reject its corresponding null hypothesis at commonly used significant levels (0.01, 0.05, and 0.1). The null hypothesis in this case states that birthweight and age are not linearly related ($\beta_1 = 0$). We can, however, reject the null hypothesis for the regression coefficient for the smoking status at 0.05 level and conclude that the result of this test is statistically significant.

For the above example, suppose we remove the variable `age` from the model and fit a new linear regression model with the variable `smoke` only. By doing so, our inference regarding the linear relationship between birthweight and smoke changes. In this case, by including `smoke` as the only explanatory variable, the estimate of the regression coefficient for this variable changes from -278 to -283, and the p-value changes from 0.010 to 0.009.

For the above example, including and removing age did not have a substantial impact on our inference regarding the relationship between birthweight and smoking status. In some situations, the impact of adding or removing an explanatory variable could be drastic. As an example, suppose we want to examine the relationship between height and percent body fat among men. To do this, we use the `bodyfat` data set (discussed in Chap. 3), which consists of measurements of percent body fat, age in years, weight in pounds, height in inches, and abdomen circumference in inches for 252 men. (Follow the steps discussed in Chap. 3 to load the data in R-Commander.)

We build two linear regression models. In the first model, we only include `height` as the explanatory variable. In the second model, we include `height` and `abdomen` both. The steps for fitting linear regression models are the same as before.

Figure 11.15 shows the results for the first model, where we only include `height` as the explanatory variable in the model. As we can see, the p-value for testing the hypothesis regarding the linear relationship between percent body fat and height is 0.15, so the result is not statistically significant at 0.1 level. However, when we include `abdomen` along with `height` in the model (Fig. 11.16), the linear relationship between percent body fat and height becomes statistically significant at any
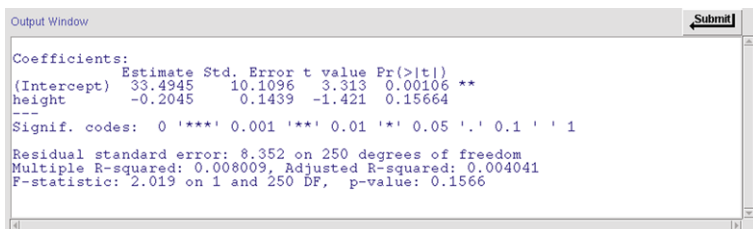
Fig. 11.15 The results of fitting a simple linear regression model to predict percent body fat with `height` as the only explanatory variable

```
Output Window                                                      [Submit]

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.31075   6.04265  -2.368   0.0186 *
height       -0.37053   0.08122  -4.562 7.95e-06 ***
abdomen       1.64707   0.07074  23.283  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.695 on 249 degrees of freedom
Multiple R-squared: 0.6878, Adjusted R-squared: 0.6853
F-statistic: 274.2 on 2 and 249 DF,  p-value: < 2.2e-16
```

Fig. 11.16 The results of fitting a multiple linear regression model to predict percent body fat by using both `height` and `abdomen` as explanatory variables

commonly used significance level; the p-value in this case reduces to an extremely small number (7.95×10^{-6}), and the estimate of the regression coefficient changes from -0.20 to -0.37.

The above results show that the importance and significance of one explanatory variable can be affected by presence or absence of other explanatory variables in the model. In this example, while height by its own does not have a statistically significant linear relationship with percent body fat, the relationship becomes quite significant among men with the same abdomen circumference. (Recall that when interpreting regression coefficient of one explanatory variable in multiple linear regression, we assume that all other explanatory variables are fixed at some specific values.)

> Our inference regarding the relationship between the response variable and an explanatory variable could depend on what other variables we include in the model. Therefore, we should choose the set of explanatory variables very carefully when we perform statistical inference using multiple linear regression models.

11.8 Advanced

In this section, we discuss linear regression models with interaction terms. We also discuss some commonly used R functions for linear regression analysis.

11.8.1 Interaction

In the previous section, we mentioned that the usual assumption in multiple linear regression models is that the effects are additive. If we believe that the effects are not additive (i.e., the effect of one explanatory variable X_1 on the response variable depends on the value of another explanatory variable X_2 in the model), we can still use linear regression models by including a new variable $X_3 = X_1 X_2$,

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon.$$

The term $X_1 X_2$ is called the **interaction term**. We refer to β_1 and β_2 as the **main effects**, and refer to β_{12} as the **interaction effect**.

As before, we use the least-squares method to estimate the model parameters,

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2.$$

When we include an interaction term in our model, we should be cautious about how we interpret model parameters. For simplicity, we assume that X_2 in the above model is binary so the value of x_2 can be either 0 or 1. When $x_2 = 0$, our estimate of the response variable is as follows:

$$\hat{y} = a + b_1 x_1.$$

The slope of the least-squares regression line is b_1. Therefore, our estimate of the response variable changes by b_1 units for one unit increase in x_1, *when we fix x_2 at zero*.

On the other hand, when $x_2 = 1$, our estimate of the response variable is

$$\hat{y} = a + b_1 x_1 + b_2 + b_{12} x_1.$$

We can rewrite the above equation as follows:

$$\hat{y} = (a + b_2) + (b_1 + b_{12}) x_1.$$

In this case, the slope of the least-squares regression line is $b_1 + b_{12}$, which means that our estimate of the response variable changes by $b_1 + b_{12}$ units for one unit increase in x_1, *when we fix x_2 at one*. Notice that the effect of X_1 on the response variable Y depends on the value of X_2.

As an example, we fit a multiple linear regression model to the birthweight data with `bwt` as the response variable. As before, we include both `age` and `smoke` as explanatory variables, but this time we include their interaction effect in the model,

$$\texttt{bwt} = \alpha + \beta_1 \texttt{age} + \beta_2 \texttt{smoke} + \beta_{12} \texttt{age} \times \texttt{smoke} + \varepsilon.$$

Repeat the steps for using R-Commander to fit multiple linear regression models, but this time enter `bwt ~ age * smoke` under `Model Formula` (Fig. 11.17). When you use "*" instead of "+" to separate the explanatory variables, R-Commander includes the main effects automatically along with the interaction effect in the model. The results are shown in Fig. 11.18. Note that the interaction term is shown as `age:smoke`. For this model, the point estimates of model parameters are $b_0 = 2406$, $b_1 = 28$, $b_2 = 798$, and $b_{12} = -47$. Therefore, we use the following equation to estimate birthweight:

$$\widehat{\texttt{bwt}} = 2406 + 28 \times \texttt{age} + 798 \times \texttt{smoke} - 47 \times \texttt{age} \times \texttt{smoke}.$$

Because `smoke` is a binary variable, the interpretation of model parameters are similar to what we discussed above. When `smoke = 0` (i.e., for nonsmoking mothers), the estimate of birthweight changes by $b_1 = 28$ grams for one year increase in mother's age. When `smoke = 1` (i.e., for smoking mothers), the estimate of birthweight changes by $b_1 + b_{12} = 28 - 47 = -19$ for one year increase in mother's age. That is, for smoking mothers, the estimate of birthweight decreases

Fig. 11.17 Fitting a multiple linear regression model with an interaction term in R-Commander. Here, birthweight (bwt) of babies is the response variable. Mother's age and smoking status (smoke) are used as explanatory variables. Note that unlike the additive model, we use "*" (instead of "+") to separate the two explanatory variables
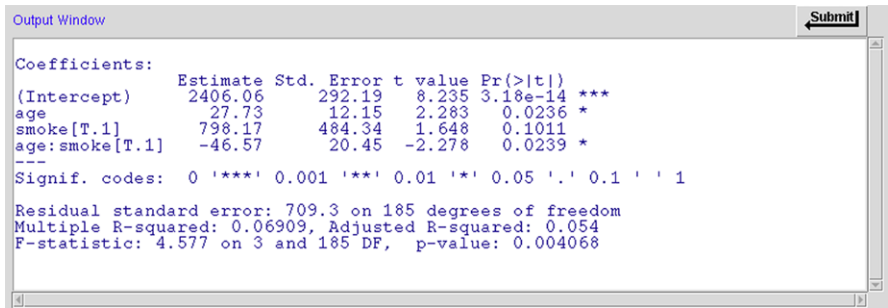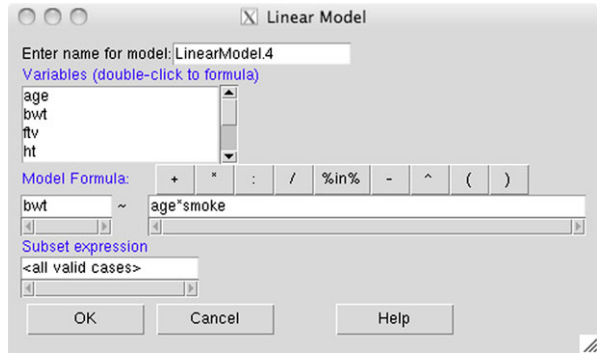
Fig. 11.18 Output of R-Commander for fitting a multiple linear regression model with an interaction term to the birthwt data to examine the linear relationship between birthweight (bwt) of babies and two characteristics of their mother, age and smoking status (smoke). The interaction term is shown as age:smoke

as mother's age increases. This is illustrated in Fig. 11.19. In this plot, nonsmoking mothers are shown as circles and smoking mothers are shown as squares. The dashed line shows the regression line for nonsmoking mothers, and the solid line shows the regression line for the smoking mothers. While the slope is positive among nonsmoking mothers, it becomes negative among smoking mothers. Compare this plot to Fig. 11.14, which we obtained for the additive (no interaction) model.

For the above model, although the estimate of the main effect for smoke is 798, we *cannot* interpret this as our estimate of the expected increase in birthweight for smoking mothers (smoke = 1) compared to nonsmoking mothers (smoke = 0) regardless of mother's age. The interpretation of smoking effect is slightly complex because it depends on mother's age, which can take many different values. Let us focus on mothers who are 23 years old at the time of pregnancy. This is the average age of mothers in our data. For nonsmoking mothers who are 23 years old, our estimate of birthweight (in grams) is

$$\widehat{\text{bwt}} = 2406 + 28 \times 23 + 798 \times 0 - 47 \times 23 \times 0$$
$$= 3050.$$

Fig. 11.19 Presenting the multiple linear regression model with an interaction term fitted to the `birthwt` data. Here, nonsmoking mothers are shown as *circles*, while smoking mothers are shown as *squares*. The *dashed line* shows the regression line among nonsmoking mothers, and the *solid line* shows the regression line among the smoking mothers
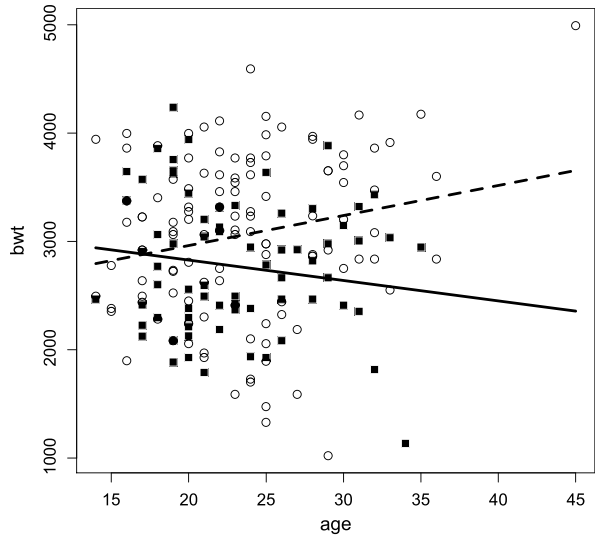
In contrast, our estimate of birthweight (in grams) for smoking mothers at the same age is

$$\widehat{\text{bwt}} = 2406 + 28 \times 23 + 798 \times 1 - 47 \times 23 \times 1$$
$$= 2767.$$

Therefore, the estimated birthweight reduces by $3050 - 2767 = 283$ grams for 23-year old smoking mothers compared to 23-year old nonsmoking mothers. Note that this estimate changes depending on mother's age.

11.8.2 Linear Regression Models in R

Fitting a linear regression model in R is straightforward. As an example, we model the relationship between percent body fat, `siri`, and height, using a simple linear regression model. The following commands install the `mfp` package using the `install.packages()` function, load it into R using the `library()` function, and make the data `bodyfat` available for analysis using the `data()` function:

```
> install.packages("mfp", dependencies = TRUE)
> library(mfp)
> data(bodyfat)
```

We set the `dependencies` to `TRUE` to install other packages that are related to `mfp` along with it.

To fit the least-squares regression model, use the `lm()` function:

```
> fit <- lm(siri ~ height, data = bodyfat)
```

The first argument of the function is the formula of the form of "response ~ explanatory variable". The second argument specifies the data set. By giving the name of the data set this way, we avoid witting the equation as `bodyfat$siri ~ bodyfat$height`.

The `fit` object now stores all the output from the linear regression model. Type `fit` to get the estimates of the estimates of α and β (i.e., regression parameters):

```
> fit

Call:
lm(formula = siri   height, data = bodyfat)

Coefficients:
(Intercept)         height
    33.4945        -0.2045
```

Using the `summary()` function, we can obtain the output similar to what R-Commander provides in above examples:

```
> summary(fit)

Call:
lm(formula = siri   height, data = bodyfat)

Residuals:
     Min        1Q  Median        3Q       Max
-19.5902   -6.7124   0.3966   6.0716   27.0919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.4945    10.1096    3.313  0.00106 **
height       -0.2045     0.1439   -1.421  0.15664

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 8.352 on 250 degrees
of freedom Multiple R-squared: 0.008009,
Adjusted R-squared: 0.004041
F-statistic: 2.019 on 1 and 250 DF, p-value: 0.1566
```

With the `names()` function, we can view all the information contained in the `fit` object:

```
> names(fit)
```

```
 [1]  "coefficients"   "residuals"
 [3]  "effects"        "rank"
 [5]  "fitted.values"  "assign"
 [7]  "qr"             "df.residual"
 [9]  "xlevels"        "call"
[11]  "terms"          "model"
```

Now we can use the $ operator to access information. For instance, suppose that we wanted the point estimates of α and β:

```
> fit$coefficients

(Intercept)        height
 33.4944938   -0.2044753
```

Likewise, the estimated response values for all people in our sample are stored in the `fitted.values` object within `fit`. Suppose that we wanted the estimates for the first five people:

```
> fit$fitted.values[1:5]

        1         2         3         4         5
19.64129  18.72115  19.94800  18.72115  18.92563
```

The differences between actual and estimated response values are stored in the `residuals` object within `fit`. The following command returns the residuals of the first five people:

```
> fit$residuals[1:5]

         1          2         3          4         5
-7.341291  -12.621152  5.351996  -8.321152  9.774373
```

Adding the least-squares line to the scatterplot is easy with the `abline()` function:

```
> plot(bodyfat$height, bodyfat$siri,
+      main = "Scatterplot for Percent Body Fat
       by Height",
+      xlab = "Height", ylab = "Percent Body Fat")
> abline(fit)
```

By default, `abline()` draws a solid line. We can set the line type to dashed line by using the option `lty=2`. The `lty` obtain defines the line type. In general, the

abline() function can be used to add a straight line to an existing plot. (You first need to create a plot before using abline.) For example, abline(h=2) draws a horizontal line two units above the origin, abline(v=-1) draws a vertical line one unit to the left of origin, and abline(a=-5, b=2) draws a line with intercept -5 and slope 2.

We can also fit a multiple linear regression model to the bodyfat data using abdomen circumference and height as explanatory variables. As before, we use the lm() function, but now we include both explanatory variables on the right-hand side of the formula. We separate the explanatory variables with plus signs:

```
> multReg <- lm(siri ~ height + abdomen, data = bodyfat)
> summary(multReg)

Call:
lm(formula = siri   height + abdomen, data = bodyfat)

Residuals:
    Min        1Q    Median        3Q       Max
-18.8513   -3.4825   -0.0156    3.0949   11.1633

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -14.31075    6.04265   -2.368    0.0186 *
height        -0.37053    0.08122   -4.562  7.95e-06 ***
abdomen        1.64707    0.07074   23.283   < 2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

Residual standard error: 4.695 on 249 degrees
of freedom Multiple R-squared: 0.6878,
Adjusted R-squared: 0.6853
F-statistic: 274.2 on 2 and 249 DF, p-value: < 2.2e-16
```

To include the interaction term between height and abdomen in the model, we can specify the regression model as siri ~ height * abdomen. In this case, R includes the main effects of the two variables automatically.

11.9 Exercises

1. We want to examine the relationship between body temperature Y and heart rate X. Further, we would like to use heart rate to predict the body temperature.
   (a) Use the "BodyTemperature.txt" data set to build a simple linear regression model for body temperature using heart rate as the predictor.

(b) Interpret the estimate of regression coefficient and examine its statistical significance.

(c) Find the 95% confidence interval for the regression coefficient.

(d) Find the value of R^2 and show that it is equal to sample correlation coefficient.

(e) Create simple diagnostic plots for your model and identify possible outliers.

(f) If someone's heart rate is 75, what would be your estimate of this person's body temperature?

2. We believe that gender might also be related to body temperature and could help us to predict its unknown values.

(a) Use the "BodyTemperature.txt" data set to build a multiple linear regression model for body temperature using heart rate and gender as predictors.

(b) How much R^2 did increase compared the above simple linear regression model?

(c) Explain the estimates of regression coefficients in plain language.

(d) Find the 95% confidence intervals for regression coefficients.

(e) If a woman's heart rate is 75, what would be your estimate of her body temperature? What would be your estimate of body temperature for a man whose heart rate is 75.

3. We would like to predict a baby's birthweight (bwt) before she is born using her mother's weight at last menstrual period (lwt).

(a) Use the birthwt data set to build a simple linear regression model, where bwt is the response variable and lwt is the predictor.

(b) Interpret your estimate of regression coefficient and examine its statistical significance.

(c) Find the 95% confidence interval for the regression coefficient.

(d) If mother's weight at last menstrual period is 170 pounds, what would be your estimate for the birthweight of her baby?

4. For the above problem, use both mother's weight at last menstrual period (lwt) and her smoking status (smoke) to predict birthweight.

(a) Interpret the estimates of regression coefficients and comment on their statistical significance.

(b) Find the 95% confidence interval for regression coefficients.

(c) If mother's weight at last menstrual period is 170 pounds and she was smoking during her pregnancy, what would be your estimate for the birthweight of her baby?

5. We want to predict percent body fat using the measurement for neck circumference.

(a) Use the bodyfat data set to build a simple linear regression model for percent body fat (siri), where neck circumference (neck) is the predictor. In this data set, neck is measured in centimeters.

(b) What is the expected (mean) increase in the percent body fat corresponding to one unit increase in neck circumference.

(c) Create a new variable, neck.in, whose values are neck circumference in inches. Rebuild the regression model for percent body fat using neck.in as the predictor.

(d) What is the expected (mean) increase in the percent body fat for one unit (1 inch $= 2.54$ centimeter) increase in `neck.in`.

(e) Compare the estimates of regression coefficient t_{obs} values and R^2 values between the two models.

6. Read the paper "A Critical Appraisal of 98.6°F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich" by Mackowiak et al. [19]. (The paper is available online at http://jama.ama-assn.org/cgi/reprint/268/12/1578.) They used a linear regression model to investigate the relationship between age and temperature. What did they find? They also used a linear regression model for the relationship between heart rate (response variable) and temperature. What was their conclusion? What was their point estimate of the regression coefficient for temperature?

Chapter 12
Clustering

12.1 Introduction

Linear regression models discussed previously are used to predict the unknown values of the response variable. In these models, the response variable has a central role; the model building process is guided by explaining the variation of the response variable or predicting its values. Therefore, building regression models is known as **supervised learning**. In contrast, building statistical models to identify the underlying structure of data (without focusing on a specific variable) is known as **unsupervised learning**. An important class of unsupervised learning is **clustering**, which is commonly used to identify subgroups within a population. In general, cluster analysis refers to the methods that attempt to divide the data into subgroups such that the observations within the same group are more similar compared to the observations in different groups.

For example, suppose that we believe that while European countries are different with respect to their protein consumption, they could be divided into several groups such that countries within the same group can be considered similar to each other in terms protein consumption. Here, we use the `Protein` data set we discussed earlier. Recall that this data set includes numerical measurements of the protein consumption from 9 different sources: `RedMeat`, `WhiteMeat`, `eggs`, `Milk`, `Fish`, `Cereals`, `Starch` (starchy foods), `nuts` (pulses, nuts, and oil-seeds), and `Fr.Veg` (fruits and vegetables). To start, suppose that we want to group countries according to their consumption of red meat (`redMeat`) and fish (`Fish`). More information about the data can be found at http://lib.stat.cmu.edu/DASL/Datafiles/Protein.html.

The core concept in any cluster analysis is the notion of similarity and dissimilarity. It is common to quantify the degree of dissimilarity based on a **distance** measure, which is usually defined for a pair of observations.

Fig. 12.1 Viewing the `Protein` data set in R-Commander. For 25 countries, there observations on the consumption of 9 different food groups

Table 12.1 Red meat and fish consumption in Albania and Austria

Countries	RedMeat	Fish
Albania	10.1	0.2
Austria	8.9	2.1

The most commonly used distance measure is the **squared distance**,

$$d_{ij} = (x_i - x_j)^2,$$

where d_{ij} refers to the distance between observations i and j, x_i is the value of random variable X for observation i, and x_j is the value for observation j.

In the `Protein` data set (Fig. 12.1), the first two countries are Albania and Austria. Suppose we want to measure their degree of dissimilarity (i.e., their distance) in terms of their consumption of red meat and fish (see Table 12.1). The squared distance between these two countries is $(10.1 - 8.9)^2 = 1.44$ in terms of red meat consumption and is $(0.2 - 2.1)^2 = 3.61$ in terms of fish consumption.

To find the overall distance between the two countries, we add the distances based on different variables:

$$d = 1.44 + 3.61 = 5.05.$$

In general, if we measure p random variables $X_1, \ldots, X_p$, the squared distance between two observations i and j in our sample is

$$d_{ij} = (x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2.$$

This measure of dissimilarity is called the **squared Euclidean distance**.

12.2 K-means Clustering

K-**means clustering** is a simple algorithm that uses the squared Euclidean distance as its measure of dissimilarity. We start by specifying the number of clusters (groups) *K*. This is the number of groups we believe exist in the population. Our goal is then to group the *n* observations in our sample into *K* clusters such that the overall measure of dissimilarity is small within groups and large between groups. Initially, we divide the observations into *K* groups randomly. Then the algorithm iteratively improves the clusters.

Let us define the **center** or **centroid** of each cluster as an imaginary observation whose measurements are the sample average of all observations in that cluster. For the food consumption example, suppose that the first cluster includes Albania and Austria only. The center of this cluster, denoted as $Center_1$, can be regarded as a fictitious country, whose red meat consumption is 9.50 (average of 10.1 and 8.9) and whose fish consumption is 1.15 (average of 0.2 and 2.1).

After randomly partitioning the observations into *K* groups and finding the center of each cluster, the *K*-means algorithm finds the best clusters by iteratively repeating these steps [11]:

1. For each observation, find its squared Euclidean distance to all *K* centers, and assign it to the cluster with the smallest distance.
2. After regrouping all the observations into *K* clusters, recalculate the *K* centers.

These steps are applied until the clusters do not change (i.e., the centers remain the same after each iteration).

Suppose that we want to cluster the countries into $K = 3$ groups based on their consumption of red meat and fish. In R-Commander, click `Statistics` → `Dimensional analysis` → `Cluster analysis` → `k-means cluster analysis`. Under `Variables`, select `Fish` and `RedMeat`. (Hold the `control` key and click on the name of variables.) Then use the slider to specify the `Number of clusters` as 3. Check the options `Print cluster summary` and `Assign clusters to the data set`. Finally, in the `Assignment variable` box, type `ClusterId` (Fig. 12.2).

R-Commander then creates a new variable `ClusterId` with the assignment of each country: 1, 2 or 3. (Try viewing the data set.) The number of countries assigned to the clusters is given in the *Output* window (Fig. 12.3). The sizes of Cluster 1, Cluster 2, and Cluster 3 are 11, 8, and 6, respectively. There is also a table showing the centroids of these cluster.

We can use a scatterplot to visualize the results of *K*-means. Use R-Commander to create a Scatterplot. Choose `Fish` under `x-variable` and `RedMeat` under `y-variable`. Uncheck all the options. Then click on `Plot by groups` and choose the newly created variable `ClusterId`. The resulting graph is shown in

Fig. 12.2 *K*-means clustering in R-Commander. Here, we want $K = 3$ clusters of the 25 countries based on `Fish` and `RedMeat` consumption. R-Commander will then provide a summary of each group in the *Output* window and create a new variable `ClusterId` with the observation's assignment
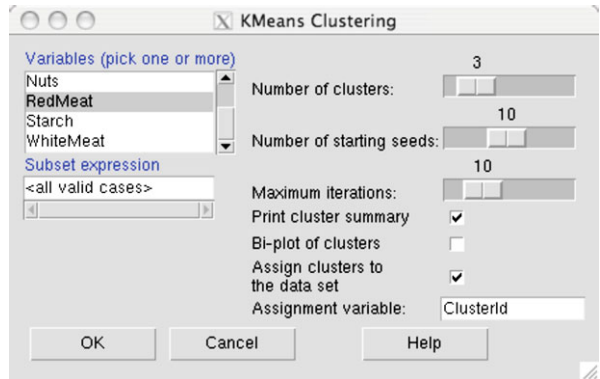
Fig. 12.3 Interpreting the results from *K*-means clustering. In the *Output* window, R-Commander provides the size and centers of the best $K = 3$ clusters, based on `Fish` and `RedMeat` consumption
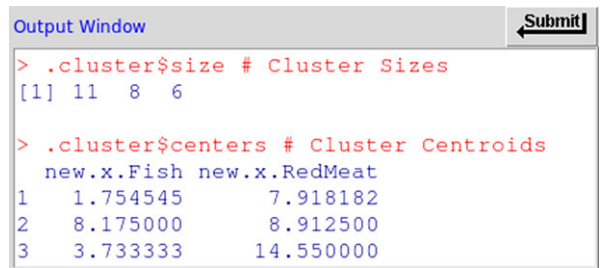
Fig. 12.4 Visualizing the results of *K*-means clustering with a scatterplot. The three clusters are represented by *red circles*, *green triangles*, and *blue crosses*. They clearly partition the countries into a group with a low consumption of fish and red meat, a group with a high consumption of fish, and a group with a high consumption of red meat
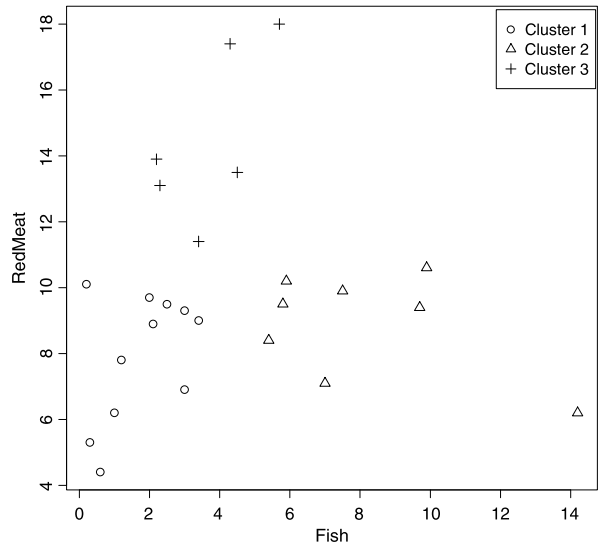
Fig. 12.4. The first cluster is represented by circles and contains countries whose consumption of both fish and red meat is relatively low. The second cluster is represented by triangles and includes countries whose consumption of both fish and

red meat is relatively high compared to the first group. Finally, the third cluster is represented by crosses and includes countries whose consumption of red meat tends to be higher compared to the other two groups.

12.3 Hierarchical Clustering

There are two potential problems with the K-means clustering algorithm. First, it is a **flat** clustering method. After observations are assigned to their clusters, they are all considered to be similar within the same cluster. That is, the observations are not further separated based on dissimilarity within a cluster. Secondly, we need to specify the number of clusters K a priori. Finding the appropriate number of clusters is not trivial, and the selected number has a substantial impact on the results.

An alternative approach that avoids these issues is **hierarchical clustering**. The result of this method is a **dendrogram** (a tree). The *root* of the dendrogram is its highest level and contains all n observations. The *leaves* of the tree are its lowest level and are each a unique observation.

There are two general algorithms for hierarchical clustering [11]:

- **Agglomerative** (bottom-up): We start at the bottom of the tree, where every observation is a cluster (i.e., there are n clusters). Then we merge two of the clusters with the smallest degree of dissimilarity (i.e., the two most similar clusters). Now we have $n - 1$ clusters. We continue merging clusters until we have only one cluster (the root) that includes all observations.
- **Divisive** (top-down): We start at the top of the tree, where all observations are grouped in a single cluster. Then we divide the cluster into two new clusters that are most dissimilar. Now we have two clusters. We continue splitting existing clusters until every observation is its own cluster.

Of the above two strategies, agglomerative algorithm is more common. Both algorithms, however, require a measure of dissimilarity between two clusters. In other words, we need to specify a distance measure for two clusters analogous to the distance measure we defined for two observations. For every pair of observations, where one is from cluster i, and the other one is from cluster j, we can find the squared Euclidean distances d_{ij}. Then we can use one of the following methods to calculate the overall distance between two clusters:

- *Single linkage* clustering uses the minimum d_{ij} among all possible pairs as the distance between the two clusters. This is the distance between two observations, one from each cluster, that are closest to each other.
- *Complete linkage* clustering uses the maximum d_{ij} as the distance between the two clusters. This is the distance between two observations, one from each cluster, that are furthest apart.

Fig. 12.5 Illustrating the difference between the single linkage method, the complete linkage method, and the centroid linkage method to determine the distance d_{ij} between the two clusters shown as circles and squares
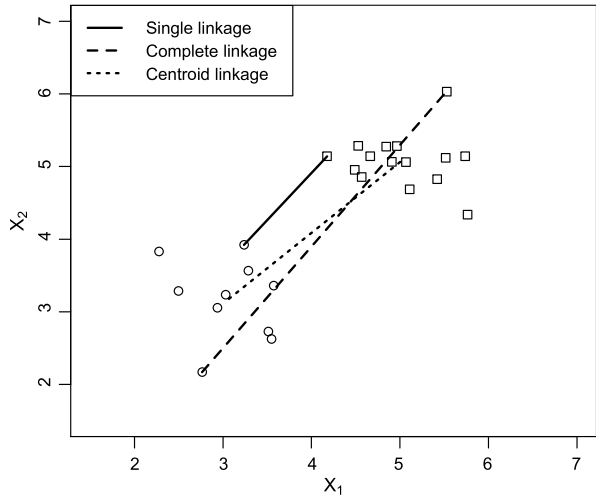
- *Average linkage* clustering uses the average d_{ij} over all possible pairs as the distance between the two clusters.
- *Centroid linkage* clustering finds the centroids of the two clusters and uses the distance between the centroids as the distance between the two clusters.

As an illustration of these methods, consider Fig. 12.5, which shows two clusters shown as circles and squares. The solid line shows the single linkage distance between the two clusters, the dashed line shows the complete linkage distance, and the dotted line shows the centroid linkage distance. Note that the dotted line connects the centers (as opposed to observations) of the two clusters. There are of course other ways for defining the distance between two clusters. However, the above measures are the most commonly used.

For example, let us perform complete linkage clustering to create a dendrogram of countries based on their protein consumption. Click `Statistics → Dimensional analysis → Cluster analysis → Hierarchical cluster analysis`. Select all nine food groups (hold the *control* key) for the `Variables`. Next, choose `Complete Linkage` as the `Clustering Method` and `Squared-Euclidean` as the `Distance Measure`. Lastly, make sure the option `Plot Dendrogram` is checked.

R-Commander then creates a dendrogram similar to the one shown in Fig. 12.6. The clusters seemed to be defined by geographic location: Balkan countries (Romania, Bulgaria, and Yugoslavia), Scandinavian countries (Finland, Norway, Denmark, and Sweden), Western European countries (UK, Belgium, France, Austria, Ireland, Switzerland, Netherlands, and West Germany), Eastern European countries (East Germany, Hungary, Czechoslovakia, Poland, Albania, USSR) and the Mediterranean countries (Portugal, Spain, Greece, Italy). (This data set was collected in 1973. Since then, some of these countries have changed or no longer exist.)

Using the results from hierarchical clustering, we can group observations into K clusters by cutting the dendrogram through K branches. For example, cutting

Fig. 12.6 The dendrogram resulting from complete linkage clustering of the 25 countries based on their protein consumption. The *dashed line* shows where to cut the dendrogram to create three clusters
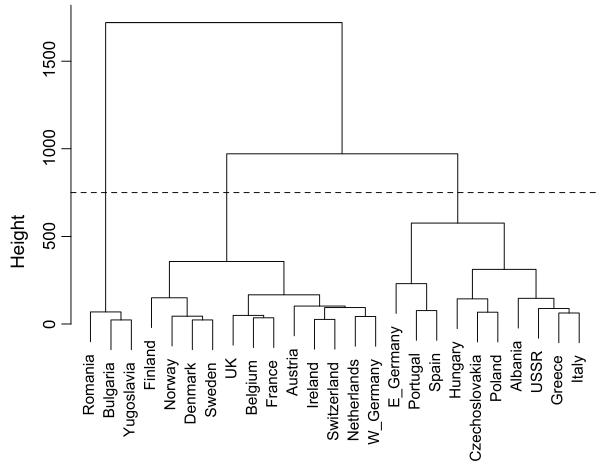
the dendrogram in Fig. 12.6 at the dashed line would create three clusters. The first cluster includes three of Balkan countries. The second cluster includes Scandinavian countries and mostly Western countries, and the last cluster mostly consists of countries from Eastern Europe and the Mediterranean.

To create *K* distinct clusters based on the result of hierarchical clustering, you can click `Statistics → Dimensional analysis → Cluster analysis → Summarize hierarchical clustering`. Next, follow the same steps, but this time choose `Add hierarchical clustering to data set` in order to create a new variable that identifies the clusters. In both cases, you need to specify the number of clusters.

12.4 Advanced

In this section, we discuss standardization of variables before clustering. We also discuss some useful commands to perform clustering in R.

12.4.1 Standardizing Variables Before Clustering

For distance-based clustering methods discussed in this chapter, it is common to standardize variable prior to clustering so that all variables contribute equally to the overall distance measure and have the same influence on the results. To this end, we divide each variable by its standard deviation as discussed in Sect. 2.6. This way, all variables have standard deviations of 1, so they become comparable.

As mentioned in Sect. 2.6, to standardize variables in R-Commander, click `Data → Manage variables in active data set → Standardize variables` and choose the variables you want to standardize. This will create a

new set of standardized variables, which have similar names to the original variables, but they all start with the prefix Z.

In general, not all variables have the same degree of importance in grouping observations into clusters. In some situations, giving all variables the same influence by standardizing them could lead to obscuring the patterns and misleading the clustering methods. Possible pitfalls of standardizing variables before clustering is illustrated in Fig. 14.5 of "The Elements of Statistical Learning" by Hastie et al. [11].

12.4.2 Clustering in R

In this section, we discuss some R functions for clustering. We start by importing the Protein data into R. Make sure that the file "Protein.txt" (available from the book website) is in your current directory, then enter the following command:

```
> Protein <- read.table("Protein.txt",
+     header = TRUE, sep = "")
```

Suppose that we want to cluster these European countries into three clusters according to their consumption of red meat and fish. We create a new object, x, that contains the columns RedMeat and Fish from the Protein data:

```
> x <- Protein[, c("RedMeat", "Fish")]
```

It is common to standardize the data prior to clustering so that the variables become comparable. We can standardize the data using the scale() function:

```
> x <- scale(x)
```

To use the K-means clustering, we use the kmeans() function:

```
> clus <- kmeans(x, centers = 3)
```

The first argument of the kmeans() function specifies the matrix that contains the data. Each row of this matrix corresponds to one of the observations (here, a European country), and each column corresponds to one of the variables. The second argument, centers, specifies the number of clusters. The output is assigned to a new object called clus.

The object clus is a list that contains the cluster ids, which is labeled as clusters, and the centroids, which are labeled as centers.

```
> clus$cluster

 [1]  3 3 1 3 3 2 2 2 1 2 3 1 3 3 2 3 2 3 2 2
```

```
[21] 1 1 3 1 3

> clus$centers

     RedMeat        Fish
1  1.4107826 -0.1618402
2 -0.2735221  1.1435597
3 -0.5705926 -0.7434033
```

We can append the cluster ids to the original data as follows:

```
> Protein$ClusterId <- clus$cluster
```

To visualize the three clusters, we can create the scatterplot of RedMeat by Fish and distinguish the clusters by using different symbols/colors similar to Fig. 12.4. We first create the plot frame without the observations as follows:

```
> plot(Protein$Fish, Protein$RedMeat,
+     type = "n", xlab = "Fish", ylab = "Red Meat",
+     xlim = c(0, 15), ylim = c(0, 20))
```

Note that we have used the option type='n', so the observations are not plotted; only the plot frame is created this way. The argument xlim and ylim are used to define the range of values for x (Fish) and y (Red Meat) axes to make sure that when we plot the observations, they do not fall outside of the plot frame. You can use the function range() for RedMeat and Fish to find the required limits.

Now we can use the function points() to add the observations to the plot. We add each cluster separately to the plot:

```
> points(Protein$Fish[Protein$ClusterId ==
+     1], Protein$RedMeat[Protein$ClusterId ==
+     1], pch = 1, cex = 1.5)
> points(Protein$Fish[Protein$ClusterId ==
+     2], Protein$RedMeat[Protein$ClusterId ==
+     2], pch = 2, cex = 1.5)
> points(Protein$Fish[Protein$ClusterId ==
+     3], Protein$RedMeat[Protein$ClusterId ==
+     3], pch = 3, cex = 1.5)
```

The first line plots those observations whose clusters id are equal to 1. For these observations the option pch=1 specifies the symbol (circles), and the option cex=1.5 specifies the size of the symbol. The second and the third lines add the second and third clusters to the plot with different symbols.

We can use the function legend() to add a legend to the plot in order to identify the clusters:

```
> legend("topright", legend = c("Cluster 1",
+     "Cluster 2", "Cluster 3"), pch = c(1,
+     2, 3))
```

The first argument provides the location (here, top right) of the legend in the plot. The second argument, legend, provides the names of the three clusters as a vector, and the third argument, pch, specifies the symbols used for each cluster. Use ?legend to learn more about this R function. Your final plot should be similar to Fig. 12.4.

To use hierarchical clustering, we use the function dist() to obtain the distance between observations (in a form of a matrix) and use the function hclust() to cluster the observations:

```
> d <- dist(x)
> clus.h <- hclust(d, method = "centroid")
```

The object d is a matrix that contains the Euclidean distances (note that we used *squared* Euclidean distances previously) of observations in x. (Recall that x itself was a matrix of standardized values of Red meat and Fish consumption.) The function hclust takes the distance matrix d as an input and performs hierarchical clustering. The argument method specifies the agglomeration method. Here, we use the centroid linkage method, where the distance between two clusters is calculated based on the distance between their centroids. For the complete list of options, enter the command ?hclust.

We have assigned the output of hierarchical clustering to a new object called clus.h. We can plot the dendrogram by simply using the plot() function with clus.h as its argument:

```
> plot(clus.h, labels = Protein$Country)
```

Here, the argument labels provides the name of each observation, and so they are identified on the dendrogram.

Now suppose that we want to use the result of the above hierarchical clustering and divide the observations into three clusters. We can first identify these clusters on the dendrogram by using the function rect.hclust():

```
> rect.hclust(clus.h, k = 3)
```

This function draws rectangles around the branches of a dendrogram highlighting the clusters. We can then obtain the cluster ids for the observations using the function cutree():

```
> clus.h.id <- cutree(clus.h, k = 3)
```

As before, we can add the cluster ids to the Protein data:

```
> Protein$HClusterId <- clus.h.id
```

12.5   Exercises

1. Use the K-means clustering method to divide Pima Indian women into three groups based on their age and BMI. Provide the centroid for cluster. Create a scatter plot, where each cluster is identified by a different symbol. How these three groups are different from each other based on their age and BMI?

2. In Sect. 3.6, we used the GBSG (German Breast Cancer Study Group) data set from the mfp package to create a new variable called rfs (recurrence-free survival) such that rfs="No" if the patient had at least one recurrence or died (i.e., cenc=1) and rfs="Yes" otherwise. Use the K-means clustering method to divide the patients into two groups based on their age, the size of tumor (tumsize), and the number of positive nodes (posnodal). Make sure the options Print cluster summary and Assign clusters to the data set are checked. Explain how the two groups are different using cluster specific summaries. R-Commander creates a new variable, which is called KMeans by default, to identify the two groups. Use this variable along with rfs to create a 2×2 contingency table where the rows show different clusters and the columns shows different values rfs. What are the sample proportion of recurrence-free survivals for the two groups. Compare the odds of recurrence-free survival between the two groups.

3. In R-Commander, click Data → Data in pacakges → Read data set from an attached package, then select the iris data from the datasets package. The data include the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of three species of iris. Use these measurements to divide the flowers into three groups. Make sure the options Print cluster summary and Assign clusters to the data set are checked. Use the centroids to explain how the three identified clusters are different. Between sepal and petal, which one seems to be more important in distinguishing the three clusters? Use the newly created variable that identifies the clusters (by default, R-Commander assigns the name KMeans to this variable) to create a contingency table where the rows are different clusters and the columns are different species. What is the connection between clusters and the type of flowers?

4. Repeat the above example, but this time use the hierarchical clustering approach with complete linkage as a measure of distance between clusters. Create three clusters (i.e., $K = 3$) based on your hierarchical cluster analysis and add a new variable to the data-identifying cluster ids. Using a contingency table, explain how the three identified clusters are related to the three species of flowers.

Chapter 13
Bayesian Analysis

13.1 Introduction

In Chap. 4, we discussed Bayes' theorem and mentioned that it is the basis of the Bayesian Statistics. In this chapter, we discuss Bayesian inference regarding the population proportion as an example for the application of Bayesian methods. To learn more about Bayesian data analysis, refer to Christensen et al. [6] or Gelman et al. [8].

13.2 A Simple Case of Bayesian Analysis for Population Proportion

We start our discussion with a simple illustrative example. Suppose that we are interested in finding the five-year survival rate (i.e., population proportion of survival) among breast cancer patients. We denote this unknown population proportion μ. For simplicity, we assume that the survival rate is either 0.75 or 0.85. Without any data, we think that both of these values are equally probable; that is, $P(\mu = 0.75) = P(\mu = 0.85) = 0.5$. Alternatively, we can write this as follows:

$$\frac{P(\mu = 0.85)}{P(\mu = 0.75)} = 1.$$

Now suppose that we take a random sample of $n = 20$ breast cancer patients from the population. We use Y to denote the number of survivals out of 20. From Chap. 5 we know that Y has a Binomial(n, μ) distribution (assuming that the patients are selected independently and they all have the same probability of survival). Therefore, if $\mu = 0.75$, the distribution of Y is Binomial(20, 0.75). If $\mu = 0.85$ on the other hand, the distribution of Y is Binomial(20, 0.85).

For our sample, we find that 18 of patients are still alive after five years: $Y = 18$. Our point estimate for the survival rate μ is therefore $p = 18/20 = 0.9$, where p is the sample proportion. Let us see how this information changes our mind about the

value of the population proportion, μ. For this, we use Bayes' theorem. Recall that Bayes' formula for two events E_1 and E_2 is

$$P(E_2|E_1) = \frac{P(E_1|E_2)P(E_2)}{P(E_1)}.$$

For our example, we can write Bayes' formula in terms of μ and Y, so E_1 corresponds to the event that $Y = 18$, and E_2 corresponds to the event that $\mu = 0.85$:

$$P(\mu = 0.85|Y = 18) = \frac{P(Y = 18|\mu = 0.85)P(\mu = 0.85)}{P(Y = 18)}.$$

Here, $P(\mu = 0.85|Y = 18)$ is the probability that the true value of the survival rate is 0.85 given the information that 18 people have survived (out of 20), $P(Y = 18|\mu = 0.85)$ is the probability that 18 people survive assuming that the true survival rate is 0.85, $P(\mu = 0.85)$ is the probability that the survival rate is in fact 0.85 (we assumed this probability is 0.5), and $P(Y = 18)$ is the probability that 18 people survive regardless of the what the probability of survival is (0.85 or 0.75).

As mentioned above, Y has Binomial(20, 0.85) distribution given that $\mu = 0.85$. Using R-Commander, we can find the probability of 18 survivals assuming that $\mu = 0.85$. Click Distributions $\rightarrow$ Discrete distributions $\rightarrow$ Binomial distribution $\rightarrow$ Binomial probabilities. Then, set Binomial trials to 20 and Probability of success to 0.85. R-Commander provides the probabilities for all possible values of Y. The probability for 18 survivals assuming that $\mu = 0.85$ is $P(Y = 18|\mu = 0.85) = 0.23$.

To find $P(Y = 18)$, we use the law of total probability (Eq. 4.7):

$$\begin{aligned}P(Y = 18) &= P(Y = 18|\mu = 0.85)P(\mu = 0.85)\\&+ P(Y = 18|\mu = 0.75)P(\mu = 0.75).\end{aligned}$$

Here, $P(Y = 18|\mu = 0.75)$ is the probability of 18 survivals assuming that the true value of the survival rate is $\mu = 0.75$. We find this probability using R-Commander by following the above steps, but this time we set Probability of success to 0.75. By doing so, we find $P(Y = 18|\mu = 0.75) = 0.07$. Therefore,

$$\begin{aligned}P(Y = 18) &= 0.23 \times 0.5 + 0.07 \times 0.5\\&= 0.15.\end{aligned}$$

Now we have all the information we need to find $P(\mu = 0.85|Y = 18)$:

$$\begin{aligned}P(\mu = 0.85|Y = 18) &= \frac{P(Y = 18|\mu = 0.85)P(\mu = 0.85)}{P(Y = 18)}\\&= \frac{0.23 \times 0.5}{0.15}\\&= 0.76.\end{aligned}$$

At the beginning (before observing any data), we believed that $\mu = 0.85$ with probability of 0.5. Knowing that 18 out of 20 people have survived, we increase this probability to 0.76.

By following similar steps, we find that $P(\mu = 0.75|Y = 18) = 0.24$. Given the observed data, we have reduced the probability of $\mu = 0.75$ from 0.5 to 0.24. Therefore, while we gave equal probabilities to both values 0.75 and 0.85 at the beginning, based on the new empirical evidence we observed, we increased the probability of $\mu = 0.85$ and decreased the probability of $\mu = 0.75$. This is intuitive, of course, because our point estimate for the survival rate is 0.9 (18 out of 20), which is closer to 0.85 than 0.75. We can use these updated probabilities and write

$$\frac{P(\mu = 0.85|Y = 18)}{P(\mu = 0.75|Y = 18)} = \frac{0.76}{0.24} = 3.2.$$

Therefore, given the observed data, the value 0.85 is 3.2 times more likely than 0.75.

13.3 Prior and Posterior Probabilities

The above example illustrates a simple application of Bayesian inference. For the population proportion μ, we refer to $P(\mu = 0.75)$ and $P(\mu = 0.85)$ as **prior probabilities**. These are probabilities we assign to possible values of μ before observing any data. Note that in practice, we might obtain these probabilities from previous studies. For example, two other research groups might have conducted similar studies in the past; one group estimated μ to be 0.75, and the other group estimated it to be 0.85, and we do not have any reason to prefer one estimate over the other. In this case, we want to conduct a new study, collect new empirical evidence, and estimate μ, but we want to take the available information regarding the value of μ into account.

We refer to $P(Y = 18|\mu = 0.85)$ as **likelihood**, i.e., how likely it is to see this specific data (18 survivals out of 20) if μ is in fact 0.85. We can express the probability of the specific data we have observed (i.e., 18 survivals out of 20) as a function of different values of μ. We refer to this function as the **likelihood function**. For the above example, the likelihood function is

$$P(Y = 18|\mu) = \begin{cases} 0.07, & \mu = 0.75, \\ 0.23, & \mu = 0.85. \end{cases}$$

We refer to the updated probability of μ after we observe the data as the **posterior probability** of μ. The posterior probabilities in the above example are $P(\mu = 0.75|Y = 18) = 0.24$ and $P(\mu = 0.85|Y = 18) = 0.76$, which are obtained after we observed 18 survivals among 20 patients. As mentioned above, we can use these posterior probabilities to write

$$\frac{P(\mu = 0.85|Y = 18)}{P(\mu = 0.75|Y = 18)} = \frac{0.76}{0.24} = 3.2.$$

This is known as the **posterior odds** (here, we find the odds of 0.85 over 0.75). Of course, based on what we discussed above, we can find the posterior odds as follows:

$$\frac{P(\mu=0.85|Y=18)}{P(\mu=0.75|Y=18)} = \frac{P(Y=18|\mu=0.85)\,P(\mu=0.85)/P(Y=18)}{P(Y=18|\mu=0.75)\,P(\mu=0.75)/P(Y=18)}$$

$$= \frac{P(Y=18|\mu=0.85)\,P(\mu=0.85)}{P(Y=18|\mu=0.75)\,P(\mu=0.75)}.$$

By canceling out the term $P(Y=18)$ from the numerator and denominator, we have

$$\frac{P(\mu=0.85|Y=18)}{P(\mu=0.75|Y=18)} = \frac{P(Y=18|\mu=0.85)}{P(Y=18|\mu=0.75)} \times \frac{P(\mu=0.85)}{P(\mu=0.75)}.$$

The term $P(\mu=0.85)/P(\mu=0.75)$ on the right-hand side of the above equation is called **prior odds** (here, we find the odds of 0.85 over 0.75). In our example, the prior odds is 1. The posterior odds is obtained by multiplying the prior odds by the following term:

$$\frac{P(Y=18|\mu=0.85)}{P(Y=18|\mu=0.75)} = \frac{0.23}{0.07}.$$

This term is in fact the ratio of two possible values for the likelihood function and is known as the **likelihood ratio**.

> The posterior odds is the product of the prior odds by the likelihood ratio.

13.4 The General Form of Bayesian Analysis for Population Proportion

For the example we have discussed so far, we have focused on a simple case where the population proportion could take one of two possible values, 0.75 and 0.85. Consequently, we specified the prior distribution by assigning probabilities to these two values only; that is, we used a discrete probability distribution with only two possible values. In general, the population proportion could take values from 0 to 1. Therefore, we need a continuous prior distribution whose range is from 0 to 1.

> The **beta** distribution, whose range is from 0 to 1, is commonly used as the prior distribution for the population proportion μ. The beta distribution is specified by two parameters, α and β, and is denoted as Beta(α, β). We refer to α and β as *shape 1* and *shape 2*, respectively. Both parameters must be positive numbers.

We can use R-Commander to plot different beta distributions by setting α and β to different values. For example, suppose that we want to plot Beta(8, 2). In R-Commander, click `Distributions` → `Continuous distributions` → `Beta distribution` → `Plot beta distribution` and set Shape 1

Fig. 13.1 Comparing the plots of the probability density function for a beta distribution with different parameter values. The *solid line* represents the pdf of Beta(1, 1). This distribution is known as the Uniform(0, 1) distribution. The *dashed line* represents the pdf of Beta(8, 2), and the *dotted line* represents the pdf of Beta(2, 8)
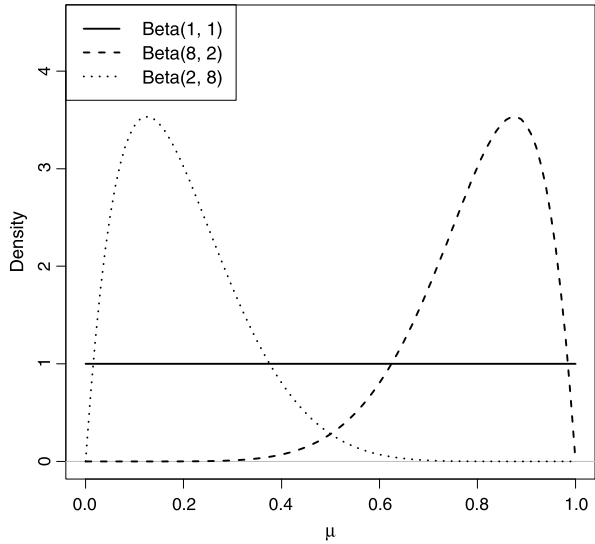
and Shape 2 to 8 and 2, respectively. Make sure the option Plot density function is checked and press OK.

Figure 13.1 shows the probability density curves for three beta distributions. The dashed curve represents the probability density function for Beta(8, 2). Notice that the density for this distribution is large around 0.8. In fact, 0.8 is the mean of this distribution.

> In general, for a beta distribution with parameters α and β, the mean is $\alpha/(\alpha + \beta)$.

For example, the mean of the Beta(2, 8) distribution (shown with a dotted curve in Fig. 13.1) is $2/(2 + 8) = 0.2$.

Now let us reconsider the breast cancer survival example. This time, instead of assuming that only two values are possible, we assume that the true population proportion could be any value from 0 to 1. In general, we always recommend to avoid making overly restrictive assumptions such as the one we used for illustrative purposes in earlier part of this chapter. That is, even if previous studies estimated the population proportion to be either 0.75 and 0.85, we still should consider all other feasible values. We could of course use the results from previous studies and assume that while the survival rate could be any value from 0 to 1, it is more likely to be around 0.8 (e.g., between 0.75 to 0.85) than, for example, around 0.2 (e.g., between 0.15 and 0.25). When specifying the prior distribution, we can use a beta distribution that reflects this assumption.

For the Beta(8, 2) distribution (shown as a dashed curve in Fig. 13.1), the probability (i.e., the area under the density curve) is high for values around 0.8, whereas

the probability is almost zero for values around 0.2. Therefore, we use Beta(8, 2) as the prior distribution for the survival rate of breast cancer patients.

Note that this prior probability distribution reflects our knowledge (based on previous studies) regarding the possible values of survival rate before we obtain new data. We update our knowledge after we observe new empirical evidence. Our updated knowledge is expressed as the posterior probability distribution, which could be drastically different from the prior probability distribution. Therefore, even though we believe in prior that the survival rate is around 0.8, a new empirical evidence could overwhelmingly change this belief. We might be even convinced that values around 0.2 are more probable than values around 0.8 if the observed data strongly suggest that. (We will illustrate this later.)

To find the posterior probability distribution, we use Bayes' theorem as before. When the prior probability distribution is continuous, as it is the case here, finding the posterior probability distribution tends to be complicated in general. For the population proportion, however, using beta prior distributions simplifies the problem of finding the posterior probability. In this case, it turns out that the posterior probability itself is a beta distribution with updated parameters.

> If we assume that the prior knowledge of the population proportion μ, can be expressed using a Beta(α, β) distribution, then the posterior distribution of μ is Beta($\alpha + y, \beta + n - y$), where n is the sample size, and y is the number of times the event of interest has been observed.

In our example, we obtained a sample of 20 patients from the population and found that 18 of them survived after 5 years. Assuming that the prior probability distribution for the breast cancer survival rate is Beta(8, 2), the posterior probability distribution for the survival rate is Beta(8 + 18, 2 + 20 − 18). We can use R-Commander to plot the probability density function for this distribution by following the steps described earlier, but this time we set `Shape 1` and `Shape 2` to 26 and 4, respectively.

Figure 13.2 shows the density curve for the posterior probability distribution, Beta(26, 4). We have included the prior probability density (dashed curve) for comparison. For this example, Beta(26, 4) reflects our updated knowledge about the survival rate among cancer patients. In the Bayesian framework, statistical inference is mainly performed based on the posterior probability distribution. This is the topic of the next section.

Before we discuss how we use posterior probability distributions for statistical inference, we use an illustrative example to show a possible (but not very common) scenario where the observed data are in extreme disagreement with our prior knowledge. For the above example, the prior probability distribution we chose was based on previous studies indicating that the survival rate is around 0.8. Now suppose that out of 20 patients we randomly sampled from the population only 4 have survived. That is, the sample proportion of survival is 0.2, which is far from 0.8. The posterior probability distribution in this case becomes Beta(8 + 4, 2 + 20 − 4) = Beta(12, 18).

Fig. 13.2 The prior
probability distribution
(*dashed curve*) for breast
cancer survival rate and the
resulting posterior probability
distribution (*solid curve*) after
observing 18 survivals among
20 patients
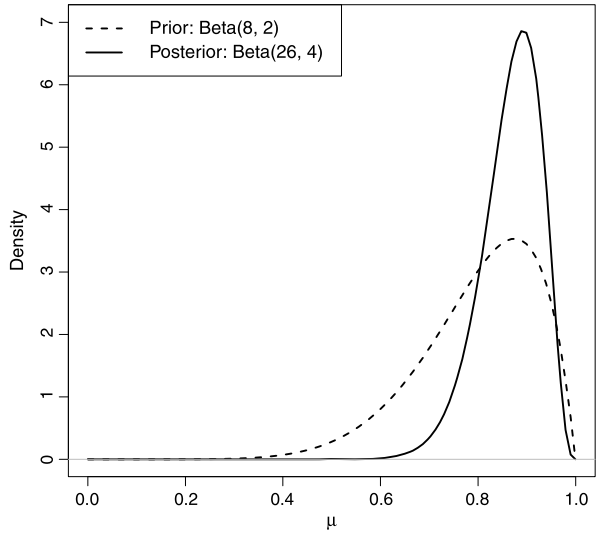
The probability density curve for this distribution is shown in Fig. 13.3. As we
see, the posterior probability distribution, which reflects our updated knowledge, is
substantially different from our prior knowledge. In this case, the distribution has
shifted toward 0.2, which is the sample proportion of survivals based on our new
data.

Posterior probability distributions combine the new observed evidence with the
results of previous studies. In the above example, the posterior probability distri-
bution Beta(12, 18) reflects a compromise between what we knew before collecting
data and what we learned from the new observed data. If the actual survival rate is in
fact 0.2, the posterior distribution would shift even further toward 0.2 as we obtain
more data by increasing the sample size, until the prior becomes completely over-
whelmed by the new empirical evidence. Until then, however, we cannot completely
ignore the knowledge we have accumulated through previous studies.

In practice, it is possible that we might not have strong prior knowledge about
the population proportion. For example, we might be the first group studying the
survival rate of breast cancer patients. We can still use a prior probability distribution
that reflects our lack of knowledge and ignorance. For the population proportion,
Beta(1, 1) is a typical distribution used in such situations. This distribution is shown
as the solid horizontal line from zero to one in Fig. 13.1. This is also known as the
Uniform(0, 1) distribution (i.e., uniform from 0 to 1). Informally, this distribution
states that all values from 0 to 1 are equally probable. More formally, the uniform
distribution indicates that the probability of any interval is equal to the length of the
interval. (Since the height of the pdf for this distribution is equal to one, the area
under the line for each interval is the same as the length of that interval.) Therefore,
the probability that the population proportion is between 0.1 and 0.2 is the same as
the probability that the population proportion is between 0.8 and 0.9. In other words,
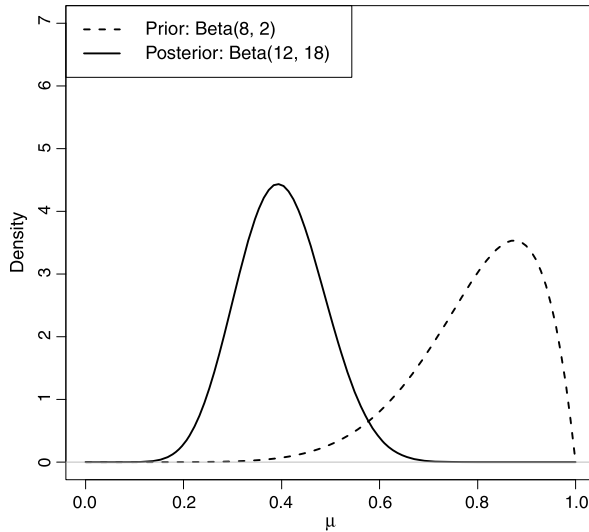
Fig. 13.3 An illustrative example, where the observed data are in extreme disagreement with our prior knowledge about the survival rate of breast cancer patients. The prior probability distribution (*dashed curve*) assigns high probabilities to the range of values around 0.8. The new data however show that only 20% of patients have survived in our sample. The resulting posterior probability distribution (*solid curve*) reflects our updated knowledge, which is substantially different from our prior knowledge

while the first interval includes small values, and the second interval includes large values, both intervals are equally probable since they have the same length.

If we decide to choose Beta(1, 1) for the population proportion of breast cancer survival and if we observe 18 survivals among 20 patients, the posterior probability distribution becomes Beta($1 + 18, 1 + 20 − 18$) = Beta(19, 3).

13.5 Bayesian Inference

In Bayesian analysis, our inference about unknown parameters (e.g., the population proportion) is based on the posterior probability distribution. Here, we discuss how we can use the posterior distribution to obtain point estimates and interval estimates. Also, we discuss some possible methods for performing hypothesis testing. As before, we focus on statistical inference regarding the population proportion.

13.5.1 Estimation

In the previous section, we used Beta(8, 2) as the prior probability distribution for the population proportion of breast cancer survival. After observing 18 survivals among 20 patients, we obtained the posterior probability distribution, Beta(26, 4).

We typically use the mean of the posterior distribution, which is known as the **posterior expectation**, as our point estimate for unknown population parameters.

As mentioned above, the mean of Beta(α, β) distribution is $\alpha/(\alpha + \beta)$. In our example, $\alpha = 26$ and $\beta = 4$ for the posterior distribution. Therefore, our point estimate for the survival rate is $26/(26 + 4) = 0.87$. This posterior expectation is close to the sample proportion $18/20 = 0.9$, and it approaches the sample proportion as we increase the sample size.

While working with point estimates is convenient, they do not reflect the uncertainty regarding our estimates. As discussed in earlier chapters, interval estimates are recommended instead. Finding interval estimates based on posterior probability distributions is quite simple. For example, suppose that we want to find an interval that includes the true value of the population proportion μ with probability 0.95. We show this interval as $[L, U]$, where L is the lower limit, and U is the upper limit of the interval. The easiest way to find such an interval is to use the 0.025-quantile as the lower limit of the interval and the 0.975-quantile as the upper limit of the interval from the posterior probability distribution. Recall that the 0.025-quantile is the value whose lower tail probability is 0.025, and the 0.975-quantile is the value whose lower tail probability is 0.975. Therefore, the probability between these two values is $0.975 - 0.025 = 0.95$.

We can use R-Commander to find these quantiles for the Beta(26, 4). Click Distributions → Continuous distributions → Beta distribution → Beta quantiles. Set the probabilities to 0.025 0.975 (use white space to separate the two probabilities; alternatively, you can find the quantiles one at a time) and set Shape 1 and Shape 2 to 26 and 4, respectively. The 0.025 and 0.975 quantiles, which are 0.73 and 0.96, respectively, will be printed in the *Output* window. Therefore, μ (population proportion of survival among breast cancer patients) falls within [0.73, 0.96] interval with probability 0.95. We call this interval the 95% probability interval or **credible interval** for μ. We can follow similar steps to find other credible intervals. For example, we can set L to the 0.05-quantile and U to the 0.95-quantile to obtain the 90% credible interval.

Note that a *credible interval* is obtained directly from a probability distribution, and we can interpret it as the range of possible values that include the true value of the unknown parameter with specified probability (e.g., 0.95). In contrast, we do not use the term "probability" in our interpretation of a *confidence interval*, unless we are referring to the procedure used to create that specific interval.

13.5.2 Hypothesis Testing

Formally, hypothesis testing in the Bayesian frameworks is regarded as a decision problem. Here, however, we discuss some simple methods for conducting hypothesis testing analogous to those methods we discussed in earlier chapters.

For the breast cancer survival example, suppose that we hypothesize that the population proportion is above 0.8; that is, $H_A : \mu > 0.8$. We can use the posterior probability distribution to find the probability of H_A. In this case, H_A is true when the population proportion is above 0.8. Therefore, to find the probability that H_A is true, we can find the upper tail probability of 0.8 from Beta(26, 4). In R-Commander, click Distributions → Continuous distributions → Beta distribution → Beta probabilities. Then, set Variable value(s) to 0.8, Shape 1 to 26, and Shape 2 to 4. Make sure the option upper tail is checked. The upper tail probability of 0.8 is then 0.86. Now, we need to decide whether this probability is large enough so that we can accept H_A. This of course depends on the specific problem at hand. In general, our decision might depend on other factors (e.g., loss function) not discussed here.

Note that in Bayesian statistics, we directly evaluate the hypothesis that has inspired our study by finding its probability. This is typically the alternative hypothesis. Of course, if we want, we can find the probability of the null hypothesis. For the above example, if we specify the null hypothesis as $H_0 : \mu \leq 0.8$, we can calculate its probability by finding the lower tail probability of 0.8 from the posterior probability distribution. In this case, the probability that the null hypothesis is true is 0.14. Consequently, we can conclude that the posterior odds of H_A compared to H_0 is $0.84/0.14 = 6.1$. That is, the alternative hypothesis is about 6 times more likely to be true compared to the null hypothesis given the observed data.

Now suppose that we want to perform a two-sided hypothesis test, where $H_A : \mu \neq 0.8$. This hypothesis states that the population proportion is different from 0.8. Of course, the probability that μ is exactly 0.8 is zero (since the probability distribution of μ is continuous), which means that the hypothesis is true with probability 1. Therefore, we need to be clear about what we mean by saying that the population proportion is different from 0.8. That is, how far away we should move from 0.8 until we consider μ different from 0.8 for all practical purposes. For the specific hypothesis testing problem we are working on, we might, for example, consider a difference of 0.02 large enough so that we consider values below 0.78 and above 0.82 different from 0.8. This corresponds to a difference we consider important in a practical sense. Recall that for significance testing methods we discussed earlier, we emphasized the importance of distinguishing between practical significance and statistical significance. In Bayesian statistics, we do not need to make such distinction; we can specify and evaluate a hypothesis based on what we consider significant in practice.

For the above example, the alternative hypothesis states that the difference between the true value of the population proportion and 0.8 is at least 0.02. Now, we can use the posterior probability distribution, Beta(26, 4), to examine this hypothesis. To this end, we can find the probability of [0.78, 0.82] interval and subtract the results from 1 to find the probability of its complement, which in this case is the probability of values outside the interval. Using R-Commander, we find that the probability of [0.78, 0.82] is 0.12 using the Beta(26, 4) distribution. (We subtract the lower tail probability of 0.78 from the lower tail probability of 0.82.) As a result, the probability of the alternative hypothesis (i.e., the difference between μ and 0.8 is at least 0.02) is $1 - 0.12 = 0.88$.

13.6 Advanced

As we saw in this chapter, the beta distribution plays an important role in Bayesian analysis of binary data. The prior for the population proportion μ is specified as a beta distribution; the posterior distribution is also a beta distribution with different parameters.

When choosing a beta distribution to express our knowledge regarding the likely values of the population proportion, we should plot the probability density function to make sure it properly reflects our knowledge. Suppose that we have decided to use Beta(8, 2) as our prior for μ, where μ is the survival rate among breast cancer patients. An easy approach to plot the probability density function is to find the values of the density function using `dbeta()` for a set of μ values:

```
> mu <- seq(from = 0, to = 1, length.out = 100)
> f <- dbeta(mu, shape1 = 8, shape = 2)
```

Note that μ only takes values between 0 and 1. The `seq` function returns a sequence of numbers from 0 to 1 and assigns them as a vector to the object `mu`. The `length.out = 100` option in `seq` specifies the length of the sequence; in this case, there are 100 values in the sequence. For each value, the function `dbeta()` returns the value of the density function. Here, `shape1` and `shape2` specify the parameter values for the beta distribution.

We can now plot the probability density function:

```
> plot(mu, f, type = "l", xlab = expression(mu),
+      ylab = "Density")
```

We have set the option `type` to "l", so the points are connected by lines. The points themselves are not shown. If we want to show both the points and the lines that connect them, we use the option `type='b'` (for **b**oth). For the label of the x-axis, we have used `expression(mu)` to create a mathematical annotation so that the label reads as μ instead of mu similar to Fig. 13.1. You can use `expression()` for other Greek letters, e.g., `expression(sigma)` adds the annotation σ (instead of sigma). You can also use `expression()` to produce subscript or superscript. For example, `expression(x[2])` generates x_2, and `expression(x^2)` generates x^2.

After we make sure that the prior appropriately represents our knowledge regarding the possible values of μ, we can update the prior based on the observed data to obtain the posterior probability distribution, which is another beta distribution in this case. In the example discussed above, we obtained a sample of 20 patients from the population and found that 18 of them survived after 5 years. Assuming that the prior probability distribution for the breast cancer survival rate is Beta(8, 2), the posterior probability distribution for the survival rate is Beta(8 + 18, 2 + 20 − 18). We can plot the posterior distribution, Beta(26, 4), the same way we plotted the prior distribution.

To find the 95% credible interval based on the posterior probability distribution, we usually use the 0.025- and 0.975-quantiles. For this, we use the `qbeta()` function:

```
> qbeta(c(0.025, 0.975), shape1 = 26, shape2 = 4)

[1] 0.7264848 0.9611052
```

Note that we have given 0.025 and 0.975 as a vector, so the output is a vector too.

To use the posterior probability distribution for hypothesis testing, we need to find the lower and upper probabilities for values specified by the hypothesis. For example, if the alternative hypothesis states that $H_A : \mu > 0.8$, we can find its posterior probability by calculating the upper tail probability for 0.8 based on Beta(26, 4). For this, we use the `pbeta()` function:

```
> pbeta(0.8, shape1 = 26, shape2 = 4,
+     lower.tail = FALSE)

[1] 0.8596195
```

By default, the `pbeta()` function returns the lower tail probability. We can obtain the upper tail probability by setting the argument `lower.tail` to FALSE.

To obtain the probability of a given interval, we subtract their corresponding lower tail probabilities as before. For example, if the null hypothesis, H_0, states that μ is in the interval [0.78, 0.82] (i.e., within 0.02 from 0.8), we can find the posterior probability of H_0 as follows:

```
> p.78 <- pbeta(0.78, shape1 = 26, shape2 = 4,
+     lower.tail = TRUE)
> p.82 <- pbeta(0.82, shape1 = 26, shape2 = 4,
+     lower.tail = TRUE)
> p.82 - p.78

[1] 0.1159820
```

Therefore, the posterior probability of H_0 is 0.12. Consequently, the posterior probability that μ is outside of the interval [0.78, 0.82] is $1 - 0.12 = 0.88$.

13.7 Exercises

1. Suppose that the smoking status, X, has Bernoulli(μ) distribution, where the prior distribution for μ is Beta(1, 10). We have interviewed a random sample of 50 people and found that 8 of them smoke regularly. Find the posterior distribution of μ given the observed data. Plot the posterior distribution and find the

95% credible interval for μ. What is the point estimate for μ based on this distribution? Compare this point estimate to the sample proportion, which is also commonly used as a point estimate for the population proportion. Suppose that we hypothesize that less than 20% of the population smoke. Use the posterior probability distribution to find the probability that our hypothesis is true.

2. For the above example, suppose that we want to conduct another study so that we can obtain more data. Use the beta distribution you found as the posterior probability distribution in the above example as your prior. This reflects what we know about μ before conducting the next study. Now suppose that in our next study, we have interviewed 30 people and found that 6 of them smoke regularly. Use the new data to obtain the posterior probability distribution. What are the mean and the 95% credible interval for μ?

3. For the above examples, suppose that we use Beta(1, 10) as our prior probability distribution for μ, but this time we wait to obtain the posterior probability distribution at the end of the second study. In this case, we have interviewed a total sample of $50 + 30 = 80$ people, where $8 + 6 = 14$ of them are identified as smokers. Compare the posterior probability distribution to what you found by updating your knowledge of μ gradually (i.e., after each study).

4. Suppose that we are interested in the proportion of population affected by diabetes among Pima Indian women. Let us represent the diabetes status of each person by random variable X, where $X = 1$ if the person has diabetes and $X = 0$ if the person does not. Then we can assume that X has a Bernoulli distribution with parameter μ. We know that the population proportion of diabetic women in the whole US is about 10%. We want to use this information to specify our prior for μ. Use R-Commander to find a beta distribution that has relatively high probability density values around 0.1. For this, plot different beta distribution by changing the parameters until you find a distribution for which the area under the probability density curve is large over the interval from 0.05 to 0.15. Then use the `Pima.tr` data (available from the `MASS` package) to find the posterior probability distribution of μ. Use the posterior probability distribution to obtain the point estimate and 95% credible interval for μ.

5. For the above question, repeat your analysis with Beta(1, 1) prior, which reflects our ignorance about the possible values of μ.

6. Suppose that we believe that the population proportion of newborn babies with low birthweight is around 0.3. Use R-Commander to find a beta probability distribution that reflects our knowledge. Use the `birthwt` data (available from the `MASS` package) to update our prior knowledge. Now suppose that someone hypothesizes that the population proportion of low birthweight babies is different from 30%, and she defines a difference of 5% as significant. What is the probability that this hypothesis is true?

Appendix A
Installing R and R-Commander

This appendix gives detailed instructions for installing R and R-Commander.

A.1 Installing R

1. Go to http://www.r-project.org/.
2. Click on the `download R` link.
3. Then select a location closest to you.
4. Click on your operating system (Linux, MacOS X, Windows) and follow directions.

If you are a **Mac** user, download the latest .dmg file and follow the instructions. Once installed, R will appear as an icon in the Applications Folder. After you install R, you should go back to the same webpage (where you obtained the latest .dmg file), click on "tools", which is located under "Subdirectories", and install the universal build of **Tcl/Tk** for **X11**. The file name contains "tcltk", three numbers representing the current version, and "x11.dmg". (Currently, the file name is "tcltk-8.5.5-x11.dmg".) This file includes additional tools necessary for building R for Mac OS X.

If you are running Windows, click on `base` and then on the link that downloads R for Windows. (In the link, the current version number appears after "R".) When the dialog box opens, click `Run`, and a "Setup Wizard" should appear. Keep clicking `Next` until the Wizard is finished. Now, you should see an icon on your desktop, with a large capital `R`.

A.2 Installing R-Commander

A.2.1 From the Command Line

You can download R-Commander from the command line by following these steps:
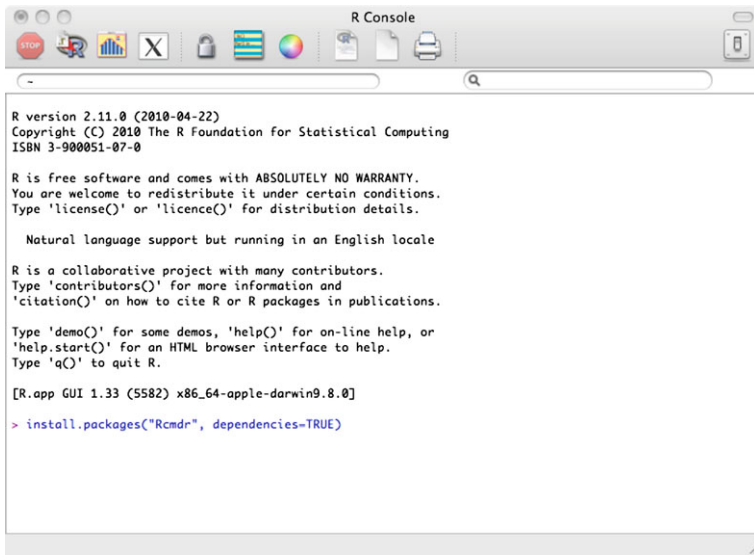
Fig. A.1 Installing R-Commander by entering the command `install.packages` `("Rcmdr", dependencies=TRUE)` in R Console

1. Once you have installed R, open it by double-clicking on the icon.
2. A window called "R Console" will open.
3. Make sure you have a working internet connection. Then, at the prompt (the > symbol), type the following command exactly and then press enter (Fig.):

   ```
   > install.packages("Rcmdr", dependencies = TRUE)
   ```

4. R may respond by asking you to select a mirror site and listing them in a pop-up box. Choose a nearby location.
5. Depending on your connection speed, the installation may take awhile. Be patient and wait until you see the prompt again before you do anything.

A.2.2 From the Menu Bar

Alternatively, you can download R-Commander from the menu bar by following these steps:

1. Open R by clicking on its icon.
2. Click "Packages & Data" from the menu bar and then click "Package Installer". This opens a window similar to Fig. .
3. Click "Get List".
4. Scroll down and select "Rcmdr".
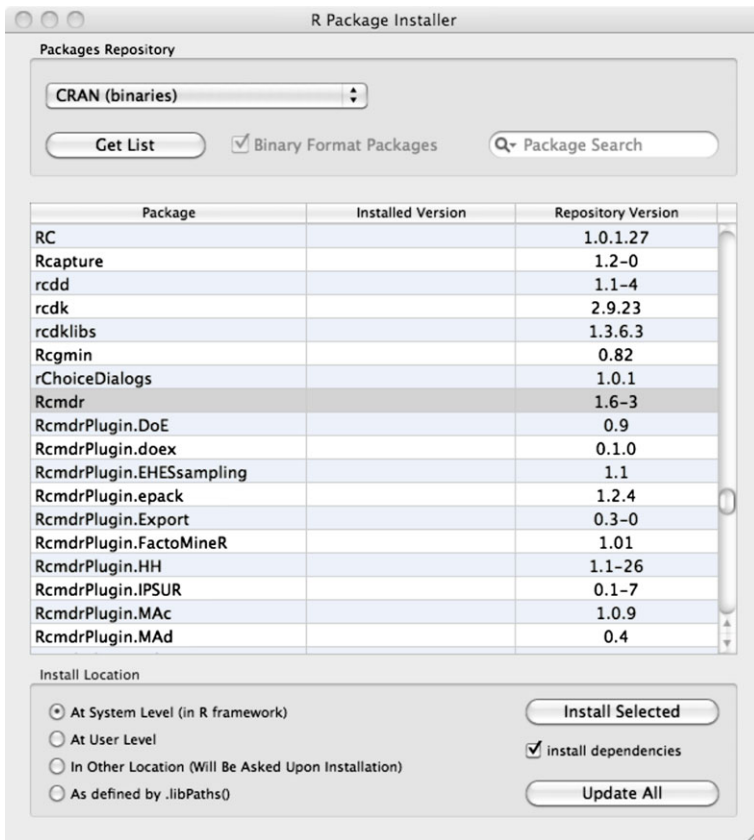
Fig. A.2 Installing R-Commander using "Package Installer" from the menu bar

5. Check "Install dependencies" and click "Install Selected".
6. R may respond by asking you to select a mirror site and listing them in a pop-up box. Choose a nearby location.
7. The installation may take awhile. Wait until you see the prompt again before you do anything.

A.3 Starting R-Commander

If R is not already open, open it by clicking on its icon. To open R-Commander, at the prompt enter the following command (Fig. A.3):

```
> library(Rcmdr)
```

Fig. A.3 Opening R-Commander by entering the command `library(Rcmdr)` in R Console
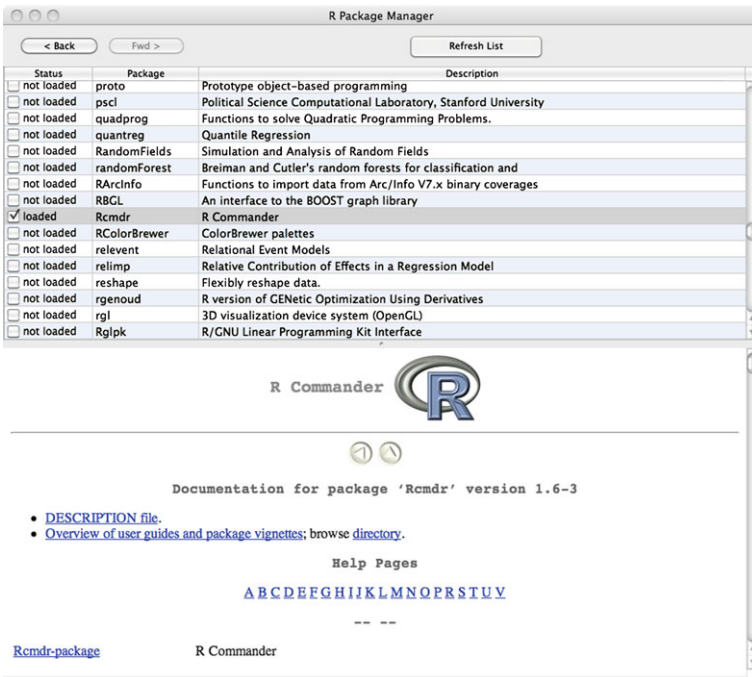
Fig. A.4 The "Package Manager" window
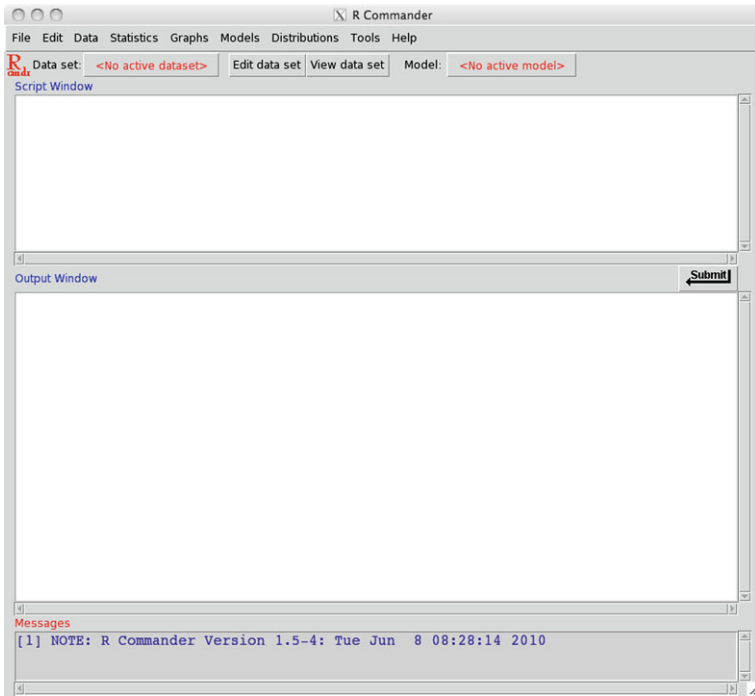
Fig. A.5   The R-Commander window

Alternatively, you can load R-Commander by clicking "Packages & Data" from the menu bar and then clicking "Package Manager". This will open a window similar to Fig. A.4. Scroll down to find R-Commander ("Rcmdr" package) and check its box to change the status from "not loaded" to "loaded".

You should see a large new window pop-up, labeled R-Commander (Fig. A.5). You are now ready to analyze your data. If you close this window while R is still open, you can start R-Commander again by entering the command Commander() in R Console. (Entering library(Rcmdr) in this situation will not work unless you close R and open it again.) Alternatively, you can use the "Package Manager" window to unload and reload R-Commander.

If you are a Mac user and your OS version is 10.4.11 or lower, you might need to open X11 manually before entering the command library(Rcmdr) in R Console. To do this, under Misc from the top bar menu choose Run X11 Server. If you do not have X11, you can install it from your install DVD or find it online.

Appendix B
Basic R

This appendix gives a brief introduction to R language for those who want to use R (instead of R-Commander) for data analysis. The introduction given here is meant to help with the initial steps toward this goal. To use R more broadly and more effectively, one needs to refer to more advanced references for R programming as well as reading the advanced sections in this book.

B.1 Starting with R

In the R Console, you can type commands directly at the prompt, >, and execute them by pressing enter. Commands can also be entered in the *Script* window in R-Commander and executed by pressing the "Submit" button. Both the R Console and R-Commander provide an interactive environment where the results are immediately shown similar to a calculator. In fact, R can be used as a calculator. The basic arithmetic operators are + for addition, – for subtraction, * for multiplication, and / for division. The ^ operator is used to raise a number (or a variable) to a power. Try executing the following commands:

```
> 65 + 32

[1] 97

> 3 * 1.7 - 2

[1] 3.1

> 4 * 3 + 6/0.2

[1] 42

> 5^2

[1] 25
```

There are also built in functions for finding the square root sqrt(), the exponential exp(), and the natural logarithm log():

```
> sqrt(430)

[1] 20.73644

> exp(-1.3)

[1] 0.2725318

> log(25)

[1] 3.218876
```

Here, the numbers in the parentheses serve as input (parameters or arguments) to the functions. Most functions have multiple parameters and options. For example, to take the base-10 logarithm of 25, we include the option base=10:

```
> log(25, base = 10)

[1] 1.39794
```

In general, you can always learn more about a function and its options (arguments) by writing its name after a question mark (e.g., ?log).

We can combine two or more functions such that the output from one function becomes the input for another function. For example, the following code combines the above log() function with the round() function, which is used to specify the number of decimal places (here, two digits):

```
> round(log(25, base = 10), digits = 2)

[1] 1.4
```

In the above code, the output of log(25, base=10) becomes the first argument of the function round(). The second argument, digits=2, specifies the number of decimal places.

B.2 Creating Objects in R

Instead of directly entering commands such as 2+3, we can create *objects* to hold values and then perform operations on these objects. For example, the following set of commands creates two objects x and y, adds the values stored in these objects, and assigns the result to the third object z:

```
> x <- 2
> y <- 3
> z <- x + y
```

In general, we use left arrow `<-` (i.e., type `<` and then `-`) to assign values to an object. Almost always, we can use the equal sign "=" instead of `<-` for assignment. For example, we could use `y = 3` to assign the value 3 to *y*, and use `z = x+y` to set *z* equal to the sum of *x* and *y*. The two options, "=" and `<-`, have some small differences, which are not discussed here.

Simply typing the name of an object displays its contents. We could also use the function `print()`:

```
> x
```

```
[1] 2
```

```
> print(y)
```

```
[1] 3
```

```
> print(z)
```

```
[1] 5
```

Object names are case sensitive. For example, `x` and `X` are two different objects. A name cannot start with a digit or an underscore `_` and cannot be among the list of reserved words such as `if, function, NULL`. We can use the period `.` in a name to separate words (e.g., `my.object`).

The objects are created and stored by their names. We can obtain the list of objects that are currently stored within R by using the function `objects()`:

```
> objects()
```

```
[1] "x" "y" "z"
```

The collection of these objects is called the *workspace*. (Note that although we are not specifying the values of arguments, we still need to type parentheses when using functions.) When closing an R session, you have the opportunity to save the objects permanently for future use. This way, the workspace is saved in a file called `.RData`, which will be restored next time you open R. If you want to save only few objects, as opposed to the entire workspace, you can use the function `save()`. For example, suppose that we only want to save the objects `x` and `y`:

```
> save(x, y, file = "myObjects.RData")
```

The above command saves `x` and `y` in a file called `myObjects.RData` in the *current working directory*. You can see where the current working directory is by

entering the command `getwd()`. If you want to save your objects in another directory, either enter the full path when specifying the file name in the `save()` function or use the menu bar to change the directory. (For Mac, this option is located under "Misc". For Windows, it is located under "File".)

After you save the `x` and `y` in `myObjects.RData`, you can load these objects for future use with the `load()` function:

```
> load("myObjects.RData")
```

Give the full address for the file if it is not located in the current working directory. Alternatively, you can change the working directory from the menu bar.

B.3 Vectors

Using objects allows for more flexibility. For example, we can store more than one value in an object and apply a function or an operation to its contents. The following commands create a *vector* object `x` that contains numbers 1 through 5 and then apply two different functions to it:

```
> x <- c(1, 2, 3, 4, 5)
> x

[1] 1 2 3 4 5

> 2 * x + 1

[1]   3   5   7   9 11

> exp(x)

[1]    2.718282    7.389056   20.085537   54.598150
[5] 148.413159
```

The `c()` function is used to combine its arguments (here, integers from 1 to 5) into a vector. Since $1, 2, \ldots, 5$ is a sequence of consecutive numbers, we could simply use the colon ":" operator to create the vector:

```
> x <- 1:5
> x

[1] 1 2 3 4 5
```

To create sequences and store them in vector objects, we can also use the `seq()` function for additional flexibility. The following commands create a vector object `y` containing a sequence increasing by 2 from -3 to 14:

```
> y <- seq(from = -3, to = 14, by = 2)
> y

[1] -3 -1  1  3  5  7  9 11 13
```

If the elements of a vector are all the same, we can use the `rep()` function:

```
> z <- rep(5, times = 10)
> z

 [1] 5 5 5 5 5 5 5 5 5 5
```

The following function creates a vector of size 10 where all its elements are un-known. In R, missing values are represented by NA (Not Available):

```
> z <- rep(NA, times = 10)
> z

[1] NA NA NA NA NA NA NA NA NA NA
```

This way, we can create a vector object of a given length and specify its elements later.

To find the length of a vector (i.e., number of elements), we use the `length()` command:

```
> length(x)

[1] 5

> length(y)

[1] 9
```

Functions `sum()`, `mean()`, `min()`, and `max()` return the sum, average, min-imum, and maximum for a vector:

```
> x

[1] 1 2 3 4 5

> sum(x)

[1] 15

> mean(x)

[1] 3
```

```
> min(x)
```

```
[1] 1
```

```
> max(x)
```

```
[1] 5
```

The elements of a vector can be accessed by providing their index using square brackets [ ]. For example, try retrieving the first element of x and the 4th element of y:

```
> x[1]
```

```
[1] 1
```

```
> y[4]
```

```
[1] 3
```

The colon : operator can be used to obtain a sequence of elements. For instance, elements 3 through 6 of y can be accessed with

```
> y[3:6]
```

```
[1] 1 3 5 7
```

To select all but the 4th element of a vector, we use negative indexing:

```
> y[-4]
```

```
[1] -3 -1  1  5  7  9 11 13
```

The above objects are all *numerical* vectors. A vector can also hold character strings delimited by single or double quotations marks. For example, suppose that have a sample of 5 patients. We can create a *character* vector storing their gender as

```
> gender <- c("male", "female", "female", "male",
+     "female")
```

(The plus sign means that the second line is the continuation of the command in the first line. You should not type it when trying the above command in R.) Retrieving the elements of the vector is as before:

```
> gender[3]
```

```
[1] "female"
```

A vector could also be *logical*, where the elements are either TRUE or FALSE (NA if the element is missing). Note that these values must be in capital letters and can be abbreviated by T and F (not recommended). For example, create a vector storing the health status of the five patients:

```
> is.healthy <- c(TRUE, TRUE, FALSE, TRUE, FALSE)
> is.healthy

[1]   TRUE  TRUE FALSE  TRUE FALSE
```

When used in ordinary arithmetic, logical vectors are coerced to integer vectors, 0 for FALSE elements and 1 for TRUE elements. For example, applying the sum() function to the above logical vector turns the TRUE and FALSE values to ones and zeros, and returns their sum, which in this case is equivalent to the number of healthy subjects:

```
> sum(is.healthy)

[1] 3
```

Use the as.integer() function to see the equivalent integer vector for is.healthy:

```
> as.integer(is.healthy)

[1] 1 1 0 1 0
```

Logical vectors are usually derived from other vectors using logical operators. For example, with the gender vector, we can create a logical vector showing which subjects are male:

```
> gender

[1] "male"   "female" "female" "male"   "female"

> is.male <- (gender == "male")
> is.male

[1]   TRUE FALSE FALSE  TRUE FALSE
```

Here, == (i.e., two equal signs) is a relational operator that returns TRUE if the two sides are equal and returns FALSE otherwise. As the second example, we create a numerical vector for the age of the five subjects and then check to see which person is 60 years old:

```
> age <- c(60, 43, 72, 35, 47)
> is.60 <- (age == 60)
```

```
> is.60
```

```
[1]  TRUE FALSE FALSE FALSE FALSE
```

The != operator, on the other hand, returns a TRUE value when the two sides are not equal:

```
> is.female <- (gender != "male")
> is.female
```

```
[1] FALSE  TRUE  TRUE FALSE  TRUE
```

```
> not.60 <- (age != 60)
> not.60
```

```
[1] FALSE  TRUE  TRUE  TRUE  TRUE
```

The other relational operators commonly applied to numerical vectors are "less than", <, "less than or equal to", <=, "greater than", >, "greater than or equal to": >=

```
> age < 43
```

```
[1] FALSE FALSE FALSE  TRUE FALSE
```

```
> age <= 43
```

```
[1] FALSE  TRUE FALSE  TRUE FALSE
```

```
> age > 43
```

```
[1]  TRUE FALSE  TRUE FALSE  TRUE
```

```
> age >= 43
```

```
[1]  TRUE  TRUE  TRUE FALSE  TRUE
```

We can also use *Boolean* operators to create new logical vectors based on existing ones. The logical NOT, !, negates the elements of a logical vector (i.e., changes TRUE to FALSE and vice versa). For example, create a is.female vector from the is.male vector:

```
> is.female <- !is.male
> is.female
```

```
[1] FALSE  TRUE  TRUE FALSE  TRUE
```

The logical AND, &, compares the elements of two logical vectors and returns TRUE only when the corresponding elements are both TRUE:

```
> is.male

[1]   TRUE FALSE FALSE   TRUE FALSE

> is.healthy

[1]   TRUE   TRUE FALSE   RUE FALSE

> is.male & is.healthy

[1]   TRUE FALSE FALSE   TRUE FALSE
```

The logical OR, |, also compares the elements of two logical vectors and returns TRUE when at least one of the corresponding elements is TRUE:

```
> is.male | is.healthy

[1]   TRUE   TRUE FALSE   TRUE FALSE
```

We can use combinations of two or more logical operators. The following commands check to see which subjects are male and less than 45 years old:

```
> is.young.male <- is.male & (age < 45)
> is.young.male

[1] FALSE FALSE FALSE   TRUE FALSE
```

Using the which() function, we can obtain the indices of TRUE elements for a given logical function:

```
> ind.male <- which(is.male)
> ind.male

[1] 1 4

> ind.young <- which(age < 45)
> ind.young

[1] 2 4
```

We can then use these indices to obtain their corresponding elements from a vector:

```
> age[ind.male]

[1] 60 35

> is.male[ind.young]

[1] FALSE   TRUE
```

We can combine the two steps:

```
> age[is.male]

[1] 60 35

> age[gender == "male"]

[1] 60 35

> is.male[age < 45]

[1] FALSE   TRUE

> gender[age < 45]

[1] "female" "male"
```

B.4 Matrices

In the above examples, we used one vector for each characteristic (e.g., age, health status, gender) of our five subjects. It is easier, of course, to store the subject information in a table format, where each row corresponds to an individual and each column to a characteristic. If all these measurements are of the same type (e.g., numerical, character, logical), a *matrix* can be used. For example, besides age, assume that for our five subjects, we have also measured BMI (body mass index) and blood pressure:

```
> BMI = c(28, 32, 21, 27, 35)
> bp = c(124, 145, 127, 133, 140)
```

Now create a matrix with the cbind() function for column-wise binding of age, BMI, and bp:

```
> data.1 = cbind(age, BMI, bp)
> data.1

     age BMI  bp
[1,]  60  28 124
[2,]  43  32 145
[3,]  72  21 127
[4,]  35  27 133
[5,]  47  35 140
```

If we had wanted a matrix where each row represented a characteristic and each column a subject, we would have used the rbind() function for row-wise binding:

```
> data.2 = rbind(age, BMI, bp)
> data.2

    [,1] [,2] [,3] [,4] [,5]
age   60   43   72   35   47
BMI   28   32   21   27   35
bp   124  145  127  133  140
```

We could obtain `data.2` by transposing (e.g., interchanging the rows and columns) `data.1` using the `t()` function:

```
> t(data.1)

    [,1] [,2] [,3] [,4] [,5]
age   60   43   72   35   47
BMI   28   32   21   27   35
bp   124  145  127  133  140
```

In general, matrices are two-dimensional objects comprised of values of the same type. The object `data.1` is a 5×3 matrix. The function `dim` returns the size (i.e., the number of rows and columns) of a matrix:

```
> dim(data.1)

[1] 5 3
```

When creating the matrix `data.1`, R automatically uses the vector names as the column names. They can be changed or accessed with the function `colnames()`:

```
> colnames(data.1)

[1] "age" "BMI" "bp"
```

Likewise, we can obtain or provide the row names using the function `row-names()`:

```
> rownames(data.1) <- c("subject1", "subject2",
+    "subject3", "subject4", "subject5")
> data.1

         age BMI  bp
subject1  60  28 124
subject2  43  32 145
subject3  72  21 127
subject4  35  27 133
subject5  47  35 140
```

To access the elements of a matrix, we still use square brackets [ ], but this time, we have to provide both the row index and the column index. For instance, the age of the third subject is

```
> data.1[3, 1]
```

```
[1] 72
```

If only a row number is provided, R returns all elements of that row (e.g., all the measurements for one subject):

```
> data.1[2, ]
```

```
age BMI  bp
 43  32 145
```

Likewise, if only a column is specified, R returns all elements of that column (e.g., all the measurements for one characteristic):

```
> data.1[, 2]
```

```
subject1 subject2 subject3 subject4 subject5
      28       32       21       27       35
```

A matrix can also be created by rearranging the elements of a vector with the `matrix` function:

```
> matrix(data = 1:12, nrow = 3, ncol = 4)
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

Here, the argument `data` specifies the numbers (vector), whose rearrangement creates the matrix. The arguments `nrow` and `ncol` are the number of rows and columns, respectively. By default, the matrix is filled by columns. To fill the matrix by rows, we must use the argument `byrow=TRUE`:

```
> matrix(data = 1:12, nrow = 3, ncol = 4, byrow = TRUE)
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
```

The length of `data` is usually equal to *nrow* × *ncol*. If there are too few elements in `data` to fill the matrix, then the elements are recycled. In the following example, all elements of the matrix are set to zero:

```
> mat <- matrix(data = 0, nrow = 3, ncol = 3)
> mat
```

```
     [,1] [,2] [,3]
[1,]    0    0    0
[2,]    0    0    0
[3,]    0    0    0
```

We can obtain or set the diagonal elements of a matrix using the `diag()` function:

```
> diag(mat) <- 1
> mat
```

```
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

Similar to vectors, we can create a matrix with missing values (NA) and specify the elements later:

```
> mat <- matrix(data = NA, nrow = 2, ncol = 3)
> mat
```

```
     [,1] [,2] [,3]
[1,]   NA   NA   NA
[2,]   NA   NA   NA
> mat[1, 3] <- 5
> mat
```

```
     [,1] [,2] [,3]
[1,]   NA   NA    5
[2,]   NA   NA   NA
```

B.5 Data Frames

In general, we use matrices when all measurements are of the same type (e.g., numerical, character, logical). Otherwise, the type of measurements could change when we mix different types. In the following example, we show this by using the `mode()` function that returns the type for a given object:

```
> mat <- cbind(age, gender, is.healthy)
> mode(mat)
```

```
[1] "character"
```

In the above example, the matrix type is character, although `age` is numeric and
`is.healthy` is logical. Note that their values will be printed in quotation marks,
which is used for character strings, if you enter `mat`.

To avoid this issue, we can store data with information of different types in a table
format similar to the format of spreadsheets. The resulting object (which includes
multiple objects of possibly different types) is called a *data frame* object. For this,
we use the function `data.frame()`:

```
> data.df = data.frame(age, gender, is.healthy,
+     BMI, bp)
> data.df

  age gender is.healthy BMI  bp
1  60   male       TRUE  28 124
2  43 female       TRUE  32 145
3  72 female      FALSE  21 127
4  35   male       TRUE  27 133
5  47 female      FALSE  35 140
```

To create a data frame this way, all vectors must have the same length. You may
notice that while `gender` is a character vector, its elements are not printed with
quotation marks as before. The reason is that character vectors included in data
frames are coerced to a different type called *factors*. A factor is a vector object
that is used to provide a compact way to represent categorical data. Each *level* of
a factor vector represents a unique category (e.g., female or male). We can directly
create factors using the `factor()` function:

```
> factor(gender)

[1] male   female female male   female
Levels: female male
```

To access elements of a data frame, we use the square brackets `[ , ]` with the
appropriate row and column indices. For example, the BMI of the 3rd subject is

```
> data.df[3, 4]

[1] 21
```

As before, we can access an entire row (e.g., all the measurements for one subject)
by only specifying the row index and an entire column (e.g., all the measurements
for one variable) by only specifying the column index. We can also access an entire
column by providing the column name:

```
> data.df[, "age"]

[1] 60 43 72 35 47
```

The $ operator also retrieves an entire column from the data frame:

```
> data.df$age
```

```
[1] 60 43 72 35 47
```

This column can then be treated as a vector and its elements accessed with the square brackets as before. For instance, try obtaining the BMI for the 3rd subject and the gender of the 2nd subject:

```
> data.df$BMI[4]
```

```
[1] 27
```

```
> data.df$gender[2]
```

```
[1] female
Levels: female male
```

B.5.1 Creating Data Frames Using a Spreadsheet-Like Environment

We can create a data frame by invoking a spreadsheet-like environment in R. For this, we start by creating an empty data frame object:

```
> new.df <- data.frame()
```

Then, we use the function fix() to edit the newly created data frame new.df.

```
> fix(new.df)
```

This way, R opens a window for data editing similar to Fig. B.1. You can use the four icons at the top of the editor to add or delete columns or rows. (Hover your mouse pointer over each icon to see its function.) For new.df, we have created four columns and three rows. The column names and the content of the data frame can be edited the same way we edit spreadsheets.

B.5.2 Importing Data from Text Files

Data are usually available in a tabular format as a delimited text file. We can import the contents of such files into R using the function read.table(). For instance, let us try importing the BodyTemperature data set (this data set is available online from the book website http://extras.springer.com):
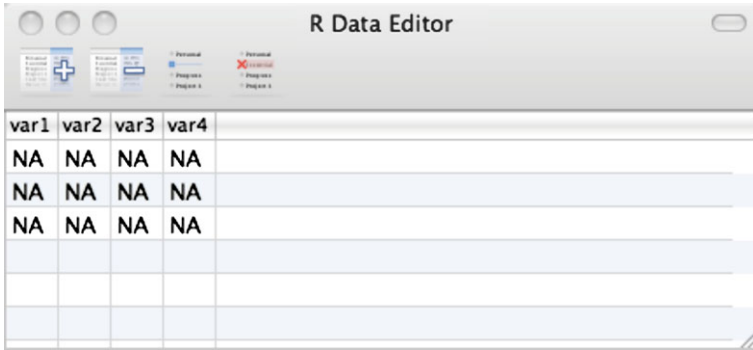
Fig. B.1 R window for data editing invoked by applying the function `fix()` to an empty data frame object

```
> BodyTemperature <- read.table(
+       file = "BodyTemperature.txt",
+       header = TRUE, sep = " ")
```

Here, we are using the `read.table()` function with three arguments. The first argument, `file="BodyTemperature.txt"`, specifies the name and location of the data file. If the file is not in the current working directory, you need to give the full path to the file or change the working directory. The `header=TRUE` option tells R that the variable names are contained in the first line of the data. Set this option to `FALSE` when this is not the case. The `sep` option tells R how the columns are separated in the text file. In this example, the columns are separated by white spaces. If the columns were separated by commas, for example, we would have used `sep=","`.

The `BodyTemperature` object is a data frame holding the contents of the "BodyTemperature.txt" file. Type `BodyTemperature` to view the entire data set. If the data set is large, it is better to use the `head()` function, which shows only the first part (few rows) of the data set:

```
> head(BodyTemperature)

  Gender Age HeartRate Temperature
1      M  33        69        97.0
2      M  32        72        98.8
3      M  42        68        96.2
4      F  33        75        97.8
5      F  26        68        98.8
6      M  37        79       101.3
```

In the `BodyTemperature` data frame, the rows correspond to subjects and the columns to variables. To view the names of the columns, try

```
> names(BodyTemperature)
```

```
[1] "Gender"      "Age"           "HeartRate"
[4] "Temperature"
```

Accessing observations in the BodyTemperature data frame is the same as be-
fore. We can use square brackets [,] with the row and column indices or the $
operator with the variable name:

```
> BodyTemperature[1:3, 2:4]

   Age HeartRate Temperature
1   33        69        97.0
2   32        72        98.8
3   42        68        96.2

> BodyTemperature$Age[1:3]

[1] 33 32 42
```

B.6 Lists

The data frames we created above (either directly or by importing a text file) include
vectors of different types, but all the vectors have the same length. To combine
objects of different types and possibly with different length into one object, we use
lists instead. For example, suppose that we want to store the above body temperature
data along with the name of investigators and students who have been involved in
the study. We can create a list as follows:

```
> our.study <- list(data = BodyTemperature,
+     investigators = c("Smith", "Jackson",
+         "Clark"), students = c("Steve", "Mary"))
> length(our.study)

[1] 3
```

We have created a list with three *components*: data, investigators, and stu-
dents. Each component has a name, which can be used to access that component:

```
> our.study$investigator

[1] "Smith"   "Jackson" "Clark"
```

The components are ordered, so we can access them using double square brackets
"[[ ]]":

```
> our.study[[2]]
```

```
[1] "Smith"    "Jackson" "Clark"
```

If the component is a matrix or a vector, we can access its individual elements as before:

```
> our.study[[2]][3]
```

```
[1] "Clark"
```

```
> our.study[[1]][2:4, ]
```

```
  Gender Age HeartRate Temperature
2      M  32        72        98.8
3      M  42        68        96.2
4      F  33        75        97.8
```

B.7 Loading Add-on Packages

A package includes a set of functions that are commonly used for a specific application of statistical analysis. R users have been creating new packages and making them publicly available on CRAN (Comprehensive R Archive Network). To use a package, you first need to download it from CRAN and install it in your local R library. For this, we can use the install.packages() function. For example, suppose that we want to perform Biodemographic analysis. The "Biodem" package, which is created by Boattini et al., provides a number of functions for this purpose. The following command downloads the Biodem package:

```
> install.packages("Biodem", dependencies = TRUE)
```

The first argument specifies the name of the package, and by setting the option dependencies to "TRUE", we install all other packages on which "Biodem" depends. The reference manual for this package is available at http://cran.r-project.org/web/packages/Biodem/Biodem.pdf.

   After we install a package, we need to load it in R in order to use it. For this, we use the library() command:

```
> library(Biodem)
```

Now we can use all the functions available in this package. We can also use all the data sets included in the package. For example, the Biodem package includes a data set called valley where every row corresponds to a different marriage record. To use this data set, we enter the following command:

```
> data(valley)
```

The data set becomes available in the workspace as a data frame.

One of the most widely used packages in R is the MASS package. This package is automatically installed when you install R. However, you still need to load it before you can use it. Enter the command library(MASS) to load this package. One of the data sets available in the MASS package is the birthwt data set, which includes the birthweight, bwt, of newborn babies.

B.8 Conditional Statements

The birthwt data set from the MASS package includes a binary variable, low, that indicates whether the baby had low birthweight. Low birthweight is defined as having birthweight lower than 2500 grams (2.5 kilograms). Suppose that we did not have this variable and we wanted to create it. First, let us load the birthwt data set into R:

```
> data(birthwt)
> dim(birthwt)

[1] 189  10
```

The data set includes 189 cases and 10 variables.

We now create an empty vector, called low, of size 189:

```
> low <- rep(NA, 189)
```

Alternatively, we could use create an empty numerical vector without specifying its length using the numeric() function:

```
> low <- numeric()
```

Note that this way we specify the type of the object. We would have used character(), or data.frame(), or list(), if we wanted to create an empty object of the type character, or data frame, or list.

Now we want to examine the birthweight of each baby, and *if* it is below 2500, we assign the value of "1" to the low variable; otherwise, we assign the value "0". The general format for an if() statement is

```
> if (condition) {
+     expression
+ }
```

If the condition is true, R runs the commands represented by expression. Otherwise, R skips the commands within the brackets { }.

Try an if() statement to set the low of the first observation:

```
> if (birthwt$bwt[1] < 2500) {
+     low[1] <- 1
+ }
```

Check the result:

```
> birthwt$bwt[1]
```

```
[1] 2523
```

```
> low[1]
```

```
[1] NA
```

Since the condition was not true (i.e., bwt is not below 2500), the expression was not executed. To assign the value "0", we can use an else statement along with the above if statement. The general format for if-else() statements is

```
> if (condition) {
+     expression1
+ } else {
+     expression2
+ }
```

If the condition is true, R runs the commands represented by expression1; otherwise, R runs the commands represented by expression2. For example, we can use the following code to decide whether the first baby in the birthwt data has low birthweight or not:

```
> if (birthwt$bwt[1] < 2500) {
+     low[1] <- 1
+ } else {
+     low[1] <- 0
+ }
> birthwt$bwt[1]
```

```
[1] 2523
```

```
> low[1]
```

```
[1] 0
```

Conditional statements can have multiple else statements to test multiple conditions:

```
> if (condition1) {
+     expression1
+ } else if (condition2) {
```

```
+       expression2
+ } else {
+       expression3
+ }
```

B.9 Loops

To apply the above conditional statements to all observations, we can use a `for()` loop, which has the general format

```
> for (i in 1:n) {
+       expression
+ }
```

Here, i is the loop counter that takes values from 1 through n. The `expression` within the loop represents the set of commands to be repeated n times. For example, the following R commands create the vector y based on the vector x one element at a time:

```
> x <- c(3, -2, 5, 6)
> y <- numeric()
> for (i in 1:4) {
+       y[i] <- x[i] + 2 * i
+ }
> y
```

```
[1]   5   2 11 14
```

For the example discussed in the previous section, we use the following loop:

```
> for (i in 1:189) {
+       if (birthwt$bwt[i] < 2500) {
+           low[i] <- 1
+       }
+       else {
+           low[i] <- 0
+       }
+ }
```

The counter starts from 1 (i.e., the first row), and it ends at 189 (i.e., the last row). At each iteration, evaluate the conditional expression `birthwt$bwt[i] < 2500`. If the expression is true, it sets the value of `low` for that row to 1, otherwise, it sets it to 0. The variable `low` you created using the above loop and conditional statements will be exactly the same as the existing variable `low` in the data frame `birthwt`.

B.10   Creating Functions

So far, we have been using R to perform specific tasks by creating objects and applying functions to them. If we need to repeat the same task over and over again under different settings, a more efficient approach would be to create a *function* which can be called repeatedly. The function we create is itself an object and is similar to existing functions in R, such as sum(), log(), and matrix(), that we have been using. The general form of creating a function is as follows:

```
> fun.name <- function(arg1, arg2, ...) {
+      expression1
+      expression2
+      ...
+      return(list = c(out1 = output1,
+      out2 = output2, ...))
+ }
```
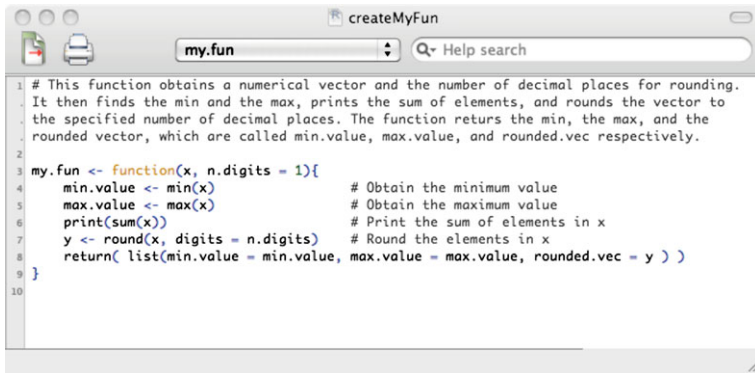
For example, suppose that we routinely need to find the min and max for a given numerical vector and print the sum of its elements. Also, we need to round the elements of the vector. However, the number of decimal places could be different for different vectors. Instead of writing the codes to create the function in *R Console*, it is better to write it in a file using a text editor so that we can modify it later. For this, click File → New Document from the menu bar. This will open a text editor. Now we can type the following commands in the text editor to create our function (Fig. B.2):

```
> my.fun <- function(x, n.digits = 1) {
+      min.value <- min(x)
+      max.value <- max(x)
+      print(sum(x))
+      y <- round(x, digits = n.digits)
+      return(list(min.value = min.value,
+      max.value = max.value, rounded.vec = y))
+ }
```

The above function takes two inputs: a numerical vector x and the number of decimal places n.digits. For the number of decimal places, we set the default to 1. If the user does not specify the number of decimal points, the function uses the default value. For x, there is no default value, so we need to provide its value every time we use this function.

The function then creates two objects, min.value and max.value, that store the min and max of x, respectively. Next, the function prints the sum of all elements. Finally, the function creates a new vector called y, which contains the rounded values of the original vector to the number of decimal place specified by n.digits.

Using return(), we specify the outputs of the function as a list. In this case, the list has three components. The first component is called "min.value", and it contains the value of the object min.value. The second component is called

Fig. B.2 Creating a function called `my.fun()` using the text editor in R

"max.value", and it contains the value of the object `max.value`. The last component is called "rounded.vec", and it contains the new vector `y`, which was created by rounding the values of the original vector.

Note that in Fig. B.2, we wrote some comments in the text editor to explain what the function does. The comments should be always preceded by the symbol "#". R regards what we write after "#" as comments and does not execute them.

When we finish typing the commands required to create the function, we save the file by clicking `File → Save As`. When prompted, choose a name for your file. For example, we called our file "CreateMyFun.R". The file will have the ".R" extension.

So far, we have just created a file that contains the command necessary to create the function. The function has not been created yet. To create the function, we need to execute the commands. For this, we can use the `source()` function to read (evaluate) the codes from the "CreateMyFun.R" file:

```
> source("CreateMyFun.R")
```

Again, give the full address for the file if it is not located in the current working directory. We can now use our function the same way we have been using any other function. The following is an example:

```
> out <- my.fun(x = c(1.2, 2.4, 5.7), n.digits = 0)

[1] 9.3

> out

$min.value
[1] 1.2

$max.value
```

```
[1] 5.7

$rounded.vec
[1] 1 2 6
```

When we run the function, it prints the sum of all elements, which is 9.3, as we requested. The outputs will be assigned to a new object called "out". Since the output was a list, `out` will be a list, and we can print its contents by using their names.

References

1. Agresti, A.: Categorical Data Analysis. Willey, New York (2002)
2. Aitchison, J., Dunsmore, I.R.: Statistical Prediction Analysis. Cambridge University Press, Cambridge (1975)
3. Allen, D.M.: Mean square error of prediction as a criterion for selecting. Technometrics **13**(3), 469–475 (1971)
4. Buijsse, B., Weikert, C., Drogan, D., Bergmann, M., Boeing, H.: Chocolate consumption in relation to blood pressure and risk of cardiovascular disease in German adults. Eur. Heart J. **31**(13), 1616–1623 (2010)
5. Cox, D.R., Hinkley, D.V.: Theoretical Statistics. Chapman and Hall, London (1974)
6. Christensen, R., Johnson, W., Branscum, A., Hanson, T.E.: Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. Texts in Statistical Science. Taylor and Francis, London (2010)
7. Fox, J.: The R Commander: a basic-statistics graphical user interface to R. J. Stat. Softw. **14**, 1–42 (2005)
8. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall, London (2003)
9. Hand, D.J., Daly, F., McConway, K., Lunn, D., Ostrowski, E.: A Handbook of Small Data Sets, 1st edn. Chapman & Hall Statistics Texts. Chapman and Hall/CRC, London (1993)
10. Harrell, F.E.: Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York (2001)
11. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, Berlin (2009). http://www-stat.stanford.edu/~tibs/ElemStatLearn/
12. Houchens, R.L., Schoeps, N.: Comparison of hospital length of stay between two insurers for patients with pediatric asthma. In: Peck, L.H.R., Goodman, A. (eds.) Statistical Case Studies: A Collaboration Between Academe and Industry, pp. 45–64. The American Statistical Society, and the Society for Industrial and Applied Mathematics, Philadelphia (1998)
13. Kahn, H.S., Cheng, Y.J., Thompson, T.J., Imperatore, G., Gregg, E.W.: Two risk-scoring systems for predicting incident diabetes mellitus in US adults age 45 to 64 years. Ann. Intern. Med. **150**, 741–751 (2009)
14. Kettunen, J.A., Harilainen, A., Sandelin, J., Schlenzka, D., Seitsalo, S., Hietaniemi, K., Malmivaara, A., Kujala, U.M.: Knee arthroscopy and exercise versus exercise only for chronic patellofemoral pain syndrome: a randomized controlled trial. BMC Med. **5**, 38 (2007)
15. Layman, P.R., Agyras, E., Glynn, C.J.: Iontophoresis of vincristine versus saline in post-herpetic neuralgia: a controlled trial. Pain **25**, 165–170 (1986)
16. Levine, P.H.: An acute effect of cigarette smoking on platelet function: a possible link between smoking and arterial thrombosis. Circulation **48**(3), 619–623 (1973)

17. Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., Ramig, L.O.: Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans. Biomed. Eng. **56**(4), 1015–1022 (2009)

18. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. Wiley-Interscience, New York (2002)

19. Mackowiak, P.A., Wasserman, S.S., Levine, M.M.: A critical appraisal of 98.6°F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. JAMA **268**, 1578–1580 (1992)

20. McGann, K.P., Marion, G.S., Camp, L., Spangler, J.G.: The influence of gender and race on mean body temperature in a population of healthy older adults. Arch. Fam. Med. **2**(12), 1265–1267 (1993)

21. Morewedge, C.K., Huh, Y.E., Vosgerau, J.: Thought for food: imagined consumption reduces actual consumption. Science **330**(6010), 1530–1533 (2010)

22. Mufunda, J.: Body mass index and blood pressure: where are we now? J. Hum. Hypertens. **21**(1), 5–7 (2007)

23. Nafiu, O.O., Burke, C., Lee, J., Voepel-Lewis, T., Malviya, S., Tremper, K.K.: Neck circumference as a screening measure for identifying children with high body mass index. Pediatrics **126**(2), 306–310 (2010)

24. Norton, P.G., Dunn, E.V.: Snoring as a risk factor for disease: an epidemiological survey. Br. Med. J. **291**, 630–632 (1985)

25. Penrose, K., Nelson, A., Fisher, A.: Generalized body composition prediction equation for men using simple measurement techniques. Med. Sci. Sports Exerc. **17**(2), 189 (1985)

26. Phillips, D.P., Barker, G.E.C.: A July spike in fatal medication errors: a possible effect of new medical residents. J. Gen. Intern. Med. **25**(8), 774–779 (2010)

27. Rubin, D.B.: Estimating causal effects from large data sets using propensity scores. Ann. Intern. Med. **127**(8 Pt 2), 757–763 (1997)

28. Sacks, F.M., Svetkey, L.P., Vollmer, W.M., Appel, L.J., Bray, G.A., Harsha, D., Obarzanek, E., Conlin, P.R., Miller, E.R., Simons-Morton, D.G., Karanja, N., Lin, P.H.: DASH-sodium collaborative research group: effects on blood pressure of reduced dietary sodium and the dietary approaches to stop hypertension (DASH) diet. DASH-sodium collaborative research group. N. Engl. J. Med. **344**(1), 3–10 (2001)

29. Scheffe, H.: The Analysis of Variance. Wiley, New York (1959)

30. Seeman, M.V.: The role of estrogen in schizophrenia. J. Psychiatry Neurosci. **21**(2), 123–127 (1996)

31. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S.: Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Greenes, R.A. (ed.) Proceedings of the Symposium on Computer Applications in Medical Care, Washington, 1988, pp. 261–265. IEEE Computer Society Press, Los Alamitos (1988)

32. Sturges, H.A.: The choice of a class interval. Am. Stat. Assoc. **21**, 65–66 (1926)

33. Taubert, D., Roesen, R., Lehmann, C., Jung, N., Schömig, E.: Effects of low habitual cocoa intake on blood pressure and bioactive nitric oxide: a randomized controlled trial. JAMA J. Am. Med. Assoc. **298**(1), 49–60 (2007)

34. Team, R.D.C.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2005)

35. Teasdale, N., Bard, C., Larue, J., Fleury, M.: On the cognitive penetrability of posture control. Exp. Aging Res. **19**(1), 1–13 (1993)

36. The Steering Committee of the Physicians' Health Study Research Group: Findings from the aspirin component of the ongoing Physicians' Health Study. N. Engl. J. Med. **318**, 262–264 (1988)

37. Venables, W.N., Ripley, B.D.: Main Package of Venables and Ripley's Mass. R Package Version, 7.3-8 edn. (2010)

38. Weerahandi, S.: Exact Statistical Methods for Data Analysis. Springer, Berlin (2003)

39. Witte, A.V., Fobker, M., Gellner, R., Knecht, S., Floel, A.: Caloric restriction improves memory in elderly humans. Proc. Natl. Acad. Sci. USA **106**(4), 1255–1260 (2009)

Index

A
`abline`, 144, 287
Agglomerative, 295
Allele, 84, 103
Alternative hypothesis, 173
Analysis of variance, 221, 225
ANOVA, 221, 257
`aov`, 231, 232
`as.integer`, 329
Association, 4, 62, 243
Average linkage, 296

B
Bar plot, 23
`barplot`, 52
Bayes' rule, 98
Bayes' theorem, 98
Bayesian, 303
Bayesian statistics, 101, 303
Bernoulli, 164, 165
Bernoulli trials, 115, 139
Beta, 141, 306
Between group variation, 2, 223
Bimodal, 29
Binomial, 115, 143
Blocking, 14, 230
`boxplot`, 54, 78
Boxplots, 38, 54
`by`, 78, 79

C
Case-control studies, 13
Categorical, 235
Causation, 4, 62, 204
`cbind`, 332, 335
Central limit theorem, 163, 195, 229
Centroid linkage, 296, 300

Chi-squared test, 248, 250
`chisq.test`, 248–250
Cluster sampling, 12
Clustering, 291
`colnames`, 249, 333
Complement, 87, 89, 109
Complete linkage, 295, 301
`complete.cases`, 57
Conditional, 93, 341
Conditional probability, 93, 94, 175
Conditional statements, 55, 341, 342
Confidence coefficient, 160
Confidence interval, 162, 261
Confidence level, 160, 161, 163
Confounding, 4, 5
Contingency table, 72, 208, 241
Continuous, 110, 131
`cor`, 77
`cor.test`, 217, 218
Correlation, 66, 68
Correlation test, 214, 217
Count data, 18, 153
Cross-sectional, 14
Cumulative distribution function, 131
`cutree`, 300

D
`data`, 51, 285
Data exploration, viii, 17
Data frame, 336
Data preprocessing, 40
Data transformation, 43
`data.frame`, 336, 341
`dbeta`, 313
`dbinom`, 144
Decision, 5, 106, 311
Decision theory, 5, 106