

Introduction to statistical data analysis with R

Matthias Kohl



Download free books at

bookboon.com

Matthias Kohl

Introduction to statistical data analysis with R



Introduction to statistical data analysis with R

1st edition

© 2015 Matthias Kohl & bookboon.com

ISBN 978-87-403-1123-5

Contents

Preface	9
1 Statistical Software R	10
1.1 R and its development history	10
1.2 Structure of R	12
1.3 Installation of R	13
1.4 Working with R	14
1.5 Exercises	17
2 Descriptive Statistics	18
2.1 Basics	18
2.2 Excursus: Data Import and Export with R	22
2.3 Import of ICU-Dataset	25
2.4 Categorical Variables	29
2.5 Metric Variables	52
2.6 Exercises	78



www.sylvania.com

We do not reinvent the wheel we reinvent light.

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM

OSRAM
SYLVANIA



Click on the ad to read more

3	Colors and Diagrams	79
3.1	Colors	79
3.2	Excursus: Export of Diagrams	86
3.3	Diagrams	89
3.4	Exercises	97
4	Probability Distributions	98
4.1	Discrete Distributions	99
4.2	Continuous Distributions	117
4.3	Exercises	136
5	Estimation	138
5.1	Introduction	138
5.2	Point Estimation	140
5.3	Confidence Intervals	157
5.4	Exercises	175



CHALLENGING PERSPECTIVES

Internship opportunities

EADS unites a leading aircraft manufacturer, the world's largest helicopter supplier, a global leader in space programmes and a worldwide leader in global security solutions and systems to form Europe's largest defence and aerospace group. More than 140,000 people work at Airbus, Astrium, Cassidian and Eurocopter, in 90 locations globally, to deliver some of the industry's most exciting projects.

An **EADS internship** offers the chance to use your theoretical knowledge and apply it first-hand to real situations and assignments during your studies. Given a high level of responsibility, plenty of learning and development opportunities, and all the support you need, you will tackle interesting challenges on state-of-the-art products.

We welcome more than 5,000 interns every year across disciplines ranging from engineering, IT, procurement and finance, to strategy, customer support, marketing and sales. Positions are available in France, Germany, Spain and the UK.

To find out more and apply, visit www.jobs.eads.com. You can also find out more on our **EADS Careers Facebook page**.

AIRBUS **ASTRIUM** **CASSIDIAN** **EUROCOPTER**

EADS



Click on the ad to read more

6	Statistical Tests	177
6.1	Introduction	177
6.2	Examples	187
6.3	Exercises	207
Software versions		209
Bibliography		210
Index		216

**Deloitte.**

© Deloitte & Touche LLP and affiliated entities.

Discover the truth at www.deloitte.ca/careers

Click on the ad to read more

List of Figures

Figure 1.1: R GUI (64-bit) on Windows (German system).	15
Figure 1.2: RStudio IDE after installation on Ubuntu Linux (German system).	16
Figure 1.3: RStudio IDE after opening a new R script on Ubuntu Linux (German system).	16
Figure 2.1: Interplay between probability theory, descriptive and inferential statistics.	19
Figure 2.2: Types of attributes and scales of measurement.	21
Figure 2.3: RStudio window for import of text files.	23
Figure 2.4: RStudio window <i>Environment</i> with a data object.	24
Figure 2.5: View of the exact structure of a dataset in RStudio.	28
Figure 2.6: Interactive context based help in RStudio.	32
Figure 2.7: Installation of R packages in RStudio.	32
Figure 2.8: The values in a box-and-whisker plot.	40
Figure 2.9: Examples of skewness.	61
Figure 2.10: Examples of kurtosis.	63
Figure 3.1: A negative example for using colors and diagrams.	80
Figure 3.2: A negative example with improved colors.	83
Figure 3.3: From a negative to a positive example.	83
Figure 3.4: RStudio window <i>Plots</i> with an example.	86
Figure 3.5: RStudio window for saving a plot as image.	87
Figure 3.6: RStudio window for saving a plot as pdf file.	87
Figure 3.7: Order the categories!	90
Figure 3.8: Once again: Order the categories!	90
Figure 3.9: And once again: Order the categories!	91
Figure 5.1: Illustration of unbiased and efficient.	141
Figure 5.2: Ratio between 95~ quantiles of t and standard normal distribution.	164
Figure 6.1: Sample size dependent on effect size.	184
Figure 6.2: Sample size dependent on variance.	184

List of Tables

Table 2.1: Overview of some basic functions for data import with R.	22
Table 3.1: Overview of devices supported by R.	88
Table 4.1: Notions from statistics and their counterparts in probability theory.	135
Table 6.1: Decision situation in case of statistical tests.	179
Table 6.2: Example of a 2×2 contingency table.	196

Preface

Statistics is everywhere today and we are steadily, knowingly or unknowingly, confronted with results of statistical procedures. Examples are internet search engines, targeted ads on websites, assessments of our creditworthiness, reference ranges of blood tests, weather forecast, election forecast, and many more. Often, statistical procedures are not appropriately applied or their results are not properly reported. Therefore, basic statistical knowledge is not only important in professional but also in everyday life and helps to distinguish between correct and incorrect information.

The basis of this book are my lecture notes of several statistics courses I gave in recent years at Furtwangen University, Campus Villingen-Schwenningen, in the framework of various bachelor and master programs as well as at Freiburg University in the framework of the international master program in biomedical sciences (IMBS).

As the title of the book already indicates, the introduction to statistical analysis happens by using the statistical software R (R Core Team (2015a)), a free software that is available for most operating systems. The R code used in the book is contained in the file www.stamats.de/RCodeEN.zip in form of text files with file extension .R. The R code of each chapter runs independent of the other chapters.

Note:

For the book several messages generated by R were wittingly suppressed to save space and to keep focus on the essentials. The suppressed messages are of no importance for the presented analyses. Conversely, you should be aware that there might be additional messages when you run the code contained in this book. This also includes innocuous warning messages.

The book was written using the software package LATEX in combination with pdfLATEX. In addition, the contributed package "knitr" (Xie (2015)) of the statistical software R was applied, which offers flexible options for combining explanations with input and output of R.

Villingen-Schwenningen August 2015

Matthias Kohl

1 Statistical Software R

The chapter includes a short introduction to the statistical software R where the following issues are covered:

- development history based on the statistical programming language S
- modular structure in form of packages
- installation on various operating systems
- installation of the integrated development environment (IDE) RStudio

Working with R in practice is introduced in the subsequent chapters in combination with the introduction to statistical data analysis.

1.1 R and its development history

The statistical software R (R Core Team (2015a)) is a free, non-commercial implementation of the statistical programming language R developed at the AT&T Bell Laboratories by Rick Becker, John Chambers and co-workers. It is a development environment and a programming language for statistics and graphics developed under GNU GPL-2/3 and therefore can be installed on arbitrary many computers without any restriction.

R is a function based language. That is, all actions are initiated by calling functions. In doing so additional parameters (arguments) are frequently passed to the functions controlling the concrete execution of the function. The function is identified by its name, the parameters by their name or also by their position. A call has the following structure (not always directly visible):

```
FunctionName(parameter1 = value1, parameter2 = value2, ..., parameterN = valueN)
```

We will see many examples in the course of the book.

We briefly summarize the development history of S and R:

05.05.1976: start of the development of version 1 of S (Chambers (2008, p. 476))

1980: release of version 2 of S (Chambers (2000))

1988: release of version 3 of S (S3) (Chambers (2000))

1992: start of the R project by Ross Ihaka and Robert Gentleman (Hornik (2008))

August 1993: first files of R published on Statlib (Ihaka (1998)).

Juni 1995: publication of the first GPL (GNU General Public License) version of R (Ihaka (1998))

05.12.1997: the R project officially becomes a GNU project (Ihaka (1997)).

1998: release of version 4 of S (S4) (Chambers (2000))

29.02.2000: R 1.0.0 released, an implementation of S3 (Hornik (2008))

04.10.2004: R 2.0.0 released, an advanced version of S4 (Chambers (2008), Hornik (2008))

22.04.2010: R 2.11.0 released, support of Windows 64bit-systems (Dalgaard (2010))

03.04.2013: R 3.0.0 released, unlimited memory allocation in case of 64bit-systems (Dalgaard (2013))

18.06.2015: R 3.2.1 released, version used for writing the book (Dalgaard (2015))

In general, there is a new release (version R x.y.0) in spring (March/April) of each year with patches released (R x.y.1, R x.y.2, etc.) over the year as necessary (R Core Team (2015c)).

SIMPLY CLEVER

ŠKODA



We will turn your CV into
an opportunity of a lifetime



Do you like cars? Would you like to be a part of a successful brand?
We will appreciate and reward both your enthusiasm and talent.
Send us your CV. You will be surprised where it can take you.

Send us your CV on
www.employerforlife.com



The base system of R is developed by the so-called R Core Development Team currently consisting of 21 members (The R Foundation (2015a)). In addition, in 2002 the R Foundation (The R Foundation (2015b)) has been founded where the R Core Development Team members participate as ordinary members. The goals of the foundation include continuation of the development of R, the investigation of new methods, teaching and training in the area of computational statistics, and organisation of assemblies and conferences focused on computational statistics.

Furthermore, an R Consortium has been founded in June 2015 under the umbrella of the Linux Foundation for a stronger support of R from industry. Members are companies such as Microsoft, Google, Oracle, and HP (The Linux Foundation (2015)).

Muenchen (2015) tries to estimate the popularity and the market share of data analysis software. The statistical software R performs well in all statistics and today plays a central and in some fields even leading role.

1.2 Structure of R

The statistical software R consists of packages that are organized in one or more libraries. There are three categories of packages. First of all, there are the **base packages** providing the basic functionality of R, which are maintained by the R Core Development Team. Currently, these are the following 14 packages: "base", "compiler", "datasets", "grDevices", "graphics", "grid", "methods", "parallel", "splines", "stats", "stats4", "tcltk", "tools", "utils"; for more information see Section 5 in the FAQs of R (Hornik (2015)).

The second group of packages, which are also part of the default installation of R, are the **recommended packages**. These packages mainly include additional, more complex statistical procedures. Currently, there are the following 15 packages: "boot", "class", "cluster", "codetools", "foreign", "KernSmooth", "lattice", "MASS", "Matrix", "mgcv", "nlme", "nnet", "rpart", "spatial", "survival" (Hornik (2015, Section 5)).

Finally, there are the **contributed packages**. Due to the open nature of R, anyone can contribute new packages anytime, which for sure is an important aspect for the success and the wide distribution of R. There is a continuously increasing developer community steadily contributing new packages to R, where the number of contributed packages grows exponentially for more than ten years now. Currently, there are already more than 9 000 packages (Muenchen (2015)). Those packages are spread over several socalled repositories. The largest number of packages are on CRAN (Comprehensive R Archive Network, <http://cran.r-project.org/>). It currently contains about 7 000 packages. Contributed packages for the analysis of genomic data are mainly part of Bioconductor (Gentleman et al. (2004), <http://www.bioconductor.org/>), which currently provides more than 1 000 packages for download. Further important repositories are Omega (<http://www.omegahat.org/>) with currently about 100 packages and GitHub (<https://github.com/>).

1.3 Installation of R

The necessary files for installing R under Windows, Mac OS X, or Linux can be downloaded from CRAN (<http://cran.r-project.org/>) or one of its mirrors. In general, the installation of R does not differ from the installation of other software on these operating systems.

Windows: The Windows installer for 32- and 64-bit can be found under <http://cran.r-project.org/bin/windows/base/>. Further information about the installation, updates or also uninstalling are included in the FAQs for Windows (Ripley and Murdoch (2015)).

Mac OS X: The necessary files for Mac OS X as well as a brief manual are given at <http://cran.r-project.org/bin/macosx/>. Similar to Windows there is also a FAQ page for Mac OS X (Iacus et al. (2015)) including additional information.

Linux: There are files for

- Debian (<http://cran.r-project.org/bin/linux/debian/>, Ranke (2015))
- OpenSUSE (<http://cran.r-project.org/bin/linux/suse/>, Steuer (2015))
- Red Hat Enterprise Linux (RHEL), CentOS, Scientific Linux, Oracle Linux (<http://cran.r-project.org/bin/linux/redhat/>, Plummer (2015))
- Ubuntu (<http://cran.r-project.org/bin/linux/ubuntu/>, Rutter (2015))

These websites include also brief manuals describing the installation.

The official and comprehensive documentation for the installation of R is the manual “R Installation and Administration” (R Core Team (2015d)). It also includes descriptions on how to install R from the source files.

1.4 Working with R

Starting R under Windows opens a simple graphical user interface (GUI) shown in Figure 1.1. One can now start to enter R commands in the R *Console* window. This works for simple computations but not for a real data analysis, which should be well documented and which we might want to repeat in the same or a slightly modified form for a different dataset. In this case it is recommended to generate a text file including the R commands. We can use any text editor for this purpose where it is common to use `.r` or `.R` as file extension. However, in programming it is common practice to go one step further and use a text editor with additional functionality or an integrated development environment (IDE).

Depending on the operating system there are several options. Grosjean (2012) has compiled an overview, which is probably not current anymore. It seems that the largest functionality is currently provided by the free and open source IDE RStudio (<http://www.rstudio.org/>). It can be installed under Linux, Windows, and Mac OS X. I currently use it for data analysis as well as in my lectures.



The advertisement features the Swedish flag and the word "Sverige" above the university's name. It highlights the university's innovation and ranking, and encourages interest in Computer Science through an English-taught master's degree. A call-to-action button invites users to click for more information. The Linköping University logo is at the bottom left, and a green callout bubble at the bottom right encourages users to click on the ad to read more.

Sweden
Sverige

Linköping University –
innovative, highly ranked,
European

Interested in Computer Science? Kick-start your career
with an English-taught master's degree.

→ Click here!

LiU LINKÖPING
UNIVERSITY

Download free eBooks at bookboon.com

14

Click on the ad to read more

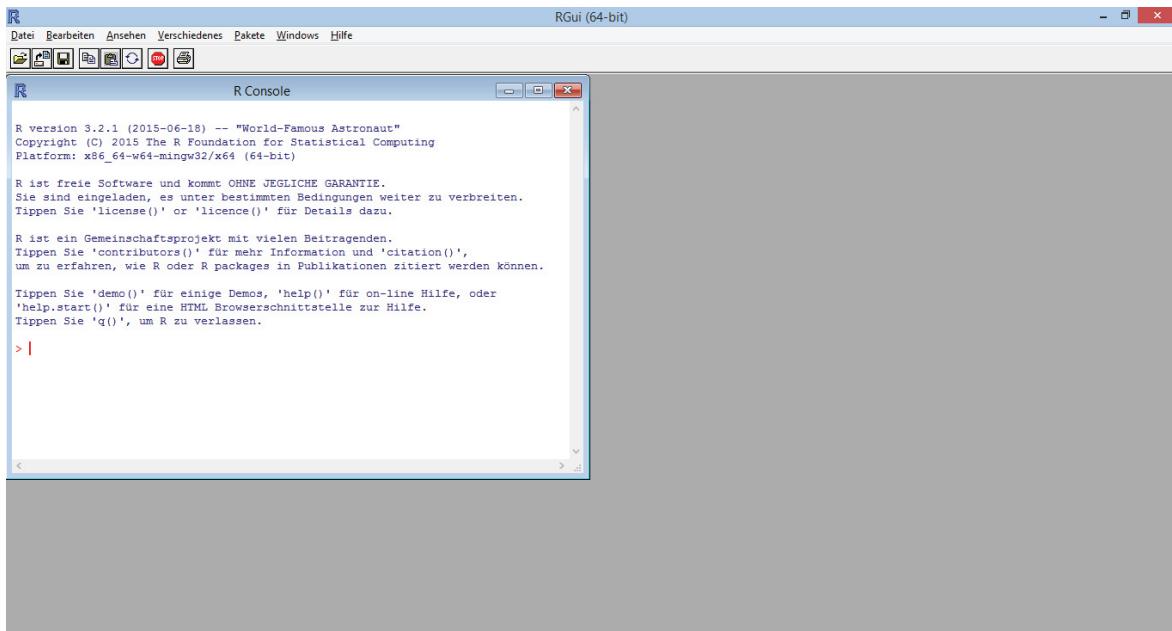


Figure 1.1: R GUI (64-bit) on Windows (German system).

Even one step further are specialized GUIs. There are also some options for R. An overview, which is probably also not current any more, is provided by Grosjean (2011).

Figure 1.2 shows the RStudio IDE after installation on my Ubuntu Linux system. It looks very similar on Windows and Mac OS X. You can see three of the four panes. On the left hand side there is the *R Console*, in which the statistical software R is running. On the top of the right hand side the windows *Environment* and *History* are shown. *Environment* shows all R objects that are currently loaded or were generated during the current session. As RStudio has just been started, the *Environment* is empty. The *History* contains an history of the R commands that have been executed. On the bottom of the right hand side there are the windows *Files*, *Plots*, *Packages*, *Help*, and *Viewer*. *Files* shows a file browser, which after the start shows the current working directory. Window *Plots* includes the plots generated in the current session and hence is empty immediately after starting RStudio. In window *Packages* all packages installed on the system are shown and can also be loaded via this window. Window *Help* provides several ways of help (local and online) for R and RStudio. Finally, in window *Viewer* local websites or web applications can be displayed.

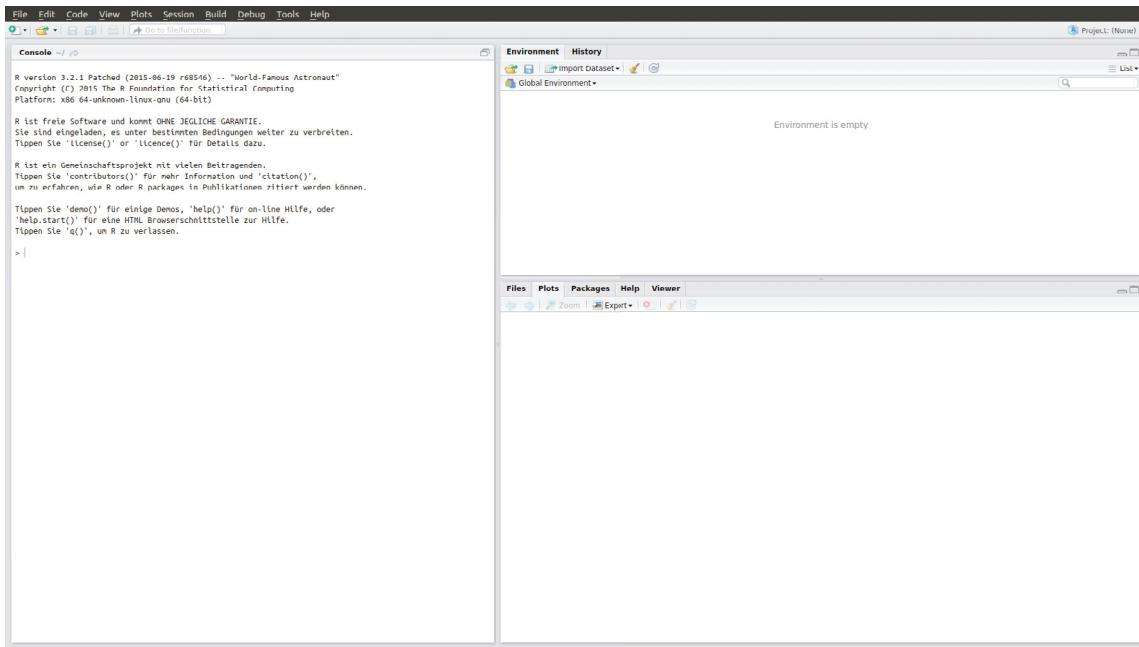


Figure 1.2: RStudio IDE after installation on Ubuntu Linux (German system).

After opening a new R script by using the menu item *File* → *New File* → *R Script*, a fourth window becomes visible (see Figure 1.3). It contains an empty and yet unnamed text file – a so-called R script. Later on, we will see that text input is supported by several interactive functions, which make it easier for beginners to write error free R code. Single R commands or also marked command blocks can be sent to the *R Console* for execution via the menu item *Run*. By means of the menu item *Source* the whole R script can be executed. The arrangement of the panes can be changed via the menu item *Tools* → *Global Options...* → *Pane Layout*. More details about RStudio will be presented in the course of this book.

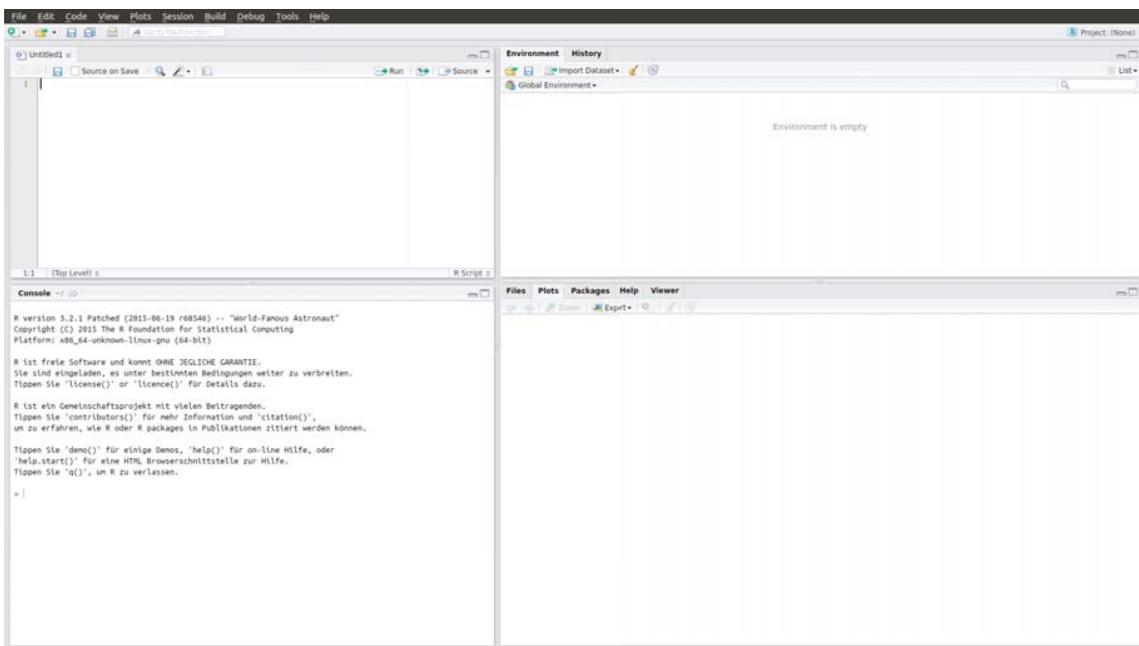


Figure 1.3: RStudio IDE after opening a new R script on Ubuntu Linux (German system).
Download free eBooks at bookboon.com

1.5 Exercises

1. Install R and RStudio on your personal computer, notebook, etc.
2. Start RStudio, open a new R script and take a close look at all opened windows and all menu items.
3. Acquaint yourself with the help options available in window *Help*.
4. Check, if the base and recommended packages are installed on your system (window *Packages*). Which R packages are checked after starting RStudio and hence are active, i.e. are loaded and can immediately be applied?

**I joined MITAS because
I wanted **real responsibility****

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com



Month 16

I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements

Download free eBooks at bookboon.com



2 Descriptive Statistics

The chapter is about descriptive statistics where the following topics are covered:

- Interplay of probability theory, descriptive and inferential statistics
- Types of attributes and scales of measurement
- Basic function for data import and export with R
- Data import of text files with RStudio
- Frequency tables, bar and pie charts
- Mode, quantile, quartile, median, range, interquartile range (IQR), MAD, box-and-whisker plot
- Cross table, ϕ -coefficient, Pearson's contingency coefficient, Cramér's V
- Spearman's P , Kendall's τ , scatter plot
- Arithmetic mean, geometric mean, standard deviation, coefficient of variation, quartile coefficient of dispersion
- histogram, density estimation
- Pearson (product-moment) correlation coefficient

The R code of this chapter is included in R script `DescriptiveStatistics.R`, which you can download from my website (link: www.stamats.de/RCodeEN.zip). The least difficulties arise, if you save my R scripts in the same folder as the data. In addition, you should use your own R script to experiment with your own R code. Please select *New File → R script* in menu item *File* of RStudio. By doing this, an empty file is opened in the editor window of RStudio. Please select a meaningful file name and save the file via *File → Save*, preferably in the folder of file `ICUData.csv`.

2.1 Basics

Figure 2.1 provides an overview of the interplay between probability theory, descriptive and inferential statistics. The starting point is a **population** or **universe** that has to be clearly characterized. The goal is to obtain some (new, important) insights about this population, e.g. which party will get how many votes in the next election or which disease occurs with which frequency. A complete survey in most cases is impossible, as for instance it would be too expensive due to the size of the population, or as the population is continuously changing over time.

The statistical way out consists of postulating **models from probability theory** where the model parameters are unknown and have to be determined. For this purpose a **representative sample** is drawn from the population, usually via random selection. The task of **descriptive statistics** is to characterize this random sample as accurately as possible. That is, descriptive statistics gains no insights about the population, but describes “only” the (randomly) selected part from it. Descriptive statistics helps to become acquainted with the data and to identify uncommon or erroneous values in the data. As a consequence, it also makes an important contribution to inferential statistics, as valid inference is only possible by knowing the data and the data quality (“garbage in, garbage out”).

The goal of **inferential statistics** is to draw inferences from a representative sample about the corresponding population. An important part is to determine (estimate) the unknown parameters of assumed probability models from the available data. In addition, the validity of existing models can be examined.

Goal: New insights about a population

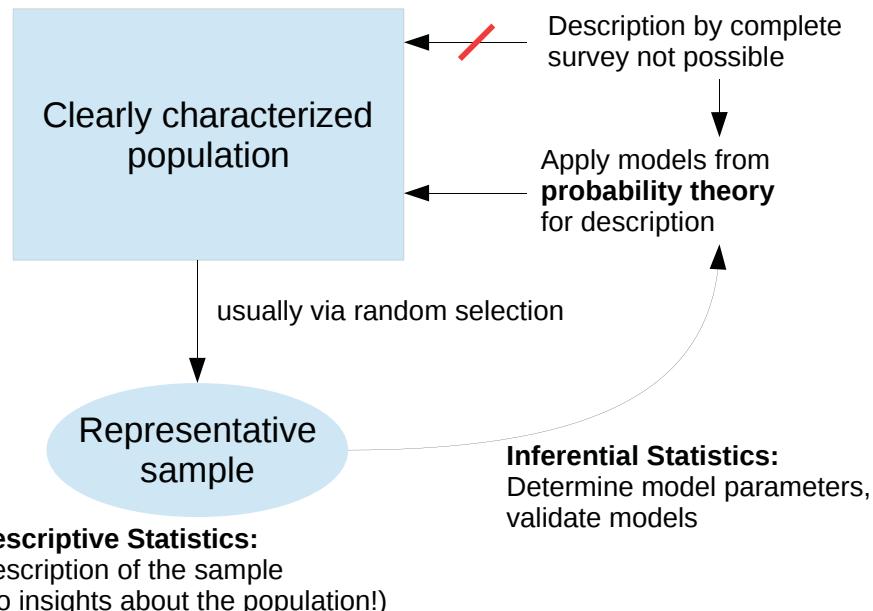


Figure 2.1: Interplay between probability theory, descriptive and inferential statistics.

Note:

We are dealing with models, i.e. we should not assume that these models exactly reflect the reality. Instead, the models under certain assumptions and at a certain time point offer a quite good description of reality. In this sense, one should interpret the following quote of the famous statistician George E.P. Box (Box and Draper (1987, p. 424)):

“Essentially, all models are wrong, but some are useful.”

The following example demonstrates that model selection is crucial for the result and that identical data under different assumptions may lead to contradictory results.

Example 2.1. In the SecondWorldWar, the goal was to better protect American bombers against fire of the German air defense. For this purpose, the location and number of bullet holes of returning airplanes were analyzed. Based on the collected information the Army concluded that the locations with extraordinary many hits should get an additional armor. A plausible result under the assumption that the German air defense especially aims at these parts of the air planes.

In contrast, the statistician Abraham Wald assumed in his analysis that the hits should be uniformly distributed over the air planes (Wald (1980)). Since this was not the case for the returning air planes, he concluded that the not returning air planes were hit at very vulnerable locations and hence crashed. Consequentially, he recommended to add armor at places where the returning air planes had no or only a few hits.

The elements of a population – which might be persons, items, etc. – are described by a number of **attributes** (variables). These attributes can be divided into several **types of attributes** as shown in Figure 2.2. The main distinction is between qualitative (categorical) and quantitative (metric) attributes.

"I studied English for 16 years but...
...I finally learned to speak it in just six lessons"

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

Download free eBooks at bookboon.com

20

Click on the ad to read more

Type of attribute	Qualitative / Categorical		Quantitative / Metric	
Scale of measurement	Nominal scale	Ordinal scale	Interval scale	Ratio scale
Examples	Gender, blood group, rhesus factor	Grades, medical scores	Temperature in °C, intelligence quotient	Temperature in Kelvin, body height
Notes	Lowest level	Order is defined	Arbitrary zero, distance is defined	Highest level, natural zero, ratio is defined
Operations	$A = B, A \neq B$	$A = B, A \neq B, A < B, A > B$	$A = B, A \neq B, A < B, A > B, d = A - B$	$A = B, A \neq B, A < B, A > B, d = A - B, r = A:B$

Figure 2.2: Types of attributes and scales of measurement.

These two categories can be divided by the so-called **scales of measurement** into nominal, ordinal, interval and ratio scaled, where nominal is the lowest and ratio scaled the highest level. In dependence of the scale of measurement, certain arithmetic operations are allowed, where the number of allowed operations increase from the left hand side (nominal) to the right hand side (ratio scaled). Therefore, it is important to know the scales of measurement of the investigated variables. Otherwise, the measured values of the variables – the so-called levels of the attributes – could for instance be wrongly described by descriptive statistical methods.

Note:

The bounds between the scales of measurements are partly fluent; e.g., in practice, a medical score with many levels is often treated like a metric variable.

The information content of variables increases with the scale of measurement. Thus, during the design of a study, one should ideally select a variable with the highest possible scale of measurement to describe an attribute. Unfortunately, this is not always possible in practice, as the measurement of more informative variables usually requires more efforts and is more expensive. As a consequence, one can not always avoid to select a less informative variable for a study.

We consider an example.

Example 2.2. Our goal is to characterize the age distribution of a sample or of the respective population. In this case, the date of birth would be more informative than age in years or age groups, where the effort to collect the data is more or less the same for all three options. Hence, the date of birth should be selected. Furthermore, this selection offers the opportunity to restrict the statistical analysis to age in years or age groups if it turns out later, that the additional information provided by date of birth is not needed or irrelevant.

2.2 Excursus: Data Import and Export with R

Before we can start with a descriptive analysis, we must first plan and conduct a study and collect data. In doing so, a variety of things have to be considered. We do not elaborate on those things here, as it would go beyond the scope of the book.

In larger studies, the collected data is often saved in specifically designed databases, in smaller studies one or several files of a spreadsheet software are usually used. In both cases, the collected data can be exported to one or several text files. Therefore, we will only consider data import from text files in this section. Beyond this, R offers a variety of options to import data such as the import of files from other statistical software packages or interfaces to databases. An overview of the various options for data import and export is included in manual “R Data Import/ Export” (R Core Team (2015b)).

The starting point for reading data from text files is function `scan`. With this function, data can be imported from the console or a text file. However, in most cases one needs not to directly apply function `scan`, but one can use function `read.table`, which is much simpler to handle. Furthermore, there are functions `read.csv`, `read.csv2`, `read.delim`, or `read.delim2` that are even more specialized; see Table 2.1.

Function name	Description
<code>scan</code>	Read data from console or a text file.
<code>read.table</code>	Read data from a text file in spreadsheet format.
<code>read.csv</code>	Special case of <code>read.table</code> with decimal point “.” and column separator “,” (“English csv-file”).
<code>read.csv2</code>	Special case of <code>read.table</code> with decimal point “,” and column separator “;” (“German csv-file”).
<code>read.delim</code>	Special case of <code>read.table</code> with decimal point “.” and column separator “\t” (tab).
<code>read.delim2</code>	Special case of <code>read.table</code> with decimal point “,” and column separator “\t” (tab).

Table 2.1: Overview of some basic functions for data import with R.

We can also use RStudio to import text files, which is especially helpful for beginners. In window *Environment* there is menu item *Import Dataset*. After selecting *From Text File...* a window opens for choosing a text file. After choosing a text file, the window shown in Figure 2.3 opens. The provided options correspond to the most important arguments of the `read.*` functions. The data is imported via one of the `read.*` functions, where the call for reading in the data is subsequently shown in figure *History*. To ensure the exact reproducibility of the import, the R code shown in figure *History* should be transferred to the current R script via the menu item *To Source*.

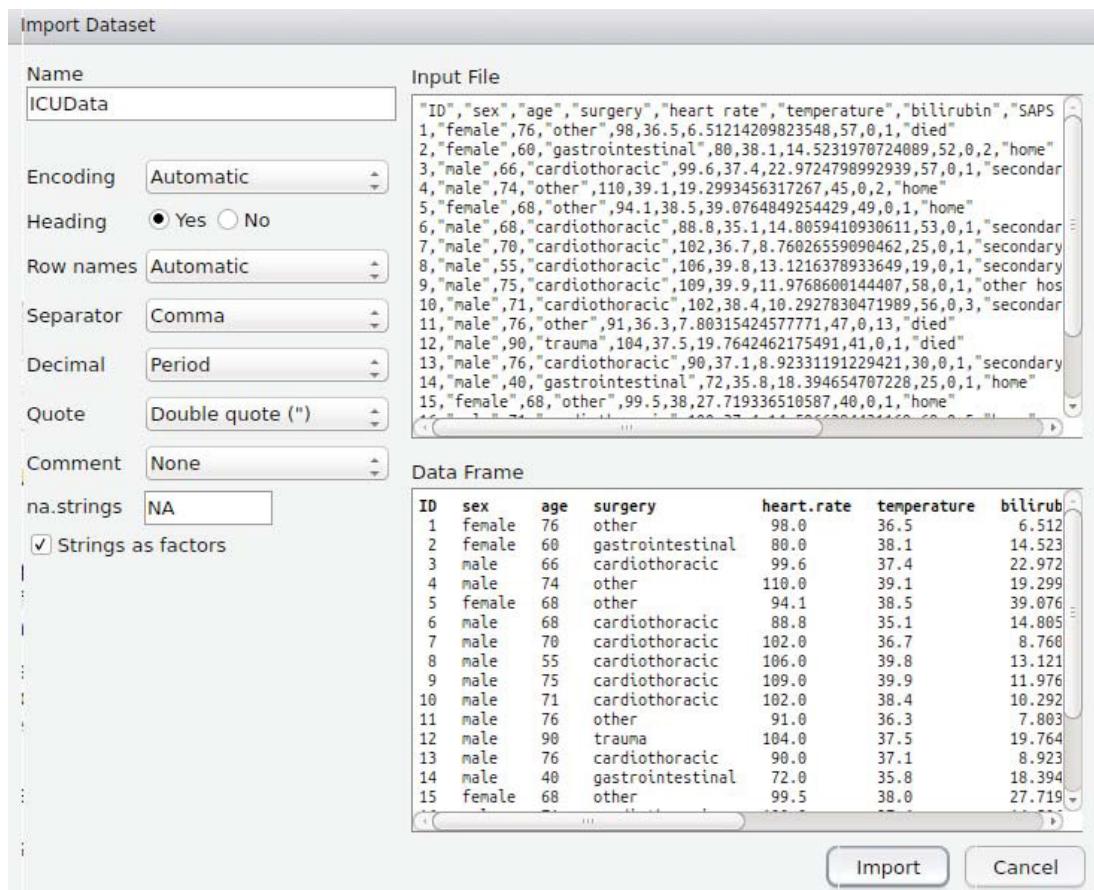


Figure 2.3: RStudio window for import of text files.

Note:

Even if the import fails, which for instance may happen if special characters are included in the file path, the R code for reading in the data is generated in window *History*. By transferring this R code to the current R script, making necessary corrections (e.g. correcting the file path) and re-running the R code one can after all import the file.

For using the result of the import for subsequent analyses, it must be assigned to some variable. The name of the variable can be specified in field *Name* (see Fig. 2.3). After the import, a data object with the chosen name is visible in window *Environment*; see Figure 2.4. The data object can be viewed in the editor window by clicking on its name.

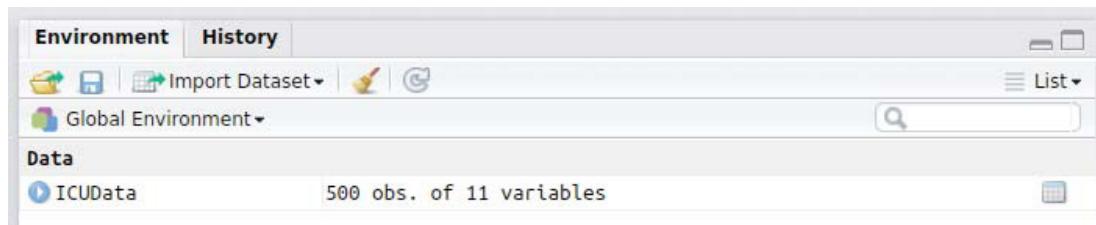


Figure 2.4: RStudio window *Environment* with a data object.

The data object is a so-called `data.frame`, the basic data structure in R for saving datasets. It is similar to a table in a spreadsheet program. The columns correspond to the variables (attributes), the rows represent the observed levels of the studied subjects.

The counterpart to the introduced `read.*` functions for exporting data are the functions `write.table`, `write.csv`, and `write.csv2`. If you work with English system settings, you should use `write.csv` for exporting data. The generated file can then be opened without problems in a current spreadsheet software.

In the past four years we have drilled
89,000 km
That's more than **twice** around the world.

Who are we?
We are the world's largest oilfield services company¹. Working globally—often in remote and challenging locations—we invent, design, engineer, and apply technology to help our customers find and produce oil and gas safely.

Who are we looking for?
Every year, we need thousands of graduates to begin dynamic careers in the following domains:

- Engineering, Research and Operations
- Geoscience and Petrotechnical
- Commercial and Business

What will you be?

Schlumberger

¹Based on Fortune 500 ranking 2011. Copyright © 2015 Schlumberger. All rights reserved.



Another form of data import is function `load`, which can be applied to load so-called `.RData`-files. These files have been generated by R function `save` or `save.image`. With these functions one can save single objects (`save`) or the entire content of an R session (`save.image`) in an `.Rdata`-file. In addition, one can specify if the file should be compressed (default) or not.

2.3 Import of ICU-Dataset

In this section, we read in the `ICUData.csv` dataset, which we will analyze in the book in various ways. It consists of data from 500 patients of an intensive care unit (ICU). The data is not from real patients, but I have generated it based on my long-term experience with data of intensive care patients. The data is similar to real data with respect to many aspects.

Please, use the following steps to import the dataset:

1. Download the dataset from my homepage and save it on your computer (Link: <http://www.stamats.de/ICUData.csv>). Avoid using special characters in the file path.
2. Start RStudio.
3. Change the working directory. Click on ... in window *Files* (at right edge) and select the folder, in which you have saved `ICUData.csv`. Next, click on *More* → *Set As Working Directory*.
4. Check the working directory by entering the following R code in window *Console*

```
1 getwd()
```

followed by the *Enter/Return*-key. The output should correspond to the folder, in which you have saved file `ICUData.csv`. If not, please repeat the above steps again.

5. Open a new R script via *File* → *New File* → *R Script*.
6. Save the (empty) R script via *File* → *Save* in the same folder, where also the file `ICUData.csv` is contained. Select an meaningful name for the file, e.g. `DescriptiveAnalysis.R`.
7. Import the ICU dataset by adding the following R code to your new R script.

```
1 ICUData <- read.csv(file = "ICUData.csv")
```

In your R script, place the cursor in the line with the above R code and click on *Run*. By doing this, the R code is copied to window *Console* and executed. There should be no output. In case there is an error message – probably

```
Error in file(file, "rt"): cannot open the connection
```

either saving the file or changing the working directory has not worked properly. Please, check steps 3 and 4 and run the R code again as described above.

As an alternative, you can use the import function of RStudio as described in Section 2.2. Please make sure that your settings match the settings visible in Figure 2.3.

8. Take a look at window *Environment* and check if there is object `ICUData` in the field *Data* (see Fig. 2.4). It must be an object of type `data.frame` with 500 observations (obs.) of 11 variables. If this is true, the import was successful.

Note:

In step 7 we have used the assignment operator `<-` to assign the result of the import via `read.csv` the name `ICUData`. That is, the data are saved in a `data.frame` with name `ICUData` and we can use this object for further analysis.

Although the import looks successful at the first glance, it is still possible that the dataset was not imported as required. Thus, I strongly recommend to check the import more precisely. First, one can use function `View` to take a closer look at the imported dataset – if it is not too large.

```
1 View(ICUData)
```

You can also achieve this by clicking on the name of the dataset in window *Environment* of RStudio. By doing this, one can for instance see, if the column names and row names (if any) were correctly transferred, if the entries in the columns are correct, and if there are empty lines or columns. As different data types look identical or very similar in this view, one should also take a closer look at the structure of the dataset. For this purpose function `str` is provided.

```
1 str(ICUData)
```

```
'data.frame': 500 obs. of 11 variables:
 $ ID : int 1 2 3 4 5 6 7 8 9 10 ...
 $ sex : Factor w/ 2 levels "female","male": 1 1 2 2 1 2 2 2 2 2 ...
 $ age : int 76 60 66 74 68 68 70 55 75 71 ...
 $ surgery : Factor w/ 5 levels "cardiothoracic",...: 4 2 1 4 4 1 1 1 1 1 ...
 ...
 $ heart.rate : num 98 80 99.6 110 94.1 88.8 102 106 109 102 ...
 $ temperature : num 36.5 38.1 37.4 39.1 38.5 35.1 36.7 39.8 39.9 38.4 ...
 $ bilirubin : num 6.51 14.52 22.97 19.3 39.08 ...
 $ SAPS.II : int 57 52 57 45 49 53 25 19 58 56 ...
 $ liver.failure: int 0 0 0 0 0 0 0 0 0 0 ...
 $ LOS : int 1 2 1 2 1 1 1 1 3 ...
 $ outcome : Factor w/ 4 levels "died","home",...: 1 2 4 2 2 4 4 4 3 4 ...
```

A similar result one can obtain in window *Environment* of RStudio by clicking on the blue arrow symbol in front of `ICUData` in the field *Data*. The result is shown in Figure 2.5.

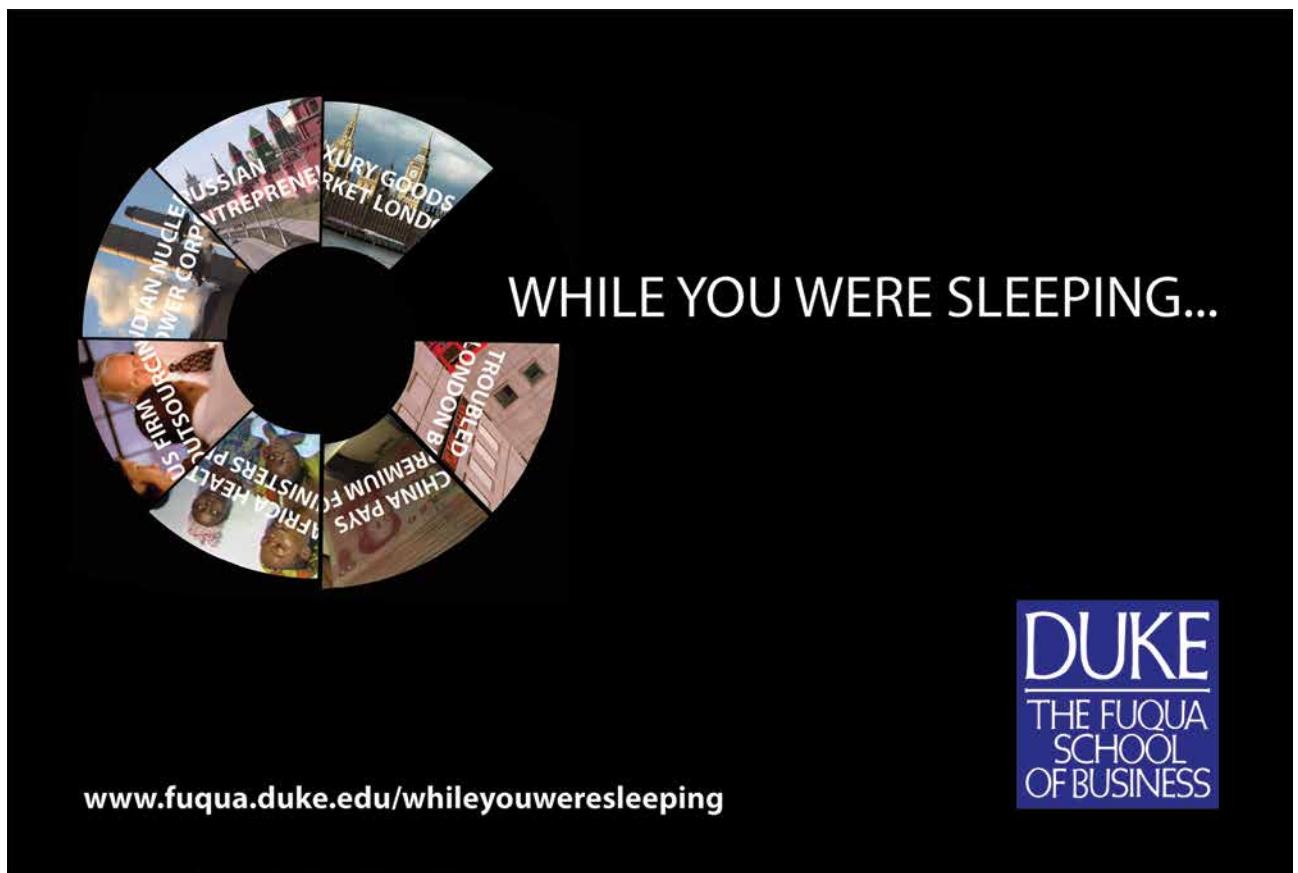
The dataset consists of the following variables:

ID: consecutive numbers (`integer`) from 1 to 500 for identification of the patients

sex: a nominal variable (`Factor`) with levels: female and male

age: age in years (`integer`)

surgery: kind of surgery, nominal variable (`Factor`) with levels: cardiothoracic, gastrointestinal, neuro, other, and trauma



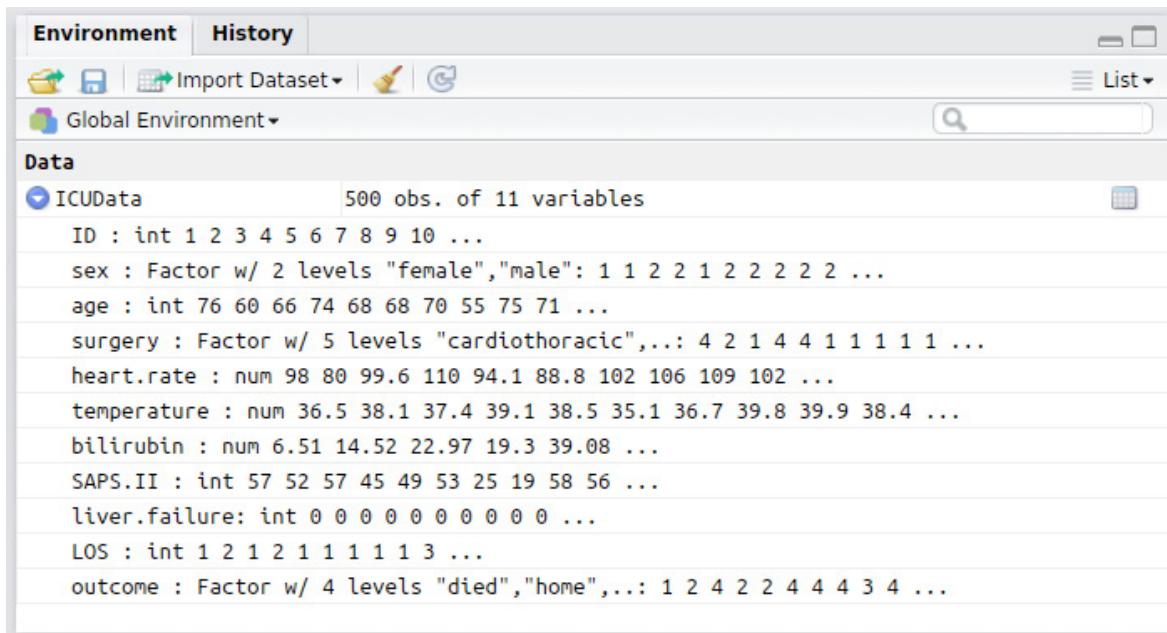


Figure 2.5: View of the exact structure of a dataset in RStudio.

heart.rate: maximum heart rate in beats per minute (`numeric` = real number) during the entire stay on the ICU.

temperature: maximum body temperature in °C (`numeric`) during the entire stay on the ICU.

bilirubin: maximum level of bilirubin in $\mu\text{mol/l}$ (`numeric`) during the entire stay on the ICU. The red dye of human blood is degraded and as an intermediate stage bilirubin emerges, a yellowish substance. Standard values are below 21 $\mu\text{mol/l}$ where higher values for instance may indicate liver problems (Wikipedia (2015b)).

SAPS.II: SAPS-II Score (`integer`) at admission to the ICU. The score reflects the physiological condition of a patient and is used to estimate the severity of disease. The higher the score the more severe is the disease. The range of values is from 0 to 163, where the values are associated with a probability of dying (Wikipedia (2015g)).

liver.failure: presence of liver failure (`integer`) where 0 and 1 indicate no and yes, respectively; that is, strictly speaking this is a nominal variable coded by numbers.

LOS: length of stay on the ICU in days (`integer`)

outcome: kind of discharge from the ICU (`Factor`). The possible levels are: died, home, other hospital, and secondary care/rehab.

Note:

The names of the variables `heart.rate`, `SAPS.II`, and `liver.failure` were changed during import. The respective column names include a blank and hence are no syntactically correct variable names. Such changes are done automatically during import. One can avoid it by setting the parameter `check.names`. The respective R code would be

```
1 ICUData <- read.csv(file = "ICUData.csv", check.names = FALSE)
```

However, `check.names = FALSE` should only be used after some experience in working with R, as it may lead to certain unwanted side effects and problems.

2.4 Categorical Variables

2.4.1 Univariate Analysis

First, we consider all variables separately (univariate) and start with nominal variables. That is, we analyze a single variable, whose levels are a set of possible names without any ordering. Examples are sex, blood group, rhesus factor, or also surgery, liver failure and outcome as in case of our ICU dataset (cf. Section 2.3).

Please first import the ICU dataset as described in Section 2.3, if you have not done it yet.

In case of nominal variables, descriptive statistics consists of calculating and visualizing absolute and relative frequencies. With the following R Code we compute the **absolute frequencies** of the kind of surgery the ICU patients obtained.

```
1 table(ICUData$surgery)
```

	cardiothoracic	gastrointestinal	neuro	other
	223	79	46	121
trauma				
	31			

The computation is done by function `table`. With symbol `$` we can access the variables of a dataset (`data.frame`). In this case, we access variable `surgery`, which includes the kind of surgery. We obtain the **relative frequencies** by dividing these numbers by the number of patients. This is also called the **empirical frequency distribution**. It is not recommended to use 500 here, even if it would be correct. It is better and more general to divide by the number of rows of the dataset, which can be obtained by function `nrow`.

```
1 table(ICUData$surgery)/nrow(ICUData)
```

cardiothoracic	gastrointestinal	neuro	other
0.446	0.158	0.092	0.242
trauma			
0.062			

That is, almost half of the patients underwent a cardiothoracic surgery. This most frequent level is also called **mode**. At second position, we have the other surgeries, followed by gastrointestinal surgeries. The smallest number of surgeries were caused by trauma, slightly more by neurological causes.

The graphical representation of relative and absolute frequencies is best done by **bar plots**. We first depict the absolute frequencies applying function **barplot**.

```
1 barplot(table(ICUData$surgery))
```

Excellent Economics and Business programmes at:


university of
groningen

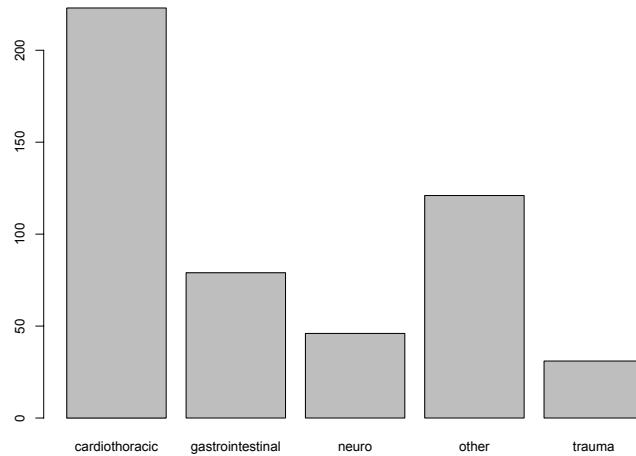



**“The perfect start
of a successful,
international career.”**

www.rug.nl/feb/education

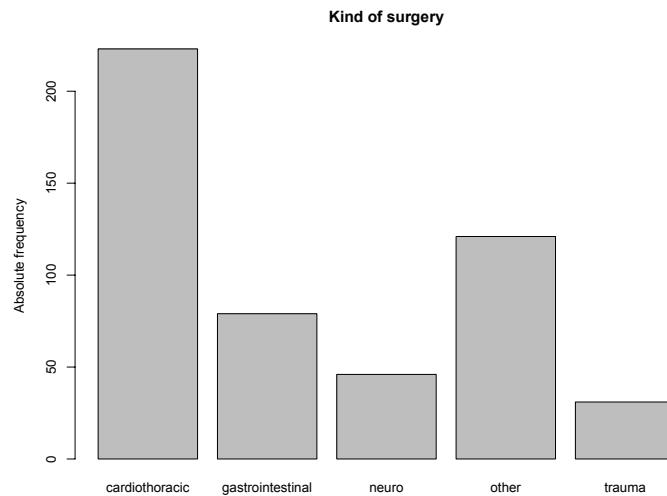
CLICK HERE
to discover why both socially
and academically the University
of Groningen is one of the best
places for a student to be





We add a title (argument `main`) and label the y axis (argument `ylab`) of the bar plot.

```
1 barplot(table(ICUData$surgery), main = "Kind of surgery",
2         ylab = "Absolute frequency")
```



There are many more arguments that can be used to further adapt the plot. We will get to know some more of them in the course of the book. Various examples of how to configure bar plots are also provided by the help page of `barplot`, which will be shown in *Window Help* of RStudio after running `?barplot`. Alternatively, one can search for help using the search field included in *window Help* of RStudio.

The most current version of RStudio (version 0.99.467, July 2015) also offers an interactive way of help. If you start writing code in an R script, the names of matching objects and, with some delay, matching help is shown; see Figure 2.6. By pressing the F1 key, the related help page opens in *window Help*.

A bar plot of the relative frequencies can be generated with a very similar R code as in case of the absolute frequencies. One just has to replace the absolute by relative frequencies. In addition to the standard graphics, there are other graphic systems implemented in R. Currently, the most frequently used system beside the standard system is probably the implementation of grammar of graphics in package "ggplot2" (Wickham (2009)). Thus, we use this system to display the relative frequencies. First of all, we have to install package "ggplot2". This can be done by running the following R code, where you need an active internet connection.

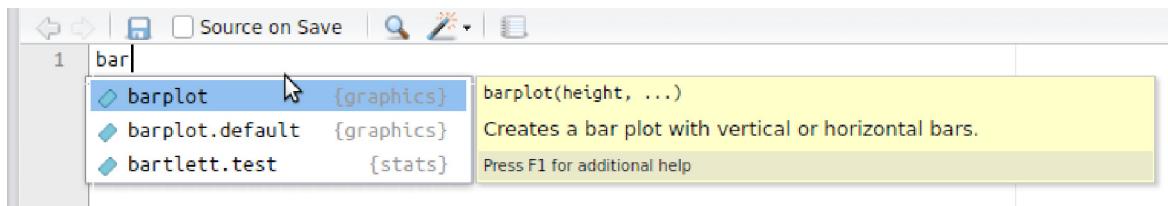


Figure 2.6: Interactive context based help in RStudio.

```
1 install.packages("ggplot2")
```

Alternatively, you can use the menu item *Install* in window *Packages* of RStudio, which opens a window for the installation; see Figure 2.7. You should only change the default settings in this window, if you are experienced in working with R. In particular, it is important to check *Install dependencies* as most of the R packages need other R packages to work properly. This option ensures that these additional packages are also installed.

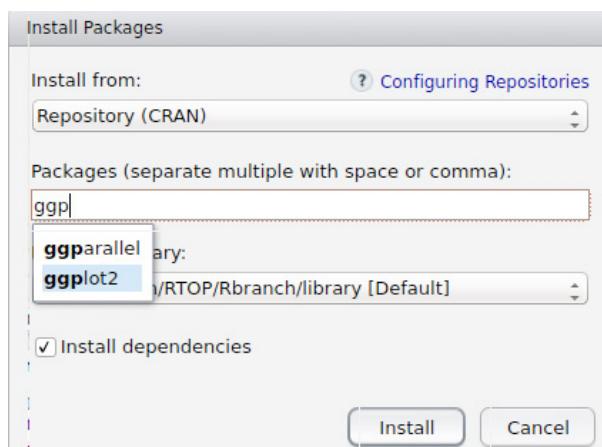


Figure 2.7: Installation of R packages in RStudio.

Note:

In case of the first installation of a contributed package, the installation might not start at once, but a windows opens, in which you have to specify a path to the library where the package shall be installed. It is recommended to use the given default setting of your operating system; that is, select and confirm this setting. A package must be installed only once and afterwards is steadily available for the user.

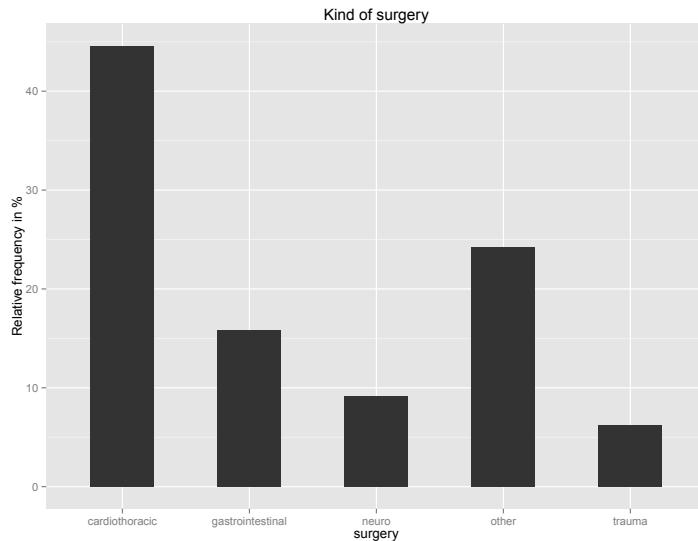
As explained in Section 1.2, there are several thousands of R packages. Thus, it makes sense that installed packages are not automatically loaded. Otherwise, your system would become more and more ponderous and slow with increasing number of installed packages. All packages except the base packages (see Section 1.2) must be explicitly loaded applying function `library`. We load package "ggplot2" (Wickham (2009)).

```
1 library(ggplot2)
```

We generate a bar plot of the relative frequencies using functions `ggplot` and `geom_bar`, where the width of the bars is reduced by argument `width`. With the help of function `aes` we can set the representation of the data. In the case at hand, we use the relative frequencies as percentages. Finally, the functions `ggtitle` and `ylab` are applied to add a title and label the y axis of the plot.

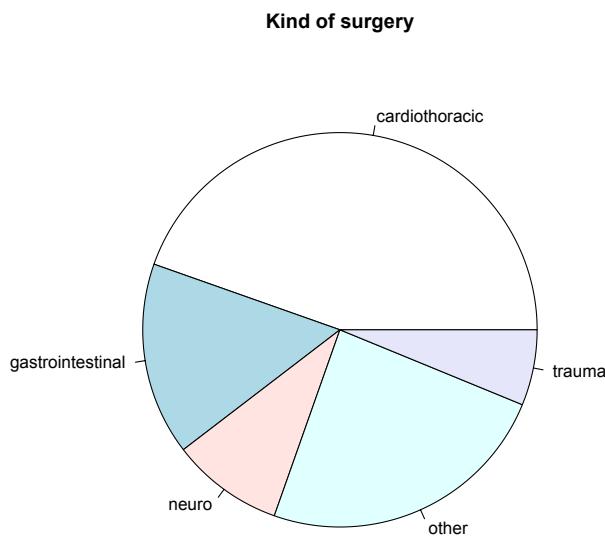
```
1 ## Assign data
2 ggplot(ICUData, aes(x=surgery)) +
3   ## Add bars with relative frequencies
4   geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
5   ## Title and label of y axis
6   ggtitle("Kind of surgery") + ylab("Relative frequency in %")
```

The advertisement features a city skyline at night with a prominent yellow speech bubble containing the text: "What if you could build your future and create the future?". The Alcatel-Lucent logo and website address "www.alcatel-lucent.com/careers" are visible at the top right. At the bottom, a quote reads: "One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



In practice, pie charts are frequently used instead of bar plots. Of course, this is also possible with R. The respective function is `pie`.

```
1 pie(table(ICUData$surgery), main = "Kind of surgery")
```



This kind of diagram has some drawbacks (see also Chapter 3). On the help page of `pie` you can read:

"Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data."

Thus, it is better to use a bar plot or dot chart to make the representation easier to read for the human eye.

Note:

The use of appropriate colors and diagrams is in more detail described in Chapter 3.

In the sequel, we additionally assume that the categories are ordered; that is, we consider ordinal variables. The ordering offers several additional ways for statistical analysis. In particular, quantiles are applicable for various purposes.

Definition 2.3 (Quantile). *Let $x_1, x_2, \dots, x_n \in \mathbb{R}$ ($n \in \mathbb{N}$) be some observations and let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ be the increasingly sorted observations. Then, the α -quantile for $\alpha \in (0, 1)$ is defined by*

$$q_\alpha = \begin{cases} x_{(\text{ceiling}(n\alpha))} & \text{if } n\alpha \notin \mathbb{Z} \\ [x_{(n\alpha)}, x_{(n\alpha+1)}] & \text{if } n\alpha \in \mathbb{Z} \end{cases} \quad (2.1)$$

The following remark includes some additional explanations about α -quantiles.

Remark 2.4.

- a) *If $n\alpha$ is no integer, the α -quantile corresponds to the ceiling ($n\alpha$)-th observation. Here “ceiling” means rounding to the next larger integer. In R there is function `ceiling`; e.g.*

```
1 ceiling(2.01)
```

```
[1] 3
```

```
1 ceiling(3.88)
```

```
[1] 4
```

If $n\alpha$ is integer, the α -quantile is not unique and all values in the bounded interval $[x_{(n\alpha)}, x_{(n\alpha+1)}]$ are valid α -quantiles. In practice, this is not satisfactory. Therefore, there is a number of proposals regarding the value of the interval that should be chosen as representative of the α -quantile. The most obvious approach probably is to use the midpoint of the interval. In R function `quantile` nine different approaches are implemented; see also Example 2.5.

- b) *Important special cases of quantiles are **percentiles** for $\alpha \in \{0.01, 0.02, \dots, 0.99, 1.00\}$, quartiles for $\alpha \in \{0.25, 0.50, 0.75\}$, and the **median** for $\alpha = 0.5$.*

Example 2.5. We consider the numbers $2, 4, 6, \dots, 20$ and want to compute the 20-th percentile, i.e. $\alpha = 0.2$. Hence, we get $n\alpha = 10 \cdot 0.2 = 2$. Therefore, the 20-th percentile is each number in the bounded interval $[x_{(2)}, x_{(3)}] = [4, 6]$. For performing this computation in R, we first have to enter the data. In the case at hand, the functions `c` (short for concatenate) or `seq` (short for sequence) can be used.

```
1 ## Concatenating numbers to a vector
2 x <- c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20)
3 ## Sequence: begin = 2, end = 20, distance = 2
4 x <- seq(from = 2, to = 20, by = 2)
```

In both cases the result is the vector `x` including the required numbers. We apply function `quantile` to the vector.

```
1 x
[1] 2 4 6 8 10 12 14 16 18 20
```


Maastricht University *Leading in Learning!*

**Join the best at
the Maastricht University
School of Business and
Economics!**



Top master's programmes

- 33rd place Financial Times worldwide ranking: MSc International Business
- 1st place: MSc International Business
- 1st place: MSc Financial Economics
- 2nd place: MSc Management of Learning
- 2nd place: MSc Economics
- 2nd place: MSc Econometrics and Operations Research
- 2nd place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

Maastricht
University is
the best specialist
university in the
Netherlands
(Elsevier)

Visit us and find out why we are the best!
Master's Open Day: 22 February 2014

www.mastersopenday.nl

```
1 ## R default
2 quantile(x, probs = 0.2)
```

```
20%
5.6
```

```
1 ## Type used by SAS software
2 quantile(x, type = 3, probs = 0.2)
```

```
20%
4
```

```
1 ## Type used by SPSS and Minitab software
2 quantile(x, type = 6, probs = 0.2)
```

```
20%
4.4
```

Note:

As Example 2.5 demonstrates, we must be aware that different software programs may give different results in case of quantiles.

We return to our ICU dataset. The medical score SAPS II is a typical example of an ordinal attribute. We first determine the median of the values via function `median`.

```
1 median(ICUData$SAPS.II)
```

```
[1] 42
```

```
1 ## also possible
2 quantile(ICUData$SAPS.II, probs = 0.5)
```

```
50%
42
```

That is, 50% of the patients have a SAPS II score ≤ 42 and 50% of the patients have a score ≥ 42 . The median is a so-called **location parameter** and does not give us any information about the variability of the values. For this purpose we can use quantiles, too. A very frequently used **scale** or **dispersion parameter** is the so-called **interquartile range** (IQR), the distance between third and first quartile (i.e. $q_{0.75} - q_{0.25}$). In R we can use function `IQR` to compute the IQR.

```
1 IQR(ICUData$SAPS.II)
```

```
[1] 26
```

Consequently, the middle 50% of our patients possess a range of 26 SAPS II points. Another option to evaluate the dispersion of the values is the **median absolute deviation** (MAD)

$$\text{MAD } (x_1, x_2, \dots, x_n) = \text{median} \{ |x_1 - M|, |x_2 - M|, \dots, |x_n - M| \} \quad (2.2)$$

where $M = \text{median} \{ x_1, x_2, \dots, x_n \}$. We obtain

```
1 M <- median(ICUData$SAPS.II)
2 median(abs(ICUData$SAPS.II-M))
```

```
[1] 13
```

Here, function `abs` computes the absolute deviations from the median. We can also use function `mad` to determine the MAD.

```
1 mad(ICUData$SAPS.II)
```

```
[1] 19.2738
```

Obviously, the result is different from our previous calculation. The reason for it is, that R as default applies the following definition

$$\text{MAD } (x_1, x_2, \dots, x_n) = 1:4826 \cdot \text{median} \{ |x_1 - M|, |x_2 - M|, \dots, |x_n - M| \} \quad (2.3)$$

By standardizing the MAD with 1:4826, the result under certain assumptions (normal distributed data) is comparable to the standard deviation, which will be introduced in Section 2.5. Function `mad` yields the unstandardized MAD by setting the standardizing constant (argument `constant`) to 1.

```
1 mad(ICUData$SAPS.II, constant = 1)
```

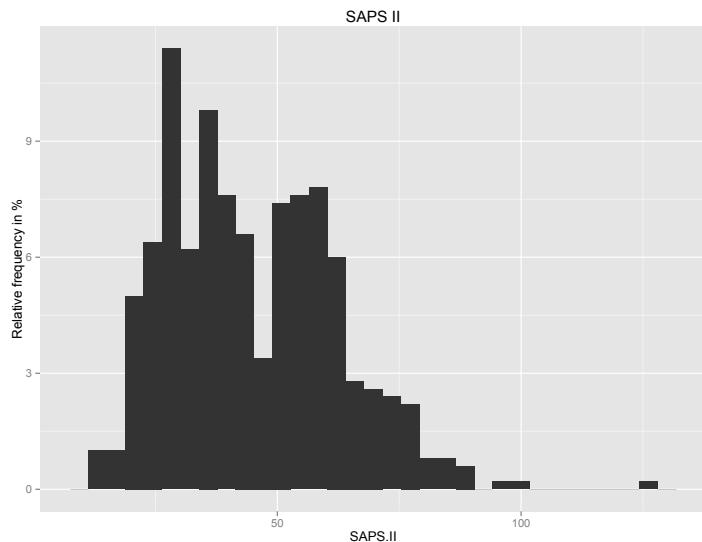
```
[1] 13
```

For depicting ordinal data we can again use bar plots.

```

1  ## Assign data
2  ggplot(ICUData, aes(x=SAPS.II)) +
3    ## Add bars
4    geom_bar(aes(y = 100*(..count..)/sum(..count..))) +
5    ## Title and label of y axis
6    ggtitle("SAPS II") + ylab("Relative frequency in %")

```



BI NORWEGIAN BUSINESS SCHOOL

Empowering People. Improving Business.

BI Norwegian Business School is one of Europe's largest business schools welcoming more than 20,000 students. Our programmes provide a stimulating and multi-cultural learning environment with an international outlook ultimately providing students with professional skills to meet the increasing needs of businesses.

BI offers four different two-year, full-time Master of Science (MSc) programmes that are taught entirely in English and have been designed to provide professional skills to meet the increasing need of businesses. The MSc programmes provide a stimulating and multi-cultural learning environment to give you the best platform to launch into your career.

- MSc in Business
- MSc in Financial Economics
- MSc in Strategic Marketing Management
- MSc in Leadership and Organisational Psychology

www.bi.edu/master

Quantiles are also the basis for one of the most important graphical display in descriptive statistics, the so-called **box-and-whisker plot**; see Figure 2.8. The box-and-whisker plot very well summarizes the information of median, IQR and range of the observations. In addition, it can be applied to identify suspicious observations (outliers).

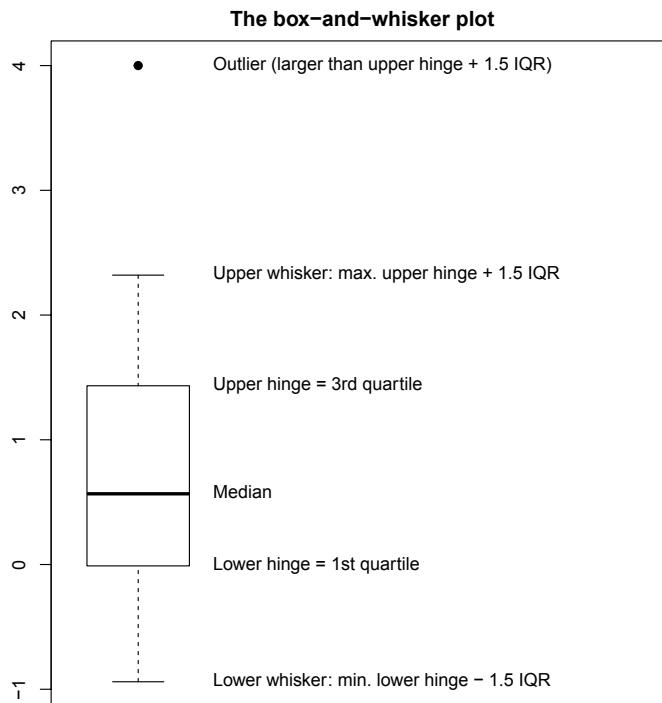
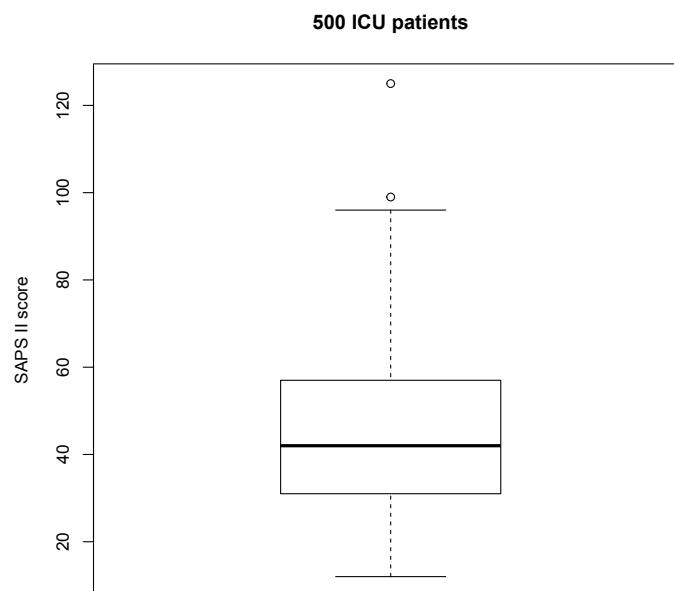


Figure 2.8: The values in a box-and-whisker plot.

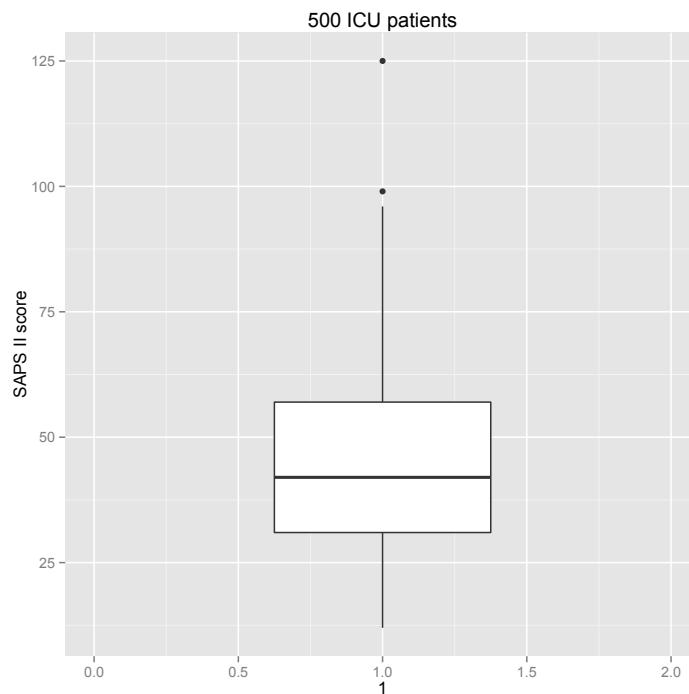
We generate a box-and-whisker plot of the SAPS II values using function `boxplot`.

```
1 boxplot(ICUData$SAPS.II, main = "500 ICU patients", ylab = "SAPS II score")
```



As we already know, the median is 42. The box of the box-and-whisker plot represents the middle 50% of the observations, which lie in the bounded interval [31, 57], whose length corresponds to the IQR, which is 26 points. Moreover, 25% of the values are smaller than 31 and accordingly, 25% of the values are larger than 57. Obviously, two patients were very severely sick with scores of 99 and 125 shown as outliers. Consequentially, the probability of surviving for these two patients was very small and hence, it is no surprise that both patients died. Nine of the ten patients with the highest SAPS II scores (≥ 83) died. We repeat the plot applying function `qplot` of package "ggplot2" (Wickham (2009)). This function is provided for generating standard plots as easy as possible.

```
1 ## Box-and-whisker plot at position x = 1
2 qplot(x = 1, y = SAPS.II, data = ICUData, geom = "boxplot",
3       xlim = c(0, 2), main = "500 ICU patients", ylab = "SAPS II score")
```



We use argument `xlim` to increase the limits of the x-axis, i.e. the box appears narrower. The limits are specified by a vector of length two, where the first coordinate corresponds to the starting point and the second coordinate to the endpoint.

Another interesting property of the α -quantile is its robustness against outliers. More precisely, up to $\alpha\%$ of the data for $\alpha \in (0, 0.5]$ and $1 - \alpha\%$ of the data for $\alpha \in [0.5, 1)$ may be outliers. This fact makes the median especially attractive as it possesses the maximum robustness.

Example 2.6. Wie again consider the sequence $2, 4, 6, \dots, 20$ and compute median and third quartile as well as 90% and 95% quantile.

```

1 x <- c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20)
2 quantile(x, probs = c(0.5, 0.75, 0.9, 0.95))

```

50%	75%	90%	95%
11.0	15.5	18.2	19.1

Now, we increase the largest number from 20 to 200, which corresponds to 10% outliers in the case at hand. We obtain

```

1 x <- c(2, 4, 6, 8, 10, 12, 14, 16, 18, 200)
2 quantile(x, probs = c(0.5, 0.75, 0.9, 0.95))

```

50%	75%	90%	95%
11.0	15.5	36.2	118.1

The 95% and also the 90% quantile are affected and are clearly increased. In contrast, median and third quartile show no change.

Another option to visualize the distribution of the data, is the so-called empirical cumulative distribution function.

Need help with your dissertation?

Get in-depth feedback & advice from experts in your topic area. Find out what you can do to improve the quality of your dissertation!

Get Help Now



Go to www.helpmyassignment.co.uk for more info

Helpmyassignment

Definition 2.7 (Empirical cumulative distribution function). Let $x_1, x_2, \dots, x_n \in \mathbb{R}$ ($n \in \mathbb{N}$) be some observations and let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ be the increasingly sorted observations. Furthermore, let $h_{(1)}, h_{(2)}, \dots, h_{(n)}$ be the associated relative frequencies. Then, the **empirical cumulative distribution function** is

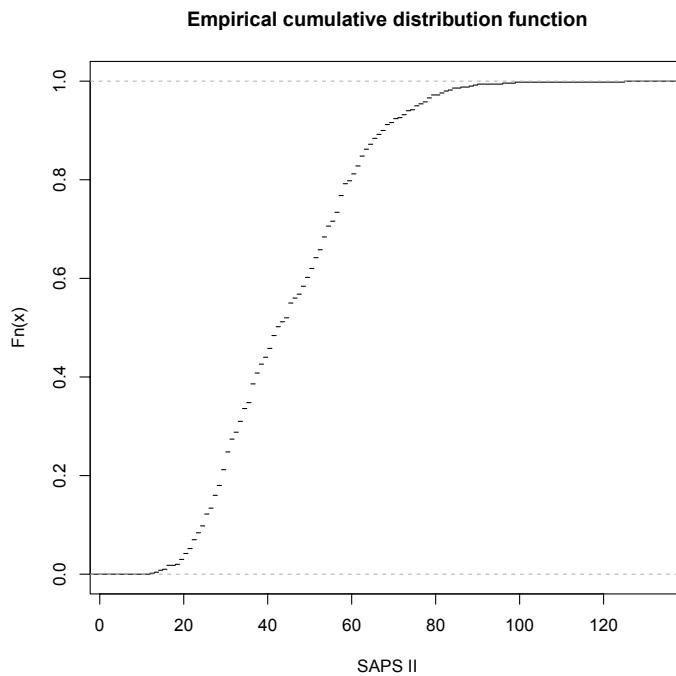
$$\hat{F}_n(x) = \begin{cases} 0 & \text{if } x < x_{(1)} \\ \sum_{i=1}^k h_{(i)} & \text{if } x_{(k)} \leq x < x_{(k+1)} \\ 1 & \text{if } x > x_{(n)} \end{cases} \quad (2.4)$$

The definition implies certain properties.

Remark 2.8. Looking at the definition, the empirical cumulative distribution function is a monotone increasing step function, which is continuous from above.

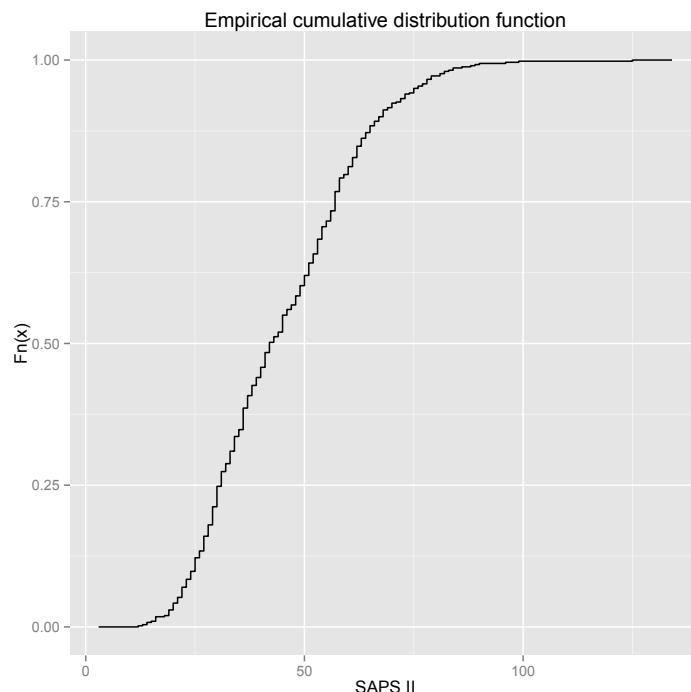
We use functions `ecdf` and `plot` to compute and plot the empirical cumulative distribution function of the SAPS II values.

```
1 plot(ecdf(ICUData$SAPS.II), xlab = "SAPS II", do.points = FALSE,
2      main = "Empirical cumulative distribution function")
```



Because of the quite large number of observations leading to a fine partition of the x-axis and many small jumps, we do not plot points (`do.points = FALSE`). The points can be used to illustrate that the function is continuous from above. We can generate a similar plot with function `qplot` of package "ggplot2" (Wickham (2009)).

```
1 qplot(ICUData$SAPS.II, stat = "ecdf", geom = "step", xlab = "SAPS II",
2       ylab = "Fn(x)", main = "Empirical cumulative distribution function")
```



2.4.2 Bivariate Analysis

So far we have analyzed the variables separately, but now we want to investigate the relationship between pairs of variables. We start with nominal variables. In this case, the analysis consists of calculating and plotting absolute or relative frequencies of all possible combinations of levels. This leads to a so-called **contingency table** or **cross table**. We analyse variables sex and surgery of the ICU dataset. We can compute the absolute frequencies of all level combinations with function `table`.

```
1 table(ICUData$sex, ICUData$surgery)
```

	cardiothoracic	gastrointestinal	neuro	other	trauma
female	61	31	19	57	7
male	162	48	27	64	24

The absolute numbers suggest that men undergo clearly more cardiothoracic surgeries than women. Since the dataset includes clearly more males than females, we should secure this hypothesis by additionally considering relative frequencies. We apply function `prop.table` to the cross table to compute relative frequencies. The argument `margin` controls if the relative frequencies are computed row- (`margin = 1`) or column-wise (`margin = 2`). In our example, we need the row-wise calculation.

```
1 prop.table(table(ICUData$sex, ICUData$surgery), margin = 1)
```

	cardiothoracic	gastrointestinal	neuro	other	trauma
female	0.34857143	0.17714286	0.10857143	0.32571429	0.04000000
male	0.49846154	0.14769231	0.08307692	0.19692308	0.07384615

For improving the representation, we use percentages and round the results via function `round` to one decimal place.

```
1 round(100*prop.table(table(ICUData$sex, ICUData$surgery), margin = 1), 1)
```

	cardiothoracic	gastrointestinal	neuro	other	trauma
female	34.9	17.7	10.9	32.6	4.0
male	49.8	14.8	8.3	19.7	7.4



Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations.

Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF



A collection of functions for descriptive statistics is included in package "DescTools" (Signorell et mult. al. (2015)), which we first have to install. You can either use the following R code

```
1 install.packages("DescTools")
```

or install it via window *Packages* of RStudio as described in the previous section. After installing, the package must first be loaded to get access to the included functions. For representing absolute and relative frequencies in a cross table we can use function `PercTable`.

```
1 library(DescTools)
2 PercTable(table(ICUData$sex, ICUData$surgery), rfrq = "010", pfmt = TRUE,
3 digits = 1)
```

	cardiothoracic	gastrointestinal	neuro	other	trauma
female	freq	61	31	19	57
	p.row	34.9%	17.7%	10.9%	32.6%
male	freq	162	48	27	64
	p.row	49.8%	14.8%	8.3%	19.7%
					7.4%

The computation of the relative frequencies is controlled by argument `rfrq`. A precise description of this argument is included in the help page of the function. By means of arguments `pfmt` and `digits` we generate percentages and round to one decimal place. The results confirm our first impression: males underwent a cardiothoracic surgery clearly more often than females. Conversely, females had remarkably more "other" surgeries.

The strength of the relationship of two (or more) nominal (or also ordinal) variables can be determined by so-called contingency coefficients.

Definition 2.9 (Contingency coefficients). *Let us assume $n \in \mathbb{N}$ observations of two variables with $l \in \mathbb{N}$ and $m \in \mathbb{N}$ levels, respectively. That is, the observed pairs of values can be represented by a matrix with l rows and m columns, where the total number of entries is $k = l \cdot m$. Furthermore, let n_i ($i = 1, \dots, k$) be the number of observations in cell i , p_i ($i = 1, \dots, k$) the theoretical probability of cell i , and hence $e_i = N \cdot p_i$ ($i = 1, \dots, k$) the expected number of observations in cell i . Then, the χ^2 -statistics is*

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i} \quad (2.5)$$

Based on χ^2 we get the following contingency coefficients

i. **ϕ -coefficient**

$$\phi = \sqrt{\frac{\chi^2}{n}} \quad (2.6)$$

ii. **Pearson's contingency coefficient**

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} \quad (2.7)$$

iii. **Cramér's V**

$$V = \sqrt{\frac{\chi^2}{n \cdot (M - 1)}} \quad M = \min\{l, m\} \quad (2.8)$$

We give some further explanations.

Remark 2.10.

- a) In practice, it is important to be aware of the maximum possible value of the computed contingency coefficient. Furthermore, a clear disadvantage of contingency coefficients is that they only measure the strength of a relationship, but are not able to identify the direction of a relationship, which for instance is of interest in case of ordinal attributes.
- b) The ϕ -coefficient attains values in the interval $[0; 1]$, where 1 is only possible under certain circumstances. If the result is 0, the two attributes are independent.
- c) The range of Pearson's contingency coefficient is $[0, \sqrt{\frac{M}{M-1}}]$ ($M = \min\{l, m\}$), where 0 indicates independence of the investigated attributes.
- d) Cramér's V attains values in the interval $[0; 1]$, where again 0 stands for independence. One speaks of weak dependence if $V \leq 0.3$, moderate dependence if $0.3 < V \leq 0.7$, and strong dependence if $V > 0.7$.

We apply functions `Phi`, `ContCoef`, and `CramerV` of package "DescTools" (Signorell et al. (2015)) to determine the strength of the relationship between sex and surgery.

```
1 ## phi coefficient
2 Phi(table(ICUData$sex, ICUData$surgery))
```

```
[1] 0.1846974
```

```
1 ## Pearson's contingency coefficient
2 ContCoef(table(ICUData$sex, ICUData$surgery))
```

```
[1] 0.1816254
```

```
1 ## Cramer's V
2 CramerV(table(ICUData$sex, ICUData$surgery))
```

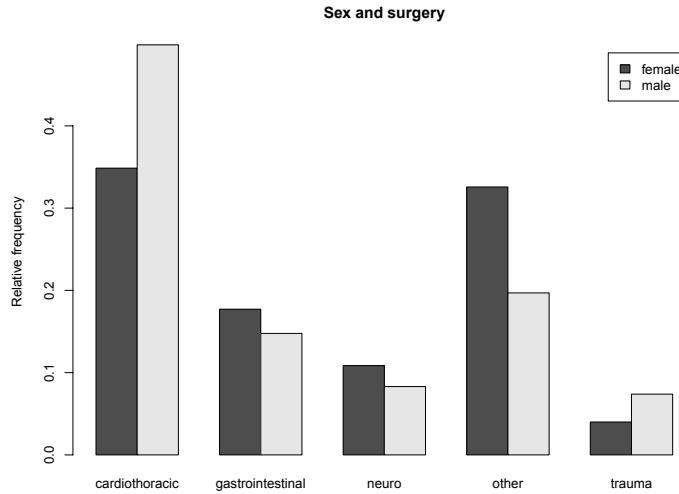
```
[1] 0.1846974
```

We obtain only a weak dependence between sex and surgery. As $M = 2$, the ϕ -coefficient and Cramér's V are identical.

Bar charts are the usual way to graphically represent contingency tables. We plot the variables sex and surgery, where we apply function `barplot` in combination with `table` and `prop.table`.

```
1 barplot(prop.table(table(ICUData$sex, ICUData$surgery)), margin = 1,
2         beside = TRUE, legend.text = TRUE, ylab = "Relative frequency",
3         main = "Sex and surgery")
```





The argument `beside = TRUE` guarantees that the bars of females and males are beside and not above each other. By `legend.text = TRUE` we obtain a legend explaining the relation between colors and sex.

In case of ordinal attributes, we can use rank correlations instead of contingency coefficients, which show not only the strength, but also the direction of a relationship. The **rank** of an observation corresponds to its position inside the sample after decreasingly sorting the observations; i.e., the largest observation has rank 1, the second largest rank 2, etc.

Definition 2.11 (Spearman's ρ). Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ($n \in \mathbb{N}$) be pairs of observations with ranks $(rx_1, ry_1), (rx_2, ry_2), \dots, (rx_n, ry_n)$. Then, **Spearman's ρ** is

$$\rho = \frac{\sum_{i=1}^n (rx_i - mr_x)(ry_i - mr_y)}{\sqrt{\sum_{i=1}^n (rx_i - mr_x)^2 \sum_{i=1}^n (ry_i - mr_y)^2}} \quad (2.9)$$

where mr_x and mr_y are the respective average ranks; i.e.

$$mr_x = \frac{1}{n} \sum_{i=1}^n rx_i \quad \text{and} \quad mr_y = \frac{1}{n} \sum_{i=1}^n ry_i \quad (2.10)$$

Spearman's ρ attains values in $[-1; 1]$, where 1 represents a perfect monotone increasing relation and -1 a perfect monotone decreasing relation.

We give some additional explanations.

Remark 2.12.

- a) If a value was observed several times (at least twice) this is called a **binding**. If there are no bindings, the computation of Spearman's ρ simplifies and it holds

$$\rho = 1 - \frac{6 \sum_{i=1}^n (rx_i - ry_i)^2}{n(n^2 - 1)} \quad (2.11)$$

- b) Beside Spearman's ρ , Kendall's τ is a frequently applied rank correlation coefficient. It compares the number of concordant and discordant pairs of observations. The result is in $[-1; 1]$. A value of 1 implies that both variables have exactly the same order, and -1 that they are in perfect inverse order. Kendall's τ is more appropriate than Spearman's ρ in case of small samples or scores with uneven scales.
- c) Rank correlations are also very useful in case of metric variables and can help to identify monotone relations.

We compute the correlation between SAPS II and length of stay (LOS), where we apply function `cor`.

```
1 ## Spearman's rho
2 cor(ICUData$SAPS.II, ICUData$LOS, method = "spearman")
```

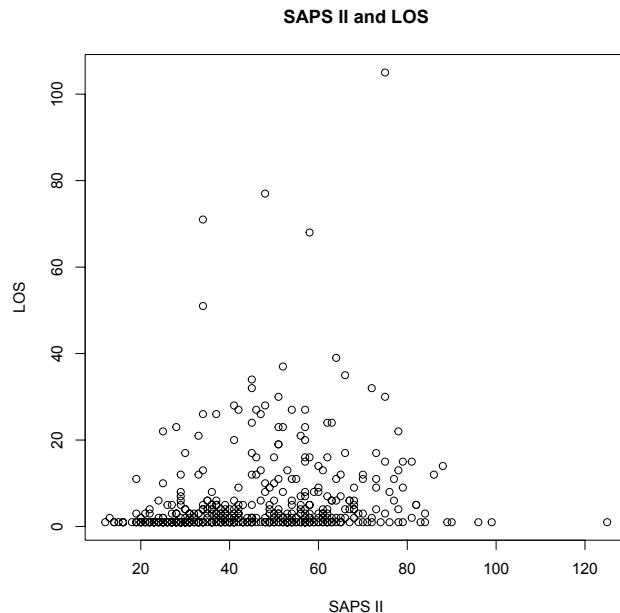
```
[1] 0.3379928
```

```
1 ## Kendall's tau
2 cor(ICUData$SAPS.II, ICUData$LOS, method = "kendall")
```

```
[1] 0.2518917
```

Expectedly, there is a positive relationship. Patients with a high SAPS II score are more severely ill and thus have to stay on the ICU for a longer time period. What works against this, is the fact that patients with a very high SAPS II value also have a high probability of dying, hence might die just after admission to ICU. We display the observed values in a **scatter plot** to check, if this is actually true. We apply function `plot`.

```
1 plot(ICUData$SAPS.II, ICUData$LOS, xlab = "SAPS II", ylab = "LOS",
2      main = "SAPS II and LOS")
```



Indeed, the patients with the highest SAPS II values have a small LOS and died quite rapidly. The ordinal or discrete structure of the attributes leads to an overlap of observations. We can use a so-called **alpha blending** to better visualize the structure of the point cloud; that is, the final color emerges from a combination of the original colors. We demonstrate this by means of package "ggplot2" (Wickham (2009)).

What do you want to do?

No matter what you want out of your future career, an employer with a broad range of operations in a load of countries will always be the ticket. Working within the Volvo Group means more than 100,000 friends and colleagues in more than 185 countries all over the world. We offer graduates great career opportunities – check out the Career section at our web site www.volvologroup.com. We look forward to getting to know you!

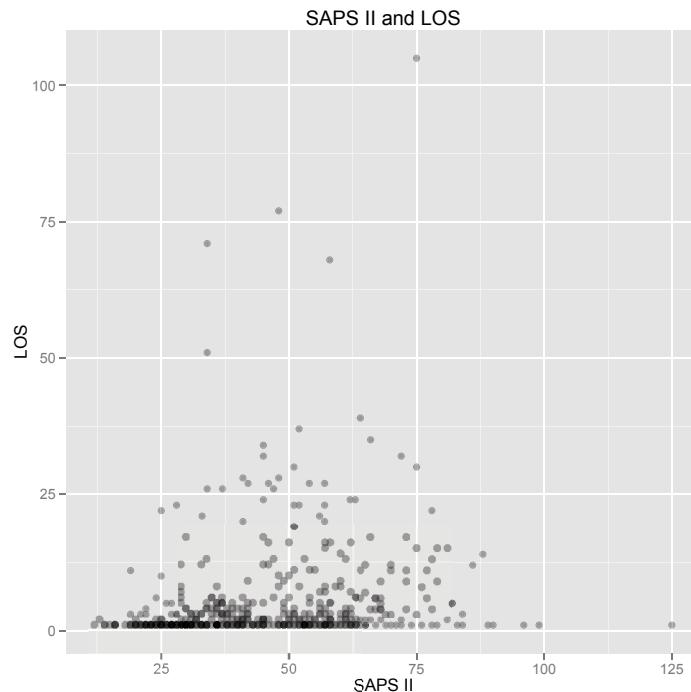
VOLVO
AB Volvo (publ)
www.volvologroup.com

VOLVO TRUCKS | RENAULT TRUCKS | MACK TRUCKS | VOLVO BUSES | VOLVO CONSTRUCTION EQUIPMENT | VOLVO PENTA | VOLVO AERO | VOLVO IT
VOLVO FINANCIAL SERVICES | VOLVO 3P | VOLVO POWERTRAIN | VOLVO PARTS | VOLVO TECHNOLOGY | VOLVO LOGISTICS | BUSINESS AREA ASIA

```

1 ggpplot(ICUData, aes(x=SAPS.II, y=LOS)) +
2   ## shape = 19: slightly larger point
3   ## alpha = 0.25: strength of blending
4   geom_point(shape=19, alpha=0.25) +
5   ## title and labels
6   ggtitle("SAPS II and LOS") + xlab("SAPS II") + ylab("LOS")

```



The darker the color the more observations overlap. In summary, we can assume a monotone increasing connection for a certain range of SAPS II scores but surely not for the full range. Therefore, the computed rank correlations should be interpreted with care.

Note:

Please, always reflect, if the results of your analysis make sense and be aware of the weaknesses of your statistical analysis. For instance, if there is no simple monotone relationship, the results of Spearman's ρ or Kendall's τ may be misleading.

2.5 Metric Variables

2.5.1 Univariate Analysis

As distances and even ratios are defined, further analyses are possible in case of metric variables. If not explicitly mentioned, the introduced analyses are possible for interval and ratio scaled variables. Probably the most frequently used statistics to describe data is the arithmetic mean.

Definition 2.13 (Arithmetic mean). Let $x_1, x_2, \dots, x_n \in \mathbb{R}$ ($n \in \mathbb{N}$) be some observations. Then, the **arithmetic mean** is

$$AM(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.12)$$

In this section, we again use our ICU dataset; see Section 2.3. We compute the arithmetic mean of the maximum body temperature during the stay on the ICU applying function `mean`.

```
1 mean(ICUData$temperature)
```

```
[1] 37.6632
```

That is, the arithmetic mean is only slightly above the normal range, where the result suggests a precision that is actually not true. The temperatures are only given with one decimal place. Consequentially, the arithmetic mean should be rounded to one decimal place. For this, we use function `round`.

```
1 round(mean(ICUData$temperature), 1)
```

```
[1] 37.7
```

It is advisable, to always compare the arithmetic mean with the median, as the median gives another description of the middle of the data and is very robust against outliers (see Example 2.6).

```
1 median(ICUData$temperature)
```

```
[1] 37.7
```

As median and arithmetic mean can be regarded as identical, it is likely, that the distribution of the maximum body temperature is quite symmetric around the arithmetic mean (resp. median). In addition, there are either no outliers or positive and negative outliers neutralize each other. We repeat the analysis using variable LOS (length of stay) given in days.

```
1 round(mean(ICUData$LOS), 1)
```

```
[1] 5.3
```

```
1 median(ICUData$LOS)
```

```
[1] 1
```

In this case, we see a clear difference between arithmetic mean and median. Either the distribution of LOS is skewed (more precisely right-skewed, see also Remark 2.23) or there are outliers pulling the arithmetic mean to the right. We will be able to distinguish these two cases below, where we consider diagrams of the data.

Another location parameter is the geometric mean, which is applied in case of relative changes. This measure of location is only meaningfully defined for strictly positive data.

Definition 2.14 (Geometric mean). *Let $x_1, x_2, \dots, x_n \in (0, \infty)$ ($n \in \mathbb{N}$) be some observations. Then, the geometric mean is*

$$GM(x_1, \dots, x_n) = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} \quad (2.13)$$

In the following remark, we describe an important connection between geometric and arithmetic mean.

Remark 2.15. *By applying the rules of logarithm we obtain*

$$\begin{aligned} AM(\log(x_1), \dots, \log(x_n)) &= \frac{1}{n} \sum_{i=1}^n \log(x_i) = \frac{1}{n} \log(x_1 \cdot x_2 \cdot \dots \cdot x_n) \\ &= \log(\sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}) \\ &= \log(GM(x_1, \dots, x_n)) \end{aligned} \quad (2.14)$$

That is, the arithmetic mean of the logarithmized observations is equal to the logarithm of the geometric mean where the base of logarithm is irrelevant. If we select the natural logarithm (\ln), we can rewrite it by applying the e -function to

$$GM(x_1, \dots, x_n) = e^{AM(\ln(x_1), \dots, \ln(x_n))} \quad (2.15)$$

If one observes processes following an exponential growth or decay, it is often easier to take the logarithm of the data and analyze the logarithmized observations. This is for instance true for the bilirubin measurements included in our ICU dataset. The base and recommended packages do not include the geometric mean, but we can apply function `Gmean` of package "DescTools" (Signorell et al. (2015)). We compute the natural logarithm of the geometric mean.

```
1 log(Gmean(ICUData$bilirubin))
```

[1] 2.847326

As our derivation in Remark 2.15 shows, the following R code must yield the same result, which is actually true.

```
1 mean(log(ICUData$bilirubin))
```

[1] 2.847326

Consequently, we may compute the geometric mean not only via function `Gmean`, but also by

```
1 exp(mean(log(ICUData$bilirubin)))
```

[1] 17.24162

where `exp` calculates the e -function. In addition, this form of computation has numerical advantages as summation is numerically more stable than calculating products. Therefore, the geometric mean is usually implemented in this way.

In practice, not only location but also dispersion of the observations is of interest. The probably most frequently applied measure of dispersion is the standard deviation, which is the square root of the variance.

Definition 2.16 (Variance, standard deviation). Let $x_1, x_2, \dots, x_n \in \mathbb{R}$ ($n \in \mathbb{N}$) be some observations. Then, the sample **variance** is

$$\text{Var}(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \text{AM}(x_1, \dots, x_n))^2 \quad (2.16)$$

and sample **standard deviation** reads

$$\text{SD}(x_1, \dots, x_n) = \sqrt{\text{Var}(x_1, \dots, x_n)} \quad (2.17)$$

We give some additional explanations.

Remark 2.17.

- a) Instead of $\frac{1}{n}$ one often uses $\frac{1}{n-1}$ for computing variance and standard deviation. This minor difference also makes the difference between descriptive and inferential statistics. With standarization $\frac{1}{n}$ we describe the sample, whereas with standardization $\frac{1}{n-1}$ we obtain an unbiased parameter estimate for the underlying population; for more details see Example 5.3. If the sample size n is not too small, we can neglect the difference in practice.
- b) Let us assume the observations were measured in unit U. Then, variance has unit U^2 and standard deviation unit U. This is one reason why standard deviation is more frequently applied in practice than variance.

We compute variance and standard deviation for the maximum body temperature. The respective functions in R are `var` and `sd` both using standardization $\frac{1}{n-1}$

```
1 var(ICUData$temperature)
```

```
[1] 3.011869
```

```
1 sd(ICUData$temperature)
```

```
[1] 1.735474
```

By multiplying the result with $\frac{n-1}{n}$, we obtain the “true” sample values.

```
1 n <- nrow(ICUData)
2 (n-1)/n*var(ICUData$temperature)
```

```
[1] 3.005846
```

```
1 (n-1) / n * sd(ICUData$temperature)
```

```
[1] 1.732003
```

Rounding to one decimal place, which should be done based on the given precision, would lead to identical results. Similarly to the comparison of arithmetic mean and median, we now compare standard deviation and the standardized MAD (cf. equation (2.3)).

```
1 sd(ICUData$temperature)
```

```
[1] 1.735474
```

```
1 mad(ICUData$temperature)
```

```
[1] 1.18608
```

There is a clear difference between both statistics. Either the temperature distribution can not be described by a distribution that is symmetric around the arithmetic mean or there are outliers distorting the standard deviation. We will identify the cause below.



In case of positive measurements, one in practice often uses the following standardized dispersion measure.

Definition 2.18 (Coefficient of variation). *Let $x_1, x_2, \dots, x_n \in [0, \infty)$ ($n \in \mathbb{N}$) be some positive observations. Then, the coefficient of variation is*

$$CV(x_1, \dots, x_n) = \frac{SD(x_1, \dots, x_n)}{AM(x_1, \dots, x_n)} \quad (2.18)$$

We give some additional explanations.

Remark 2.19.

- a) *The coefficient of variation is a dimensionless quantity, which is frequently given in percent; that is, percental dispersion with reference to the arithmetic mean. Consequentially, it should only be applied to ratio scaled variables.*
- b) *There are variants of the coefficient of variation based on quantiles. One option is based on median and MAD*

$$medCV(x_1, \dots, x_n) = \frac{MAD(x_1, \dots, x_n)}{\text{median}(x_1, \dots, x_n)} \quad (2.19)$$

Alternatively, one can use quartiles leading to the so-called **quartile coefficient of dispersion**

$$QCD(x_1, \dots, x_n) = \frac{IQR(x_1, \dots, x_n)}{\text{median}(x_1, \dots, x_n)} \quad (2.20)$$

We apply these standardized dispersion measures to the maximum body temperature.

```
1 sd(ICUData$temperature) / mean(ICUData$temperature)
```

```
[1] 0.04607877
```

```
1 mad(ICUData$temperature) / median(ICUData$temperature)
```

```
[1] 0.03146101
```

```
1 IQR(ICUData$temperature) / median(ICUData$temperature)
```

```
[1] 0.0397878
```

We get only minor variations around the arithmetic mean respectively, median in the range of about 3–5%.

In the following definition we give the standard deviation for the geometric mean.

Definition 2.20 (Geometric standard deviation). *Let $x_1, x_2, \dots, x_n \in (0, \infty)$ ($n \in \mathbb{N}$) be some positive observations. Then, the **geometric standard deviation** is*

$$SD_{GM}(x_1, \dots, x_n) = e^{\sqrt{\frac{1}{n} \sum_{i=1}^n (\ln(x_i) - \ln(GM(x_1, \dots, x_n)))^2}} \quad (2.21)$$

We briefly motivate this definition.

Remark 2.21. *It holds*

$$SD(\ln(x_1), \dots, \ln(x_n)) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\ln(x_i) - AM(\ln(x_1), \dots, \ln(x_n)))^2} \quad (2.22)$$

By using the connection (2.14) and by analogously introducing the geometric standard deviation, we get

$$\begin{aligned} SD(\ln(x_1), \dots, \ln(x_n)) &= \sqrt{\frac{1}{n} \sum_{i=1}^n (\ln(x_i) - \ln(GM(x_1, \dots, x_n)))^2} \\ &= \ln(SD_{GM}(x_1, \dots, x_n)) \end{aligned} \quad (2.23)$$

By applying the e -function, Definition 2.20 follows. Furthermore, the expression below the sigma sign may be rewritten as

$$\ln(x_i) - \ln(GM(x_1, \dots, x_n)) = \ln\left(\frac{x_i}{GM(x_1, \dots, x_n)}\right) \quad (2.24)$$

We check equation (2.23) using the bilirubin values of our ICU dataset, where we use function `Gsd` of package "`DescTools`" (Signorell et mult. al. (2015)) for computing the geometric standard deviation.

```
1 log(Gsd(ICUData$bilirubin))
```

```
[1] 0.7238379
```

```
1 sd(log(ICUData$bilirubin))
```

```
[1] 0.7238379
```

In addition to location and scale measures, shape measures are used in case of metric variables. A shape measure of symmetry is skewness.

Definition 2.22 (Skewness). Let $x_1, x_2, \dots, x_n \in \mathbb{R}$ ($n \in \mathbb{N}$) be some observations. Then, the **skewness** is

$$Skew(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - AM(x_1, \dots, x_n)}{SD(x_1, \dots, x_n)} \right)^3 \quad (2.25)$$

If $Skew(x_1, \dots, x_n) < 0$, the data distribution is **left-skewed**, if $Skew(x_1, \dots, x_n) > 0$, it is **right-skewed**.

We give some additional explanations.

Remark 2.23.

- a) By centering the date with respect to the arithmetic mean and standardizing it by the standard deviation, which is also called **z-transformation**, one gets the so-called **z-score**, a dimensionless score. As skewness is defined based on the z-score, it is also a dimensionless measure.
- b) The skewness of a distribution can also be identified by using arithmetic mean and median. If $AM(x_1, \dots, x_n) < \text{median}(x_1, \dots, x_n)$, the distribution is left-skewed. Conversely, if $AM(x_1, \dots, x_n) > \text{median}(x_1, \dots, x_n)$ the distribution is right-skewed; see also Figure 2.9.



The advertisement features a large, illuminated geometric structure composed of numerous red and orange light panels forming a complex lattice pattern against a dark background. A small URL, "www.sylvania.com", is visible in the top left corner of the main image area.

We do not reinvent
the wheel we reinvent
light.

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM

OSRAM
SYLVANIA



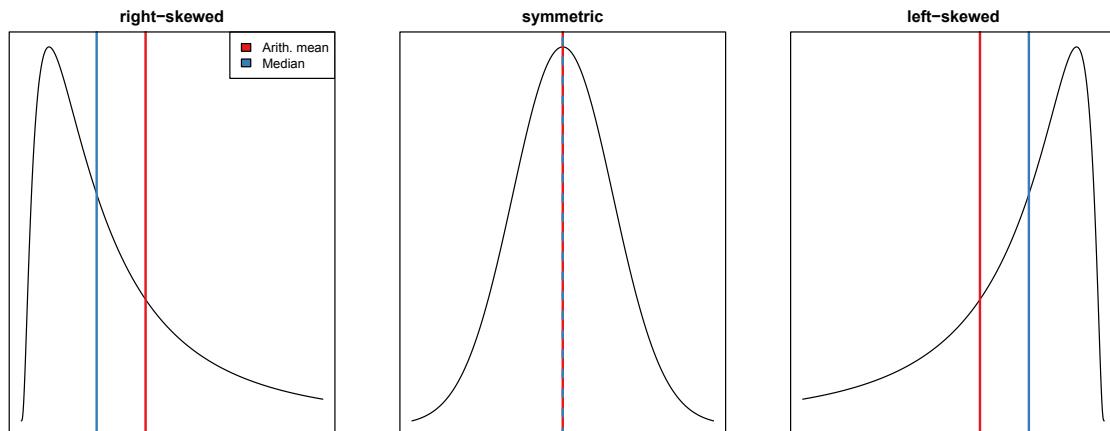


Figure 2.9: Examples of skewness.

We compute the skewness of the maximum body temperature by applying function `Skew` of package "DescTools" (Signorell et mult. al. (2015)).

```
1 Skew(ICUData$temperature)
```

```
[1] -8.77457
```

The result, which indicates a left-skewed distribution, contradicts our observation above, where median and arithmetic mean were (more or less) identical giving evidence for a symmetric distribution. A closer look at the measured temperatures shows that patient 398 had an abnormally low maximum (!) body temperature of 9.1°C (measurement- or transcription error?). We repeat the computation without patient 398. For accessing the maximum body temperature of patient 398, we can use square brackets [and his index.

```
1 ## Patient 398
2 ICUData$temperature[398]
```

```
[1] 9.1
```

A negative index means that this index is omitted. We obtain

```
1 Skew(ICUData$temperature[-398])
```

```
[1] 0.3142909
```

Now, the skewness is very small and confirms our first impression. The distribution of the values, without patient 398, is quite symmetric around the arithmetic mean. Furthermore, omitting patient 398 also clearly reduces the standard deviation

```
1 sd(ICUData$temperature[-398])
[1] 1.173187
```

which is now very close to the standardized MAD.

Note:

Single outliers may have a strong influence on certain statistical procedures and may clearly distort the results. Examples are arithmetic mean, variance/standard deviation, and skewness. Therefore, it is important to always investigate the data with respect to suspicious values.

We compute the skewness for length of stay (LOS). Based on arithmetic mean and median, we concluded above that the distribution must be right-skewed. Thus, we would expect a positive value of skewness.

```
1 Skew(ICUData$LOS)
```

```
[1] 4.880826
```

Indeed, the result confirms our first analysis.

Another shape measure is the kurtosis.

Definition 2.24 (Kurtosis). Let $x_1, x_2, \dots, x_n \in \mathbb{R}$ ($n \in \mathbb{N}$) be some observations. Then, the **kurtosis** is

$$Kurt(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - AM(x_1, \dots, x_n)}{SD(x_1, \dots, x_n)} \right)^4 - 3 \quad (2.26)$$

If $Kurt(x_1, \dots, x_n) < 0$, the data distribution is **platykurtic**, if $Kurt(x_1, \dots, x_n) > 0$, it is **leptokurtic**.

We give some additional explanations.

Remark 2.25. The reference for defining the kurtosis is the normal distribution (see Section 4.2). By subtracting 3 in the above definition, the normal distribution has kurtosis 0. If we observe a negative kurtosis, the distribution is flatter and less curved than the normal distribution. If the kurtosis is positive, the distribution is steeper and more curved than the normal distribution; see also Figure 2.10.

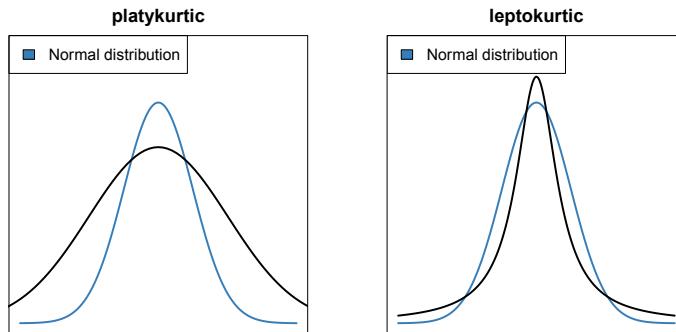


Figure 2.10: Examples of kurtosis.

We compute the kurtosis of the maximum body temperature of the ICU patients using function `Kurt` of package "DescTools" (Signorell et mult. al. (2015)). Due to the strong impact of patient 398 on skewness, we compare the kurtosis with and without this patient.

```
1 Kurt(ICUData$temperature)
```

```
[1] 144.4649
```

```
1 Kurt(ICUData$temperature[-398])
```

```
[1] 0.3431707
```

Once again, we see how large the influence of one single observation may be. We conclude that the distribution is not extremely leptokurtic, but except for one observation can be quite well described by a normal distribution. We determine the kurtosis of length of stay (LOS).

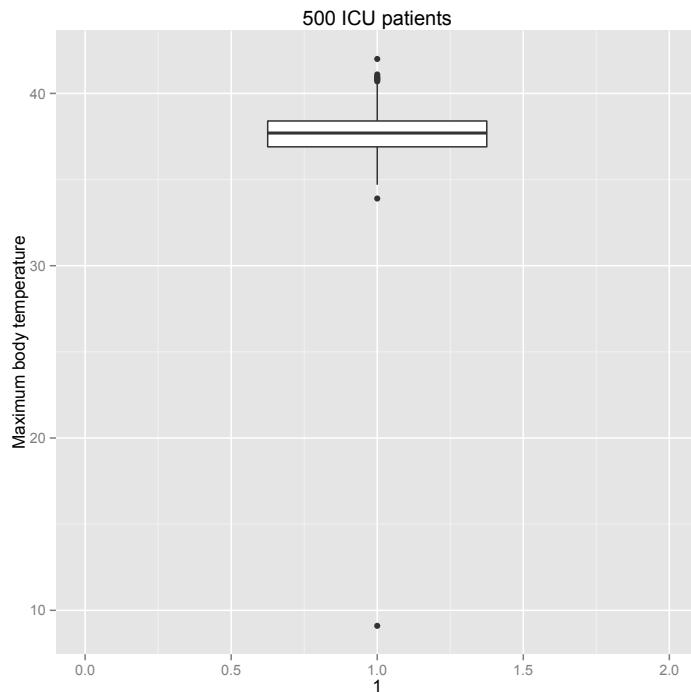
```
1 Kurt(ICUData$LOS)
```

```
[1] 33.59482
```

That is, the distribution of LOS is leptokurtic.

We proceed with various options for plotting metric variables. We start with a box-and-whisker plot of the maximum body temperature.

```
1 ## Box-and-whisker plot at position x = 1
2 qplot(x = 1, y = temperature, data = ICUData, geom = "boxplot",
3        xlim = c(0, 2), main = "500 ICU patients",
4        ylab = "Maximum body temperature")
```



We see some minor outliers and the value of patient 398, which extremely differs from all other observations.

CHALLENGING PERSPECTIVES

Internship opportunities

EADS unites a leading aircraft manufacturer, the world's largest helicopter supplier, a global leader in space programmes and a worldwide leader in global security solutions and systems to form Europe's largest defence and aerospace group. More than 140,000 people work at Airbus, Astrium, Cassidian and Eurocopter, in 90 locations globally, to deliver some of the industry's most exciting projects.

An **EADS internship** offers the chance to use your theoretical knowledge and apply it first-hand to real situations and assignments during your studies. Given a high level of responsibility, plenty of learning and development opportunities, and all the support you need, you will tackle interesting challenges on state-of-the-art products.

We welcome more than 5,000 interns every year across disciplines ranging from engineering, IT, procurement and finance, to strategy, customer support, marketing and sales. Positions are available in France, Germany, Spain and the UK.

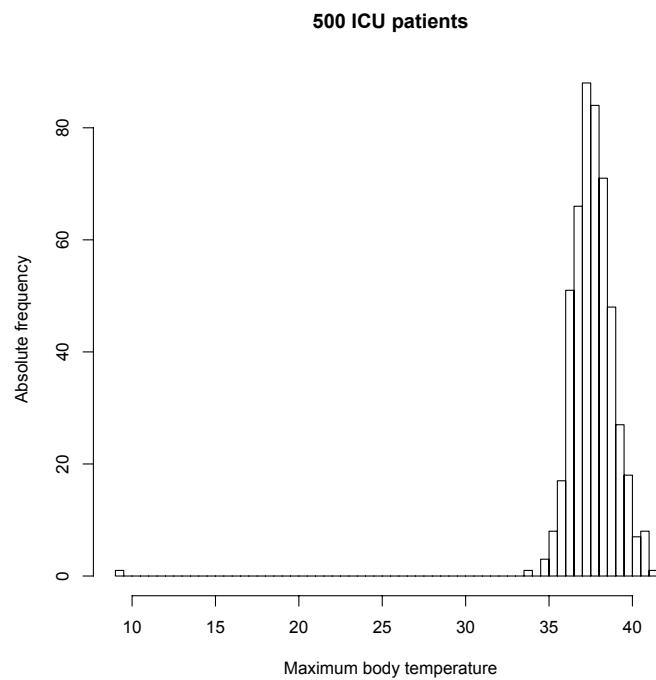
To find out more and apply, visit www.jobs.eads.com. You can also find out more on our **EADS Careers Facebook page**.

AIRBUS **ASTRIUM** **CASSIDIAN** **EUROCOPTER**

EADS

We further analyse the distribution using histograms. A **histogram** is a special kind of bar chart, that is obtained by splitting the range of a metric variable in consecutive intervals. For each interval the absolute or relative frequency of the included observations is visualized by a bar. For choosing the intervals, there are some rules of thumb, which are used by software programs to automatically select a number of equal length intervals. However, in most cases it is better to select the intervals by hand and choose a division that fits to the context. We generate a histogram of the maximum body temperatures, where we use intervals of length 0.5°C . We can specify the intervals by argument `breaks`.

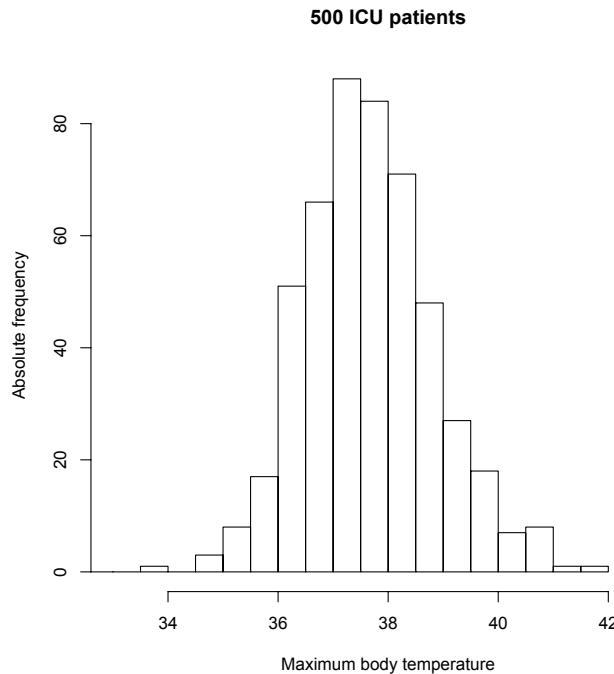
```
1 hist(ICUData$temperature, breaks = seq(from = 9.0, to = 42, by = 0.5),
2      main = "500 ICU patients", xlab = "Maximum body temperature",
3      ylab = "Absolute frequency")
```



Again, we clearly see the extreme value of patient 398.

To get a better view of the distribution, we can either remove the value of patient 398 or restrict the range of the x-axis by argument `xlim`. We select the second option.

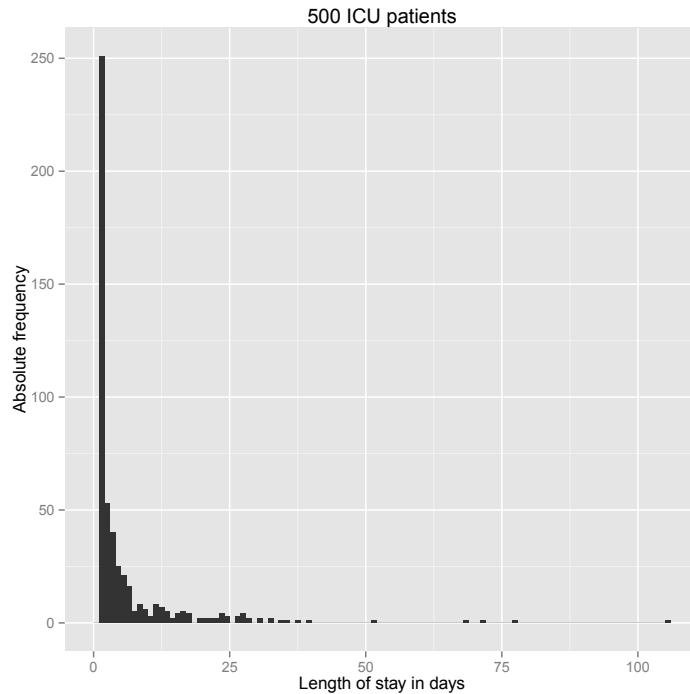
```
1 hist(ICUData$temperature, breaks = seq(from = 9.0, to = 42, by = 0.5),
2      main = "500 ICU patients", xlab = "Maximum body temperature",
3      ylab = "Absolute frequency", xlim = c(33,43))
```



The plot confirms our previous computations; i.e., the distribution is quite symmetric around the arithmetic mean and the distribution of the maximum body temperature in the ICU population (except for patients with strong undercooling/hypothermia) is probably well described by a normal distribution.

Next, we take a look on length of stay, where we use function `qplot` of package "ggplot2" (Wickham (2009)) to generate a histogram. As length of the intervals we use one day, which we can specify by argument `binwidth`.

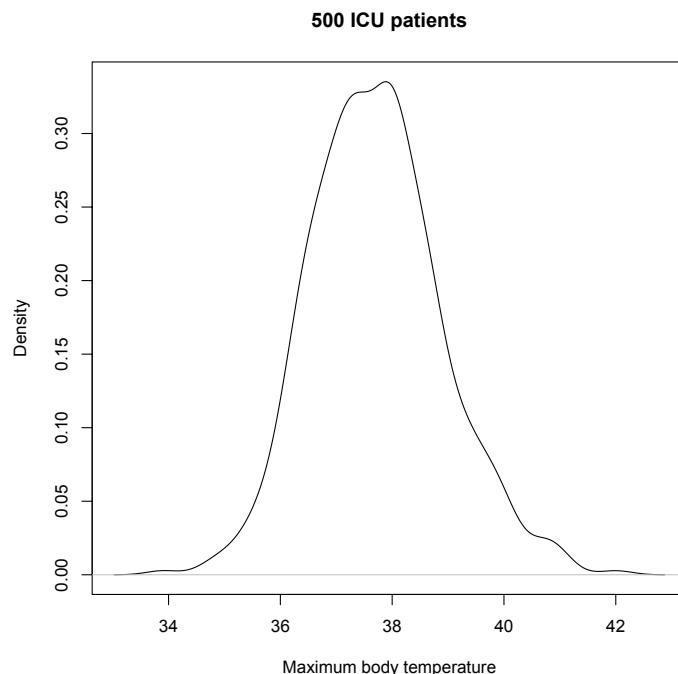
```
1 qplot(LOS, data = ICUData, geom = "histogram", binwidth = 1,
2       xlab = "Length of stay in days", ylab = "Absolute frequency",
3       main = "500 ICU patients")
```



The figure confirms our previous computations. We get a clearly right-skewed and quite spiky distribution. The majority of patients had a LOS of only a few days. The maximum LOS was 105 days.

Alternatively, we can visualize the distribution of the observed values by means of their estimated density. The empirical **density** may be regarded as a smoothed version of a histogram. In R we can apply function `density` to compute the density (more precisely: the kernel density estimation). The result can be visualized via function `plot`. We consider the maximum body temperature and omit patient 398.

```
1 plot(density(ICUData$temperature[-398]), xlab = "Maximum body temperature",
2      ylab = "Density", main = "500 ICU patients")
```



We get a density that is quite symmetric around the arithmetic mean.

The advertisement features a circular water splash against a white background. Inside the circle, the text "360° thinking." is written in blue, with "thinking." in green. Below the circle, the Deloitte logo is displayed in blue, followed by the tagline "Discover the truth at www.deloitte.ca/careers". A small copyright notice at the bottom right reads "© Deloitte & Touche LLP and affiliated entities."

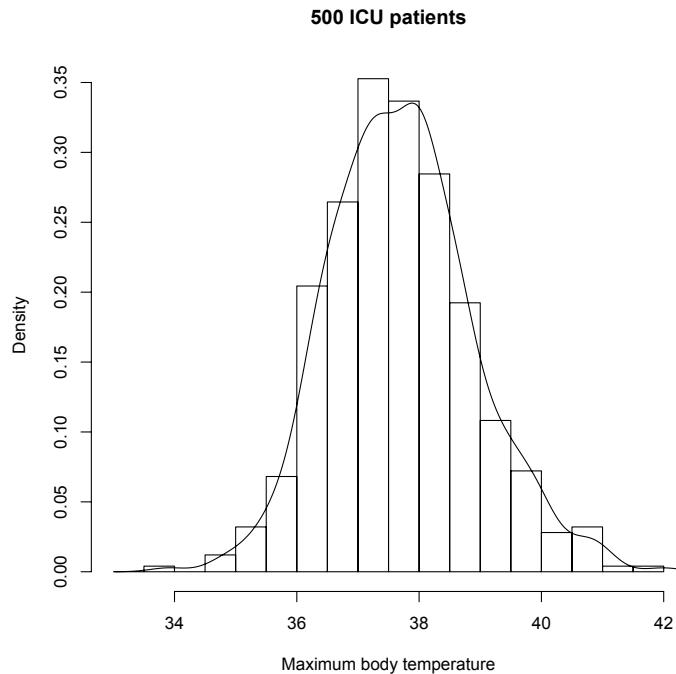
Deloitte.

Discover the truth at www.deloitte.ca/careers

© Deloitte & Touche LLP and affiliated entities.

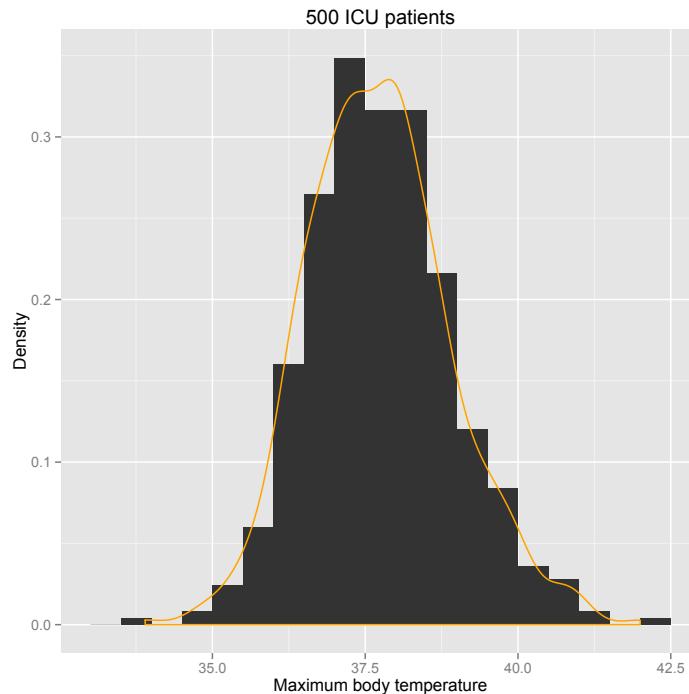
If we want to display histogram and density together, we must use argument `freq = FALSE` in the call of function `hist`. With this setting the density scale is used for plotting the histogram. The function `lines` adds a line to an already existing plot and can be used to add the estimated density to the histogram.

```
1 hist(ICUData$temperature[-398], breaks = seq(from = 33, to = 42, by = 0.5),
2       xlab = "Maximum body temperature", ylab = "Density", freq = FALSE,
3       main = "500 ICU patients")
4 lines(density(ICUData$temperature[-398]))
```



The density curve adapts well to the histogram. We use function `ggplot` in combination with functions `geom_histogram` and `geom_density` to generate a similar plot with package "ggplot2" (Wickham (2009)). With functions `ggtitle`, `xlab` and `ylab` we add a title and label x and y axis.

```
1 ggplot(ICUData[-398,], aes(x=temperature)) +
2   geom_histogram(aes(y=..density..), binwidth = 0.5) +
3   geom_density(color = "orange") + ylab("Density") +
4   xlab("Maximum body temperature") +
5   ggtitle("500 ICU patients")
```

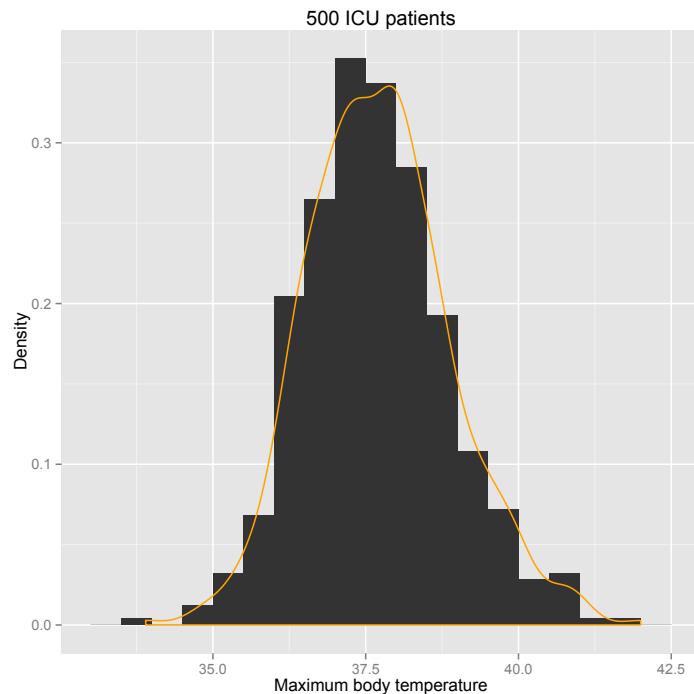


The estimated densities are very similar or even identical in both figures, however the histograms differ. That happens, because in case of `geom_histogram` one considers intervals that are open to the righthand side and closed to the left-hand side, whereas in case of `hist` it is the other way round, i.e. open to the left and closed to the right. We can achieve this, by additionally setting `right = TRUE` in function `geom_histogram`.

```

1 ggplot(ICUData[-398,], aes(x=temperature)) +
2   geom_histogram(aes(y=..density..), binwidth = 0.5, right = TRUE) +
3   geom_density(color = "orange") + ylab("Density") +
4   xlab("Maximum body temperature") +
5   ggtitle("500 ICU patients")

```



Now, both figures present identical results.

SIMPLY CLEVER

ŠKODA



We will turn your CV into
an opportunity of a lifetime



Do you like cars? Would you like to be a part of a successful brand?
We will appreciate and reward both your enthusiasm and talent.
Send us your CV. You will be surprised where it can take you.

Send us your CV on
www.employerforlife.com

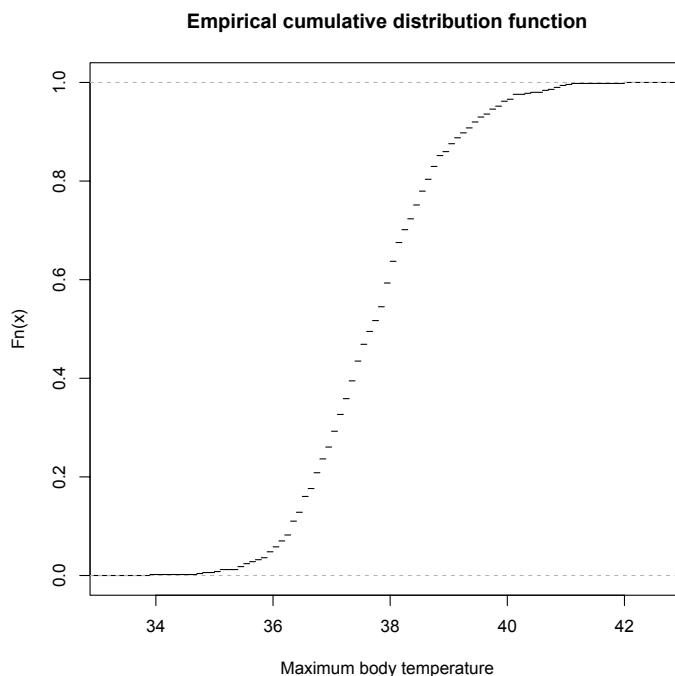


Note:

In case of `ggplot`, we not only omit the temperature of patient 398 but by `[-398,]` remove all data of patient 398. More precisely, we remove row 398 from the dataset.

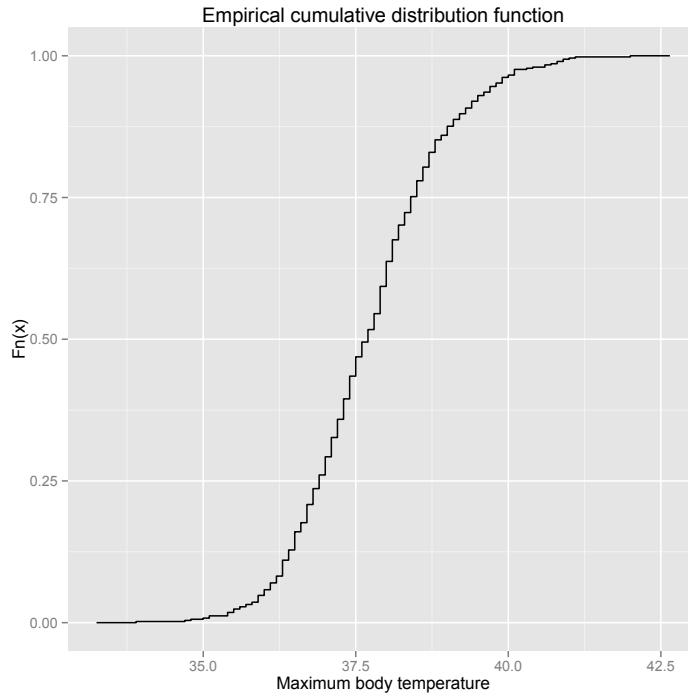
We may also visualize the distribution of the maximum body temperature by means of the empirical cumulative distribution function (cf. Definition 2.7). We first apply functions `ecdf` and `plot` where we again omit patient 398.

```
1 plot(ecdf(ICUData$temperature[-398]), xlab = "Maximum body temperature",
2      main = "Empirical cumulative distribution function", do.points = FALSE)
```



Because of the large number of small jumps, we do not plot points (i.e., `do.points = FALSE`). We can also generate an analogous figure by means of function `qplot` of package "ggplot2" (Wickham (2009)).

```
1 plot(ecdf(ICUData$temperature[-398]), xlab = "Maximum body temperature",
2      main = "Empirical cumulative distribution function", do.points = FALSE)
```



Another important application of these way to display the empirical distribution of the data, is to compare it with the distribution of an assumed probability model. In doing so, a graphical validation of an assumed model is possible. We will investigate this in more detail in Chapter 5.

2.5.2 Bivariate Analysis

The strength and direction of the relationship between metric variables can be described by means of correlation, similar to the case of ordinal data (see Section 2.4.2). Beside rank correlations one can use the Pearson correlation.

Definition 2.26 (Pearson correlation). *Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \in \mathbb{R}^2$ be some pairs of observations. Then, the **Pearson (product-moment) correlation (coefficient)** is*

$$r = \frac{\sum_{i=1}^n (x_i - AM(x_1, \dots, x_n))(y_i - AM(y_1, \dots, y_n))}{\sqrt{\sum_{i=1}^n (x_i - AM(x_1, \dots, x_n))^2 \sum_{i=1}^n (y_i - AM(y_1, \dots, y_n))^2}} \quad (2.27)$$

The Pearson correlation may attain values in $[-1; 1]$ where 1 represents a perfect positive linear relation and -1 a perfect negative linear relation.

We give some additional explanations.

Remark 2.27.

- a) The assumption that there is a linear relation and hence, the Pearson correlation is appropriate to describe the strength of the relationship is a rather strong assumption. Rank correlations are more flexible, as they can be used to describe monotone relationships.
- b) A closer look at the equations shows that Spearman's ρ (cf. Definition 2.11) is nothing else but the Pearson correlation of the ranks.
- c) By adding $\frac{1}{n}$ to the numerator of the Pearson correlation, the numerator is identical to the sample covariance of the two variables. By expanding the denominator analogously, it becomes the product of the two standard deviations. That is, the Pearson correlation can be regarded as a normalized covariance.

We investigate the connection between maximum body temperature and maximum heart rate.

```
1 cor(ICUData$temperature , ICUData$heart.rate)
```

```
[1] 0.1763067
```

There is a weak positive relation; i.e., with increasing body temperature also the heart rate tends to increase.



The advertisement features a photograph of two young women with long hair, smiling and peeking from behind a red door or wall. The woman on the left is wearing a black leather jacket over a white top. The woman on the right is wearing a dark blue blazer over a purple top. To the left of the photo, there is promotional text and the university's logo.

Sweden Sverige

Linköping University – innovative, highly ranked, European

Interested in Computer Science? Kick-start your career with an English-taught master's degree.

→ [Click here!](#)

L.U LINKÖPING UNIVERSITY

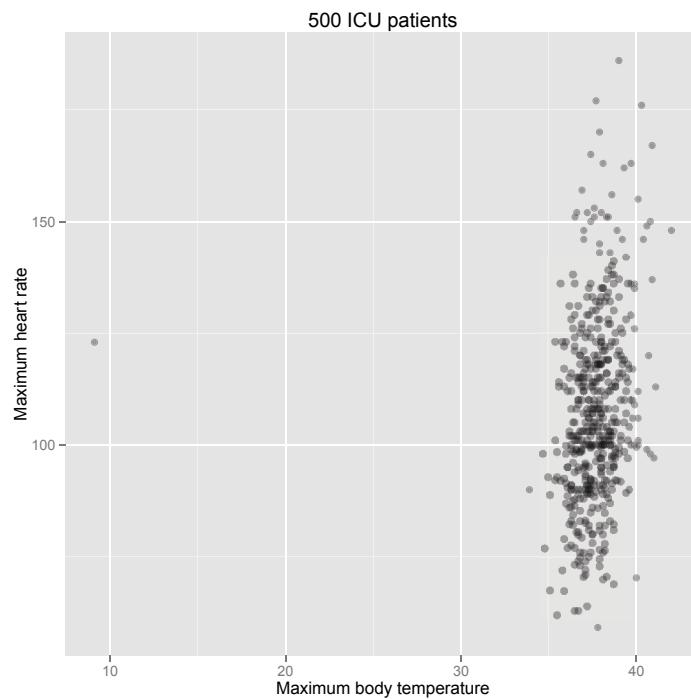


We plot the data by means of a scatter diagram and thereby also verify, if the assumption of a linear relation is justified.

```

1 ggplot(ICUData, aes(x=temperature, y=heart.rate)) +
2   ## shape = 19: somewhat larger point
3   ## alpha = 0.25: strength of alpha blending
4   geom_point(shape=19, alpha=0.25) +
5   ## title and axes labels
6   ggtitle("500 ICU patients") + xlab("Maximum body temperature") +
7   ylab("Maximum heart rate")

```



As before, we clearly see the extreme value of patient 398. We repeat the analysis without this patient.

```
1 cor(ICUData$temperature[-398], ICUData$heart.rate[-398])
```

```
[1] 0.2978033
```

By omitting a single value, the Pearson correlation is almost twice as large. That is, single outliers may have a strong influence on Pearson correlation. We investigate the influence of outliers on Spearman's ρ and Kendall's τ .

```
1 ## Spearman's rho
2 cor(ICUData$temperature , ICUData$heart.rate , method = "spearman")
```

```
[1] 0.2659957
```

```
1 cor(ICUData$temperature [-398] , ICUData$heart.rate [-398] , method = "spearman")
[1] 0.2707241
```

```
1 ## Kendall's tau
2 cor(ICUData$temperature , ICUData$heart.rate , method = "kendall")
```

```
[1] 0.1826903
```

```
1 cor(ICUData$temperature [-398] , ICUData$heart.rate [-398] , method = "kendall")
[1] 0.1858804
```

I joined MITAS because
I wanted **real responsibility**




The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com

Month 16
I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements

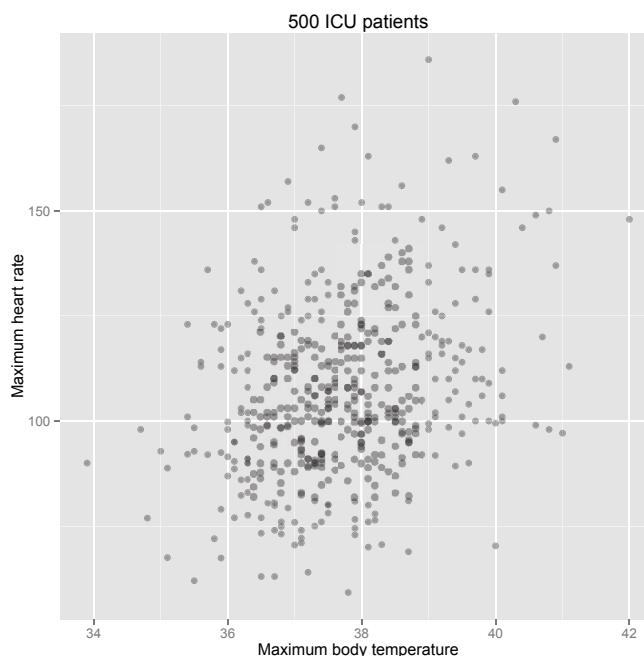



Both rank correlations change only slightly. Thus, the transition to ranks generates a certain robustness against outliers comparable to quantiles. The following scatter plot (without patient 398) suggests that there is at least a monotone relation between maximum body temperature and maximum heart rate. There might even be a linear relation.

```

1 ggplot(ICUData[-398,], aes(x=temperature , y=heart.rate)) +
2   ## shape = 19: somewhat larger point
3   ## alpha = 0.25: strength of alpha blending
4   geom_point(shape=19, alpha=0.25) +
5   ## title and axes labels
6   ggtitle("500 ICU patients") + xlab("Maximum body temperature") +
7   ylab("Maximum heart rate")

```



Note:

The popular saying: "A picture is worth a thousand words" applies also to statistics. Hence, always try to look at your data. It serves as a check of the data, e.g. for identifying wrong or erroneous observations or outliers as well as for confirming computed results.

2.6 Exercises

Use the ICU dataset and always briefly describe your results.

1. Compute absolute and relative frequencies for variable `outcome`.
2. Use a bar chart to visualize the relative frequencies for variable `outcome`. Apply the standard function `barplot` as well as the functions of package "ggplot2" (Wickham (2009)).
3. Determine the 95% quantile, median, inter quartile range, MAD, arithmetic mean, standard deviation, coefficient of variation, skewness, and kurtosis of variable `heart.rate`. What do the results tell you about the distribution of the values?
4. Draw a box-and-whisker plot as well as a histogram combined with a density plot of variable `heart.rate`. Apply the standard functions as well as the functions of package "ggplot2" (Wickham (2009)).
5. Investigate in a suitable manner the relation between variables `liver.failure` and `outcome`. Plot the data in an appropriate way.
6. Check in a suitable manner, if there is a connection between variables `age` and `SAPS.II`. Plot the data in an appropriate way.

"I studied English for 16 years but...
...I finally learned to speak it in just six lessons"

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



3 Colors and Diagrams

This rather short chapter deals with the correct use of colors and the generation of diagrams, which represent the available data in a most suitable way. It covers the following topics:

- Recommendations for handling colors
- Use of predefined color palettes
- Export of diagrams
- Recommendations for generating diagrams according to E. Tufte

The R code of this chapter is included in file `Colors.R`, which can be downloaded from my website (link: www.stamats.de/RCodeEN.zip). As described at the beginning of Chapter 2, you should additionally use your own R script and experiment with your own R code.

3.1 Colors

As we have seen in the last chapter, diagrams play a crucial role in understanding the available data. By the proper use of colors, the expressiveness and aesthetics of a graphic can be clearly improved. Figure 3.1 shows a negative example. Neither the type of diagram nor the colors are appropriately chosen. The selected type of diagram makes it hard to tell the exact proportions (as absolute or relative frequencies). The colors red and blue are very intense and not adapted to the answers. Furthermore, the graphic does not make clear, if there is another category between “We don’t care about colors” and “We love colors”. The category might exist, but nobody selected it or the category was not provided.

At the DSC conference 2003, Ross Ihaka (Ihaka (2003)) specified the following options for handling colors:

1. Avoid colors.
2. Determine colors by experimentation.
3. Use “good taste” or expertise.
4. Use fixed palettes designed by an expert.
5. Look for guiding principles.

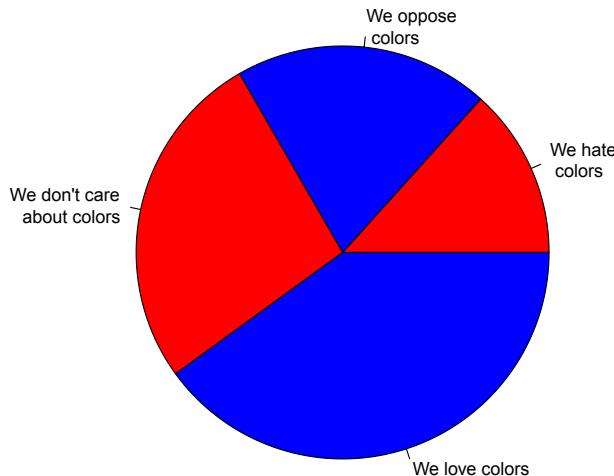


Figure 3.1: A negative example for using colors and diagrams.

We briefly comment on these options:

- Ad 1:** Of course, one can try to avoid colors, but colors can be very helpful and can clearly improve the expressiveness of a graphic.
- Ad 2:** The determination of colors by experimentation is usually very time consuming.
- Ad 3:** This requires a special talent for colors or a respective experience in handling colors.
- Ad 4:** Good idea!
- Ad 5:** Where can we find such guiding principles?

The following basic principles in handling colors are given in Zeileis et al. (2009):

- The colors should not be unappealing.
- The colors in a statistical graphic should cooperate with each other.
- The colors should work everywhere.

A project that applies these principles is ColorBrewer (Harrower and Brewer (2003)). It provides color palettes for various purposes on the website www.colorbrewer2.org. These color palettes can be applied in R by package "RColorBrewer" (Neuwirth (2014)). We install the package. We can either use

```
1 install.packages("RColorBrewer")
```

or install the package via window *Packages* of RStudio; for more details we refer to Section 2.4.1. We load the package and take a look at the various color palettes using function `display.brewer.all`. First, we consider the qualitative color palettes, which can be used for displaying categorical variables.

```
1 library(RColorBrewer)
2 display.brewer.all(type = "qual")
```



In this case, it is important that there is no color that dominates the others. All colors should appeal equally “important”. The second group of color palettes provides colors for attributes, whose levels range from unimportant or uninteresting to important or interesting.

In the past four years we have drilled
89,000 km

That's more than **twice** around the world.

Who are we?
We are the world's largest oilfield services company¹. Working globally—often in remote and challenging locations—we invent, design, engineer, and apply technology to help our customers find and produce oil and gas safely.

Who are we looking for?
Every year, we need thousands of graduates to begin dynamic careers in the following domains:

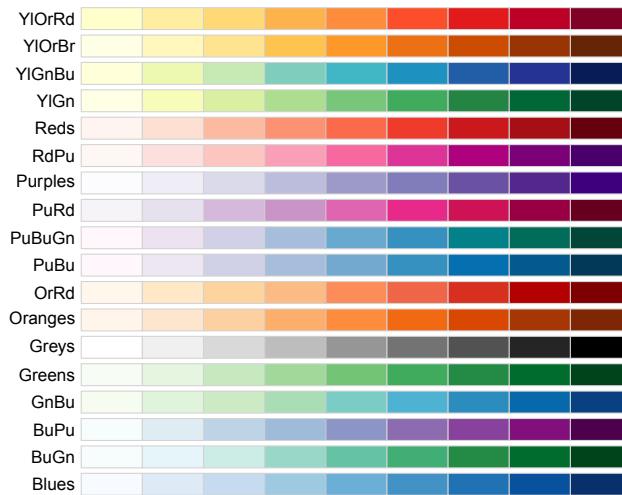
- **Engineering, Research and Operations**
- **Geoscience and Petrotechnical**
- **Commercial and Business**

What will you be?

Schlumberger

¹Based on Fortune 500 ranking 2011. Copyright © 2015 Schlumberger. All rights reserved.

```
1 display.brewer.all(type = "seq")
```



Finally, there is a third group of color palettes for variables with a range from negative to neutral to positive.

```
1 display.brewer.all(type = "div")
```



On the website www.colorbrewer2.org one can additionally choose colors by the following criteria:

- colorblind safe
- print friendly
- photocopy safe
- LCD friendly

Figure 3.2 compares the introductory negative example with a corresponding pie chart, where the colors are adapted to the categories. For the new diagram ColorBrewer palette RdYlGn was applied. A further improvement of the diagram could either consist of labeling the pieces by (absolute or relative) frequencies or by transferring the results to a bar chart. In particular, in case of a bar chart one could make clear that there is a category “We tolerate colors” by adding a bar of height 0; see Figure 3.3.

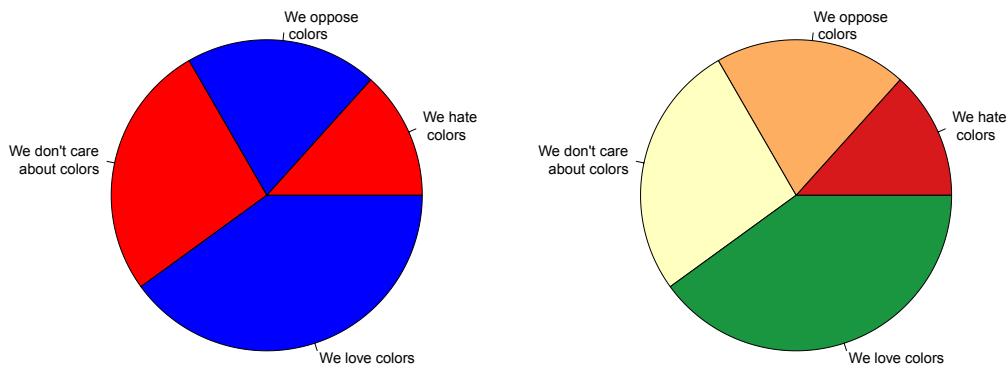


Figure 3.2: A negative example with improved colors.

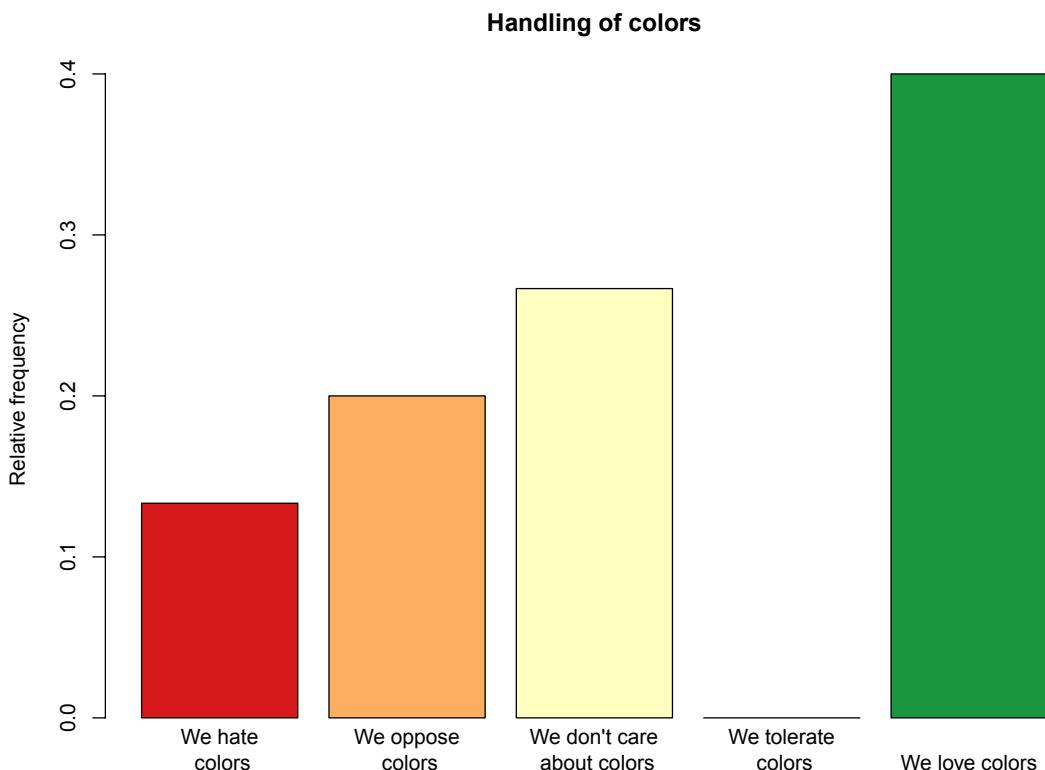


Figure 3.3: From a negative to a positive example.

We plot the absolute frequencies of the types of surgeries as in Section 2.4.1, where we additionally use color palette Set1 of ColorBrewer. For this, we first generate a vector of colors by applying function `brewer.pal` of package "RColorBrewer" (Neuwirth (2014)).

```
1 ## n = 5 colors of palette with name Set1
2 cols <- brewer.pal(n = 5, name = "Set1")
3 cols
```

```
[1] "#E41A1C" "#377EB8" "#4DAF4A" "#984EA3" "#FF7F00"
```

That is, the colors are saved in hexadecimal code. R includes several functions for various color spaces, which can be used to determine the hexadecimal code of colors. For instance, if the red-green-blue (RGB) code is known, one can use function `rgb`.

```
1 rgb(red = 228, green = 26, blue = 28, maxColorValue = 255)
```

```
[1] "#E41A1C"
```

A large number of colors can also be specified by their names. More precisely, there are 657 colors in R that are saved by their names. One can use function `colors` to display these colors and function `col2rgb` to determine their red-green-blue code; e.g.

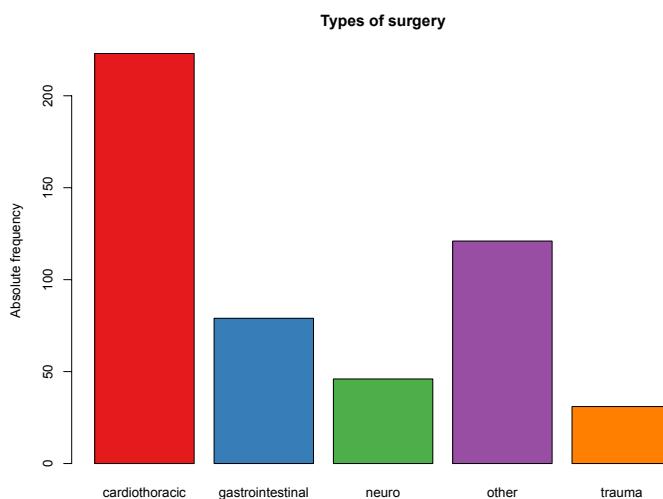


```
1 col2rgb("royalblue")
```

	[,1]
red	65
green	105
blue	225

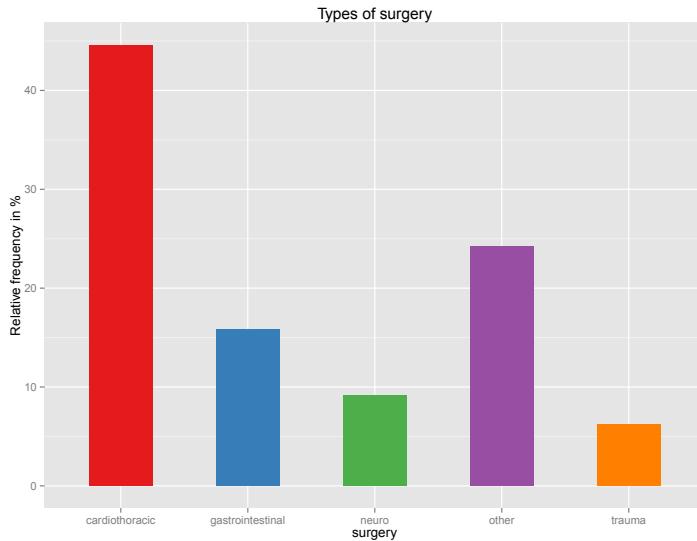
The standard functions for plotting data all have argument `col`, which can be used to specify colors. We now generate the bar chart.

```
1 barplot(table(ICUData$surgery), main = "Types of surgery",
2         ylab = "Absolute frequency", col = cols)
```



Package "ggplot2" (Wickham (2009)) also provides various ways of using colors. We generate the bar chart of the relative frequencies included in Section 2.4.1, where we this time color the bars via ColorBrewer palette Set1.

```
1 ## Define data
2 ggplot(ICUData, aes(x=surgery)) +
3   ## Add bars of relative frequencies
4   geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5,
5             ## Fill bars with color
6             fill = cols) +
7   ## Title and label of y-axis
8   ggttitle("Types of surgery") + ylab("Relative frequency in %")
```



3.2 Excursus: Export of Diagrams

In RStudio the generated diagrams are shown in window *Plots* and can be exported by menu item *Export* to various graphic formats or pdf; see Figure 3.4. By clicking on *Save as Image...* a new window opens,

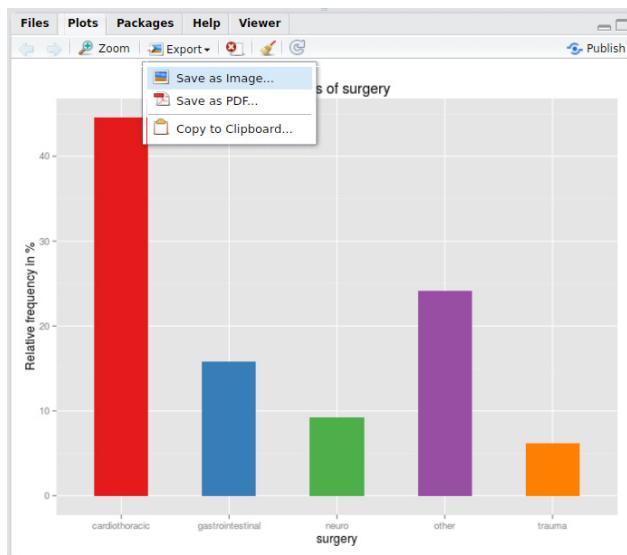
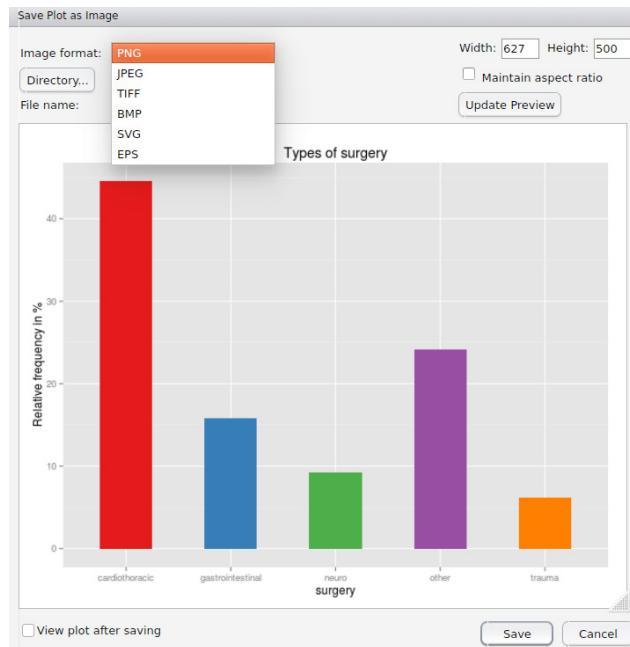
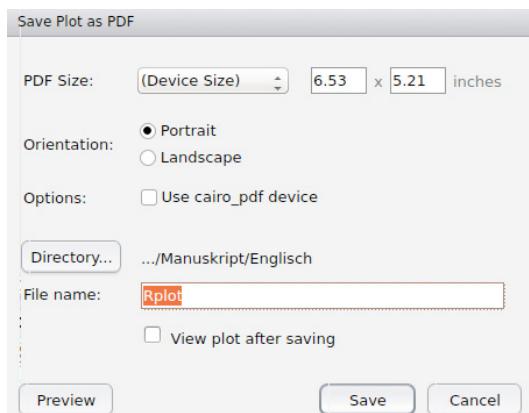


Figure 3.4: RStudio window *Plots* with an example.

in which the size of the image, the file name, the folder and the graphic format can be chosen (see Figure 3.5). It depends on the operating system and maybe additionally installed graphics software, which graphic formats are available. By choosing *Save as PDF...*, the window shown in Figure 3.6 opens. One can specify the size, the file name and the folder.

**Figure 3.5:** RStudio window for saving a plot as image.**Figure 3.6:** RStudio window for saving a plot as pdf file.

For a quick import of plots for example in a document or email, one can use menu item *Copy to Clipboard*.... In most cases however, it is preferable to first save the graphic as image or pdf and then import the generated file.

The described options are convenient and quick and in most cases lead to the wanted result. But, especially in case of complex graphics there may be problems under certain circumstances. Moreover, sometimes it is necessary to adapt further parameters during the export such as resolution, compression rate or font size. In such a situation, one directly has to use the export functions of R. As already mentioned, the available devices depend on the operating system and maybe additionally installed software. The main functionality is provided by base package "grDevices" (R Core Team (2015a)). In addition, there are some contributed packages offering further options. Table 3.1 contains an overview of common devices supported by R.

Function name	Description
bmp	Bitmap (bmp) a standard format of raster graphics in Microsoft Windows.
jpeg	Compressed image files of raster graphics, very common in Internet.
png	Portable network graphics (png) for lossless compressed image files of raster graphics, usually more appropriate for statistical graphics than jpeg.
tiff	Tagged image file format (tiff) especially used for high-resolution printable raster graphics.
pdf	Portable document format (pdf) a very common file format that embeds graphics as vector graphics.
postscript	PostScript (ps) a vector graphics format frequently used for printing, especially the further developed Encapsulated PostScript (eps) is of interest for graphics.
svg	Scalable vector graphics (svg) a vector graphics format for web browsers.

Table 3.1: Overview of devices supported by R.

Excellent Economics and Business programmes at:


/


**university of
groningen**



**“The perfect start
of a successful,
international career.”**

www.rug.nl/feb/education

CLICK HERE
to discover why both socially
and academically the University
of Groningen is one of the best
places for a student to be

Note:

Raster graphics are based on a grid of pixels, where every pixel has a certain color. The best possible way to display such images is the resolution, in which they were generated. In case of rescaling, especially enlarging, the quality of these images declines. In contrast, vector graphics are based on a description of the image and can be rescaled without any problems. In addition, vector graphics often require clearly less memory.

The export of a plot to some file always consists of the following three steps:

1. Open the desired device.
2. Generate the plot.
3. Close the device using function `dev.off()`.

As an example, we generate a png image.

```

1 ## 1. Open the device
2 ## height and width in number of pixels
3 png(file = "Example_Image.png", height = 640, width = 640)
4 ## 2. Generate the plot
5 barplot(table(ICUData$surgery), main = "Type of surgery",
6         ylab = "Absolute frequency", col = cols)
7 ## 3. Close the device
8 dev.off()

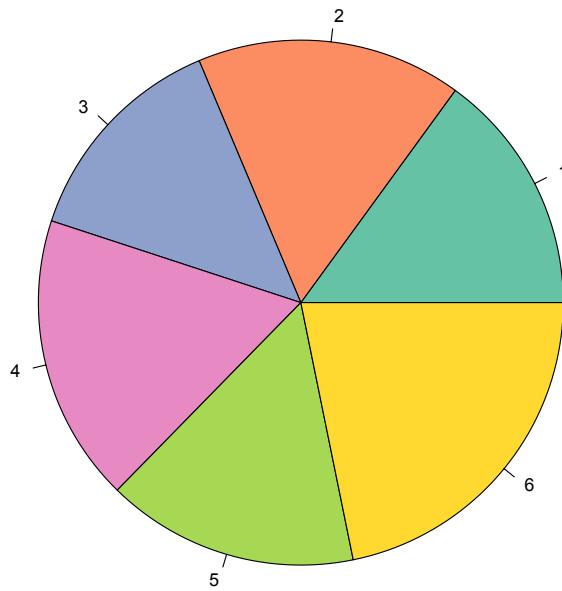
```

After running this code, there is an image file called `Example_Image.png` in the current working directory, which includes the generated plot.

Besides single images, one can even generate movies with R; e.g., package "animation" (Xie (2013)) provides various options and contains some interesting examples.

3.3 Diagrams

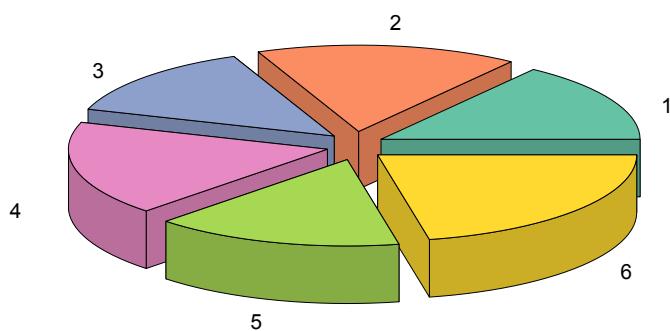
The negative example of Section 3.1 (see Figure 3.1) confirms the statement in Section 2.4.1, that pie charts are not the best option for displaying information. Please, try to order the categories shown in Figure 3.7 or even try to determine the plotted frequencies. This gets even worse by introducing a further dimension in form of a three-dimensional pie chart; see Figure 3.8.

**Figure 3.7:** Order the categories!

Note:

The third dimension in three-dimensional pie and bar charts, which are frequently used nowadays, leads to a perspective distortion. Moreover, it contradicts one of the recommendations of E. Tufte given below, as the number of information carrying dimensions (= 3) is larger than the dimension of the plotted data (= 2).

In contrast, the order of the categories is immediately visible by using a bar chart as in Figure 3.9.

**Figure 3.8:** Once again: Order the categories!

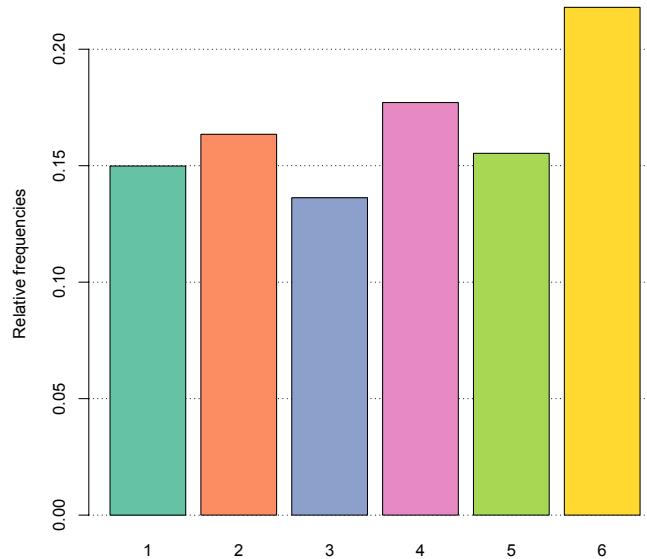


Figure 3.9: And once again: Order the categories!

The following recommendations go back to Eduard Tufte (see Globus (1994)):

- The numbers, that can be measured off the graphic, should be directly proportional to the numerical quantities represented by them.
- Use a clear, detailed and complete labeling to avoid a graphical bias and ambiguity.

..... Alcatel-Lucent 

www.alcatel-lucent.com/careers

What if
you could
build your
future and
create the
future?

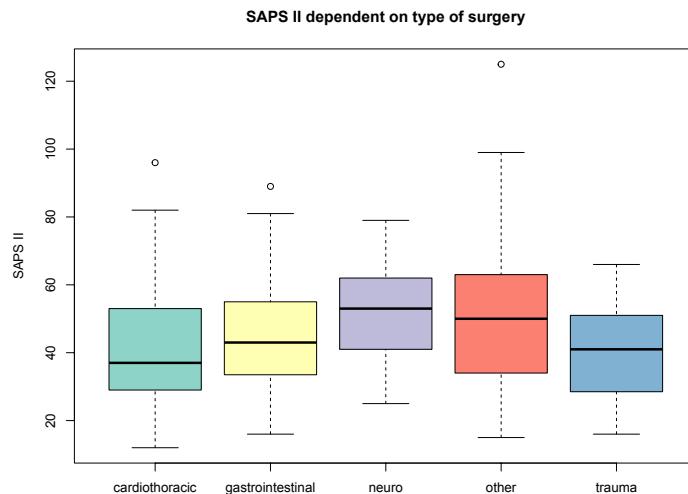


One generation's transformation is the next's status quo.
In the near future, people may soon think it's strange that
devices ever had to be "plugged in." To obtain that status, there
needs to be "The Shift".

- Explanations of the data should be given on the graphic itself.
- Important events in the data should be labeled.
- It is important to show the variation of the data and not of the design.
- The number of information carrying dimensions should not exceed the dimension of the data.
- Never use graphics outside of their context.

In the sequel, we present some more examples for using diagrams in combination with colors. We start with a plot of the SAPS II scores for the different types of surgery, where we use box-and-whisker plots. As there is no obvious order between the types of surgery, we choose a qualitative color palette, in this case `Set3` of ColorBrewer (Harrower and Brewer (2003)). First, we apply function `boxplot`.

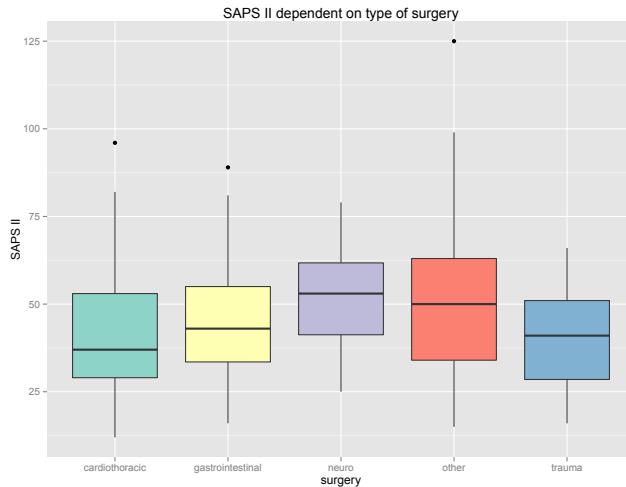
```
1 cols <- brewer.pal(n = 5, name = "Set3")
2 boxplot(SAPS.II ~ surgery, data = ICUData, ylab = "SAPS II",
3         main = "SAPS II dependent on type of surgery", col = cols)
```



For splitting the scores by types of surgery, we have used a so-called formula. The expression `SAPS.II ~ surgery` means that the left-hand side `SAPS.II` has to be considered in dependence of the right hand side `surgery`. Not surprisingly, we see the largest range in case of other surgeries and the values in case of neurological surgeries tend to be higher.

We repeat the plot using package "`ggplot2`" (Wickham (2009)). The colors can be specified by argument `fill` of function `geom_boxplot`.

```
1 ## Define data
2 ggplot(ICUData, aes(x = surgery, y = SAPS.II)) +
3   ## Box-and-whisker plot with colors
4   geom_boxplot(fill = cols) +
5   ## Labeling
6   ylab("SAPS II") + ggtitle("SAPS II dependent on type of surgery")
```

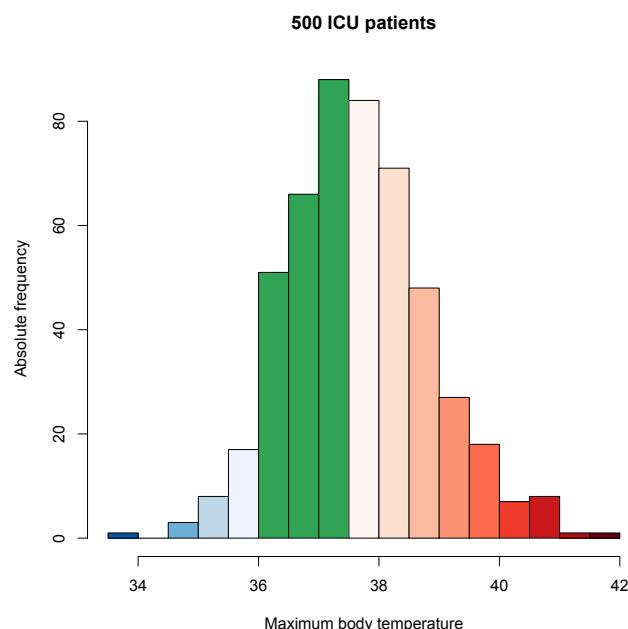


The use of colors is often also useful in case of histograms. We repeat the histogram of the maximum body temperature generated in Section 2.5.1. We split the maximum body temperature in the following three intervals: $< 36^{\circ}\text{C}$ (too low), $36\text{--}37.5^{\circ}\text{C}$ (normal), $> 37.5^{\circ}\text{C}$ (too high). For the first interval, consisting of five sub-intervals, we use ColorBrewer palette Blues and revert the order of the colors with function `rev`. For the normal range, consisting of three sub-intervals, we use color green (more precisely: #31A354) and replicate the color with function `rep`. For the third interval, consisting of nine sub-intervals, we select ColorBrewer palette Reds. For getting a better overview, we omit patient 398.

```

1 cols1 <- rev(brewer.pal(5, "Blues"))
2 cols2 <- rep("#31A354", 3)
3 cols3 <- brewer.pal(9, "Reds")
4 hist(ICUData$temperature[-398], breaks = seq(from = 33.5, to = 42, by = 0.5),
5      main = "500 ICU patients", ylab = "Absolute frequency",
6      xlab = "Maximum body temperature", col = c(cols1, cols2, cols3))

```



We generate a similar figure by means of package "ggplot2" (Wickham (2009)). Here, there is an additional (empty) sub-interval on the left- and right-hand side. Thus, we have to add one more color in category one (too low) and three (too high).

```

1 cols1 <- c(cols1[1], cols1)
2 cols3 <- c(cols3, cols3[9])
3 ggplot(ICUData[-398,], aes(x=temperature)) +
4   geom_histogram(binwidth = 0.5, right = TRUE,
5                 fill = c(cols1, cols2, cols3)) +
6   ylab("Absolute frequency") + xlab("Maximum body temperature") +
7   ggtitle("500 ICU patients")

```


Maastricht University *Leading in Learning!*

**Join the best at
the Maastricht University
School of Business and
Economics!**



Top master's programmes

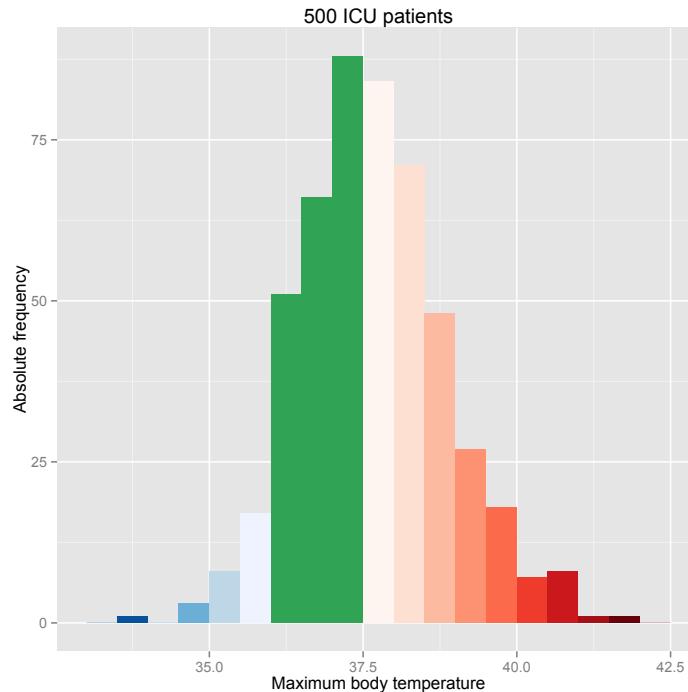
- 33rd place Financial Times worldwide ranking: MSc International Business
- 1st place: MSc International Business
- 1st place: MSc Financial Economics
- 2nd place: MSc Management of Learning
- 2nd place: MSc Economics
- 2nd place: MSc Econometrics and Operations Research
- 2nd place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

Visit us and find out why we are the best!
Master's Open Day: 22 February 2014

www.mastersopenday.nl

Maastricht
University is
the best specialist
university in the
Netherlands
(Elsevier)



The use of colors is also helpful in case of scatter diagrams and can for example be used to visualize a third variable in addition to the variables on x and y axis. First, we apply function `plot`. We generate a vector of colors that has entry red (more precisely: `#E41A1C`) for females and entry blue (more precisely: `#377EB8`) for males. For this, we start with an empty vector generated by function `character`. Accordingly, it is a vector that can include letters or strings. By using the square brackets `[`, the vector is filled with red at positions of female patients and with blue at positions of male patients. The sign `==` is a so-called **logical operator** that can be used to check for equality.

```

1 ## Generate empty vector
2 colsSex <- character(nrow(ICUData))
3 ## Fill with colors
4 colsSex[ICUData$sex == "female"] <- "#E41A1C"
5 colsSex[ICUData$sex == "male"] <- "#377EB8"

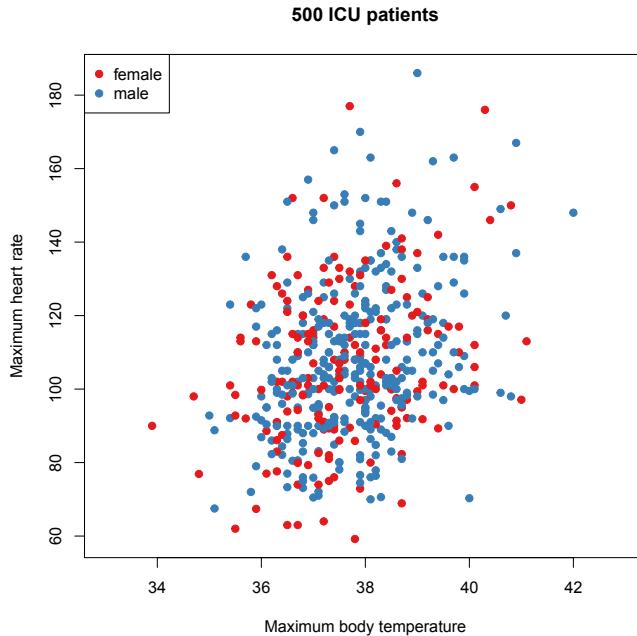
```

Using argument `pch = 19` (`pch=point character`), we select a thicker point as plot symbol. The possible plot symbols are specified in the help page of function `points`. We restrict the x axis to the interval `[33; 43]` to obtain a better overview. We also add a legend to the plot via function `legend` to explain the meaning of the colors.

```

1 plot(x = ICUData$temperature , y = ICUData$heart.rate , pch = 19,
2       xlab = "Maximum body temperature" , ylab = "Maximum heart rate",
3       main = "500 ICU patients" , col = colsSex , xlim = c(33,43))
4 legend(x = "topleft" , legend = c("female" , "male") , pch = 19,
5        col = c("#E41A1C" , "#377EB8"))

```



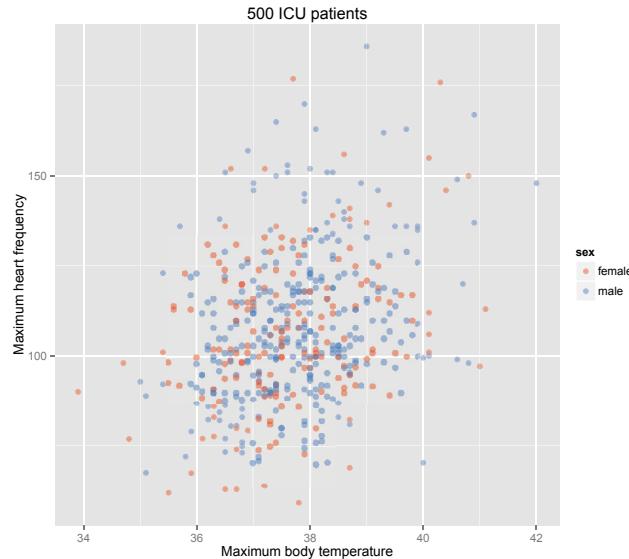
The observations of females and males are quite uniformly distributed over the whole scatter diagram, which indicates that there is no influence of sex on maximum body temperature and maximum heart rate.

In case of package "ggplot2" (Wickham (2009)), it is very easy to additionally use alpha blending. Furthermore, the assignment of colors to the sexes is much easier and can be done by applying function `scale_colour_manual`.

```

1 ggpplot(ICUData[-398,], aes(x=temperature , y=heart.rate , colour=sex)) +
2   ## shape = 19: somewhat larger point
3   ## alpha = 0.4: strength of blending
4   geom_point(shape=19, alpha=0.4) +
5   ## colors
6   scale_colour_manual(values = c("#E41A1C" , "#377EB8")) +
7   ## labeling
8   ggtitle("500 ICU patients") + xlab("Maximum body temperature") +
9   ylab("Maximum heart frequency")

```



3.4 Exercises

Use the ICU dataset.

1. Generate a bar chart to plot the relative frequencies of variable outcome. Use the standard function `barplot` as well as the functions of package "ggplot2" (Wickham (2009)) in combination with color palette `Set2` of package "RColorBrewer" (Neuwirth (2014)).
2. Draw box-and-whisker plots of variable age where you split the values by variable outcome. Use the standard function `boxplot` as well as the functions of package "ggplot2" (Wickham (2009)) in combination with appropriate colors for the boxes.
3. Generate a histogram of variable `heart.rate`. Consider the range from 70 to 100 as normal. Use appropriate colors for the histogram and apply the standard function `hist` as well as the functions of package "ggplot2" (Wickham (2009)). Save the plots as png images.
4. Draw a scatter diagram of the heart rate dependent on age and additionally mark females and males by colors. Use the standard function `plot` as well as the functions of package "ggplot2" (Wickham (2009)). Save the plots in pdf files.

4 Probability Distributions

We need models of probability theory to be able to infer from a sample to the underlying population (cf. Section 2.1). The basis of such models are probability distributions, where in the simplest case, the probability distributions are already the models that shall be investigated. In this case, the goal of inferential (parametric) statistics consists of estimating the unknown parameters of the assumed probability distributions from the given data.

This chapter introduces the probability, cumulative distribution, and quantile functions of discrete and (absolutely) continuous probability distributions. It covers the following probability distributions:

- Bernoulli distribution $\text{Bernoulli}(p)$
- Binomial distribution $\text{Binom}(m, p)$
- Hypergeometric distribution $\text{Hyper}(m, n, k)$
- Negative binomial distribution $\text{Nbinom}(r, p)$
Special cases: Pascal distribution, Pólya distribution, geometric distribution
- Poisson distribution $\text{Pois}(\lambda)$
- Normal distribution $\text{Norm}(\mu, \sigma^2)$
- Log-normal distribution $\text{Lnorm}(\mu, \sigma)$

**Empowering People.
Improving Business.**

BI Norwegian Business School is one of Europe's largest business schools welcoming more than 20,000 students. Our programmes provide a stimulating and multi-cultural learning environment with an international outlook ultimately providing students with professional skills to meet the increasing needs of businesses.

BI offers four different two-year, full-time Master of Science (MSc) programmes that are taught entirely in English and have been designed to provide professional skills to meet the increasing need of businesses. The MSc programmes provide a stimulating and multi-cultural learning environment to give you the best platform to launch into your career.

- MSc in Business
- MSc in Financial Economics
- MSc in Strategic Marketing Management
- MSc in Leadership and Organisational Psychology

www.bi.edu/master

**EFMD
EQUIS
ACCREDITED**

- Gamma distribution Gamma (σ, α)
Special cases: Exponential distribution, Erlang distribution, χ^2 distribution
- Weibull distribution Weibull (σ, α)
- Distributions arising in connection with normal distributions: χ^2 distribution Chisq (n), t distribution t (n), F distribution F (m, n)

The R code of this chapter is included in file `ProbabilityDistributions.R`, which can be downloaded from my website (link: www.stamats.de/RCodeEN.zip). It is advisable to use an additional R script for your own R code. More details are given at the beginning of Chapter 2.

4.1 Discrete Distributions

We consider a function X , which attains its values in the space of natural numbers with certain probabilities. Such a function X is called a **discrete random variable**. The values of a random variable are called realisations.

We can uniquely describe the **discrete probability distribution** or **discrete distribution** of a random variable X by specifying the **probability** $P(X = k)$ of all possible values $k \in \mathbb{N}$ of X . The function

$$d(k) = P(X = k) \quad (4.1)$$

is called **probability mass function** of X . The function

$$p(k) = P(X \leq k) = \sum_{i=0}^k P(X = i) = \sum_{i=0}^k d(i) \quad (4.2)$$

is called **cumulative distribution function** of X . Its inverse is the **quantile function**

$$q(p) = \min \{k \in \mathbb{N} \mid p(k) \geq p\} \quad p \in [0, 1] \quad (4.3)$$

Important parameters of a distribution, which can also be used for its characterization, are **expectation** and **variance**. The expectation of X , $E(X)$ for short, is the value of X that we can expect in mean. It holds

$$E(X) = \sum_{k \in \mathbb{N}} k \cdot d(k) \quad (4.4)$$

i.e., the possible levels of X are multiplied by their probabilities and added. The variance of X , $Var(X)$ for short, is the expected value of the quadratic deviations from the expectation

$$Var(X) = \sum_{k \in \mathbb{N}} (k - E(X))^2 \cdot d(k) \quad (4.5)$$

Often, the square root of the variance is considered, which is called **standard deviation** of X , $\sigma_X = \sqrt{\text{Var}(X)}$ for short.

In this section, several important discrete distributions are introduced.

Bernoulli distribution

The simplest discrete distribution is the so-called Bernoulli distribution, for which two applications are sketched in the following example.

Example 4.1.

- a) We consider the production of bulbs, where 1% of the bulbs are defective. That is, we can describe the production process by a discrete random variable X , which may attain the values 0 = defective and 1 = not defective. This leads to the following probability mass function

$$P(X = 0) = 0.01 \quad \text{and} \quad P(X = 1) = 1 - 0.01 = 0.99 \quad (4.6)$$

- b) In a randomized controlled clinical trial two interventions are compared where 65% of the patients are randomly assigned to intervention I and accordingly, 35% of the patients to intervention II. This procedure can be described by the discrete random variable X , which attains value 0 = intervention I with probability 65% and value 1 = intervention II with probability 35%, respectively. It yields the following probability mass function

$$P(X = 0) = 0.65 \quad \text{and} \quad P(X = 1) = 1 - 0.65 = 0.35 \quad (4.7)$$

The probability distribution that underlies both examples is defined as follows.

Definition 4.2 (Bernoulli distribution). *Let X be some discrete random variable, that may only attain values 0 and 1. Then, the probability mass function of the distribution of X is*

$$d(k) = P(X = k) = p^k(1 - p)^{1-k} = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases} \quad (4.8)$$

where $p \in [0, 1]$. The distribution is called **Bernoulli distribution** with parameter p , abbreviated by $X \sim \text{Bernoulli}(p)$.

Binomial distribution

The Bernoulli distribution can be generalized to the so-called binomial distribution. The following example shows two possible applications.

Example 4.3.

- a) We again consider the production of bulbs, where 1% of the bulbs is defective and want to check the quality of the last batch. For this purpose, we randomly draw **with** replacement a sample of size $m = 20$ bulbs from the batch (=population). Let X be the random variable describing the number of defective bulbs. By means of the distribution of X , we can for instance specify how likely it is to draw exactly one defective bulb. We get

$$P(X = 1) = 20 \cdot 0.01 \cdot 0.99^{19} = 16.5\% \quad (4.9)$$

Because of the 20 draws, there are 20 possibilities to draw a defective bulb, which happens with a probability of 0:01. In the remaining 19 draws a properly functioning bulb is drawn with probability 0:99 in each draw.

- b) In 2014, the prevalence of diabetes among adults amounted to about 9% (WHO (2015b)), where **prevalence** is the proportion of a population that has a disease. We conduct a trial and randomly draw **with** replacement a sample of 50 persons. The number of persons having diabetes in our sample is denoted by the random variable X . How likely is it, that our sample contains at least two persons with diabetes? In this case, it is simpler to consider the so-called complementary event: the sample contains no or exactly one person with diabetes. We obtain

$$P(X = 0) = 0.91^{50} = 0.9\% \quad \text{and} \quad P(X = 1) = 50 \cdot 0.09 \cdot 0.91^{49} = 4.4\% \quad (4.10)$$

Need help with your dissertation?

Get in-depth feedback & advice from experts in your topic area. Find out what you can do to improve the quality of your dissertation!

Get Help Now



Go to www.helpmyassignment.co.uk for more info



Helpmyassignment

Thus, the wanted probability reads

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - P(X = 0) - P(X = 1) = 1 - 0.009 - 0.044 = 94.7\% \quad (4.11)$$

In general, we get the following discrete probability distribution.

Definition 4.4. We consider a box (urn) with black and white balls, where the proportion of white balls is equal to $p \in [0, 1]$. We randomly draw m -times ($m \in \mathbb{N}$) with replacement from this box and describe the number $k \in \{1, 2, \dots, n\}$ of drawn white balls by the random variable X . Then, the probability mass function of X reads

$$d(k) = P(X = k) = \binom{m}{k} p^k (1-p)^{m-k} \quad (4.12)$$

where

$$\binom{m}{k} = \frac{m!}{k!(m-k)!} \quad (4.13)$$

is the binomial coefficient and ! indicates factorials. This distribution is called **Binomial distribution** with parameters m and p , abbreviated by $X \sim \text{Binom}(m, p)$.

We give some additional explanations.

Remark 4.5.

- a) The factorial of $k \in \mathbb{N}$ is defined as

$$k! = \begin{cases} 1 & \text{if } k = 0 \\ 1 \cdot 2 \cdot \dots \cdot k & \text{if } k \geq 1 \end{cases} \quad (4.14)$$

- b) A closer look at the probability mass functions of the Bernoulli and the binomial distribution shows $\text{Bernoulli}(p) = \text{Binom}(1, p)$.

- c) Expectation and variance of $\text{Binom}(m, p)$ are

$$E(X) = m \cdot p \quad \text{Var}(X) = m \cdot p \cdot (1-p) \quad (4.15)$$

The statistical software R includes the probability mass functions, cumulative distribution functions and quantile functions of many discrete probability distributions. In general, the names of these basic functions always consist of a prefix and some abbreviation of the name of the probability distribution. The possible prefixes are

d: probability mass function

p: cumulative distribution function

q: quantile function

r: function for generating (pseudo) random numbers

Therefore, it is sufficient to know the abbreviation of the distribution to apply the respective functions.

In case of the binomial distribution, the abbreviation is `binom` and the respective functions are: `dbinom`, `pbinom`, `qbinom`, and `rbinom`. The parameters m and p of the binomial distribution are called `size` and `prob` in R.

We compute the probabilities of Example 4.3 using R.

```
1 ## a) Exactly one defective bulb
2 dbinom(1, size = 20, prob = 0.01)
```

```
[1] 0.1652337
```

```
1 ## b) No person with diabetes
2 dbinom(0, size = 50, prob = 0.09)
```

```
[1] 0.008955083
```

```
1 ## b) Exactly one person with diabetes
2 dbinom(1, size = 50, prob = 0.09)
```

```
[1] 0.04428338
```

For determining the probability that at least two persons with diabetes are drawn, it is easier to apply the cumulative distribution function.

```
1 ## b) At least two persons with diabetes: 1-P(X ≤ 1)
2 1 - pbinom(1, size = 50, prob = 0.09)
```

```
[1] 0.9467615
```

Alternatively and numerically somewhat more precise, we can compute this probability by using argument `lower.tail = FALSE`. Then, the probability $P(X > k)$ instead of $P(X \leq k)$ is computed.

```
1 ## b) At least two persons with diabetes: P(X > 1)
2 pbinom(1, size = 50, prob = 0.09, lower.tail = FALSE)
```

```
[1] 0.9467615
```

By means of the quantile function, we can for instance determine how many defective bulbs we can at most expect with a probability of 99%.

Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations.

Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF

```
1 qbinom(0.99 , size = 20, prob = 0.01 )
```

```
[1] 2
```

Consequently, if there are more than two defective bulbs during quality control, it may indicate a quality problem; i.e., a larger proportion of defective bulbs. Because it is very unlikely (< 1%) to draw three or more defective bulbs, if there are only 1% defective bulbs in the batch.

Function `rbinom` can be used to generate random numbers. If we adapt this to our diabetes example, every random number represents a trial, more precisely, the number of persons with diabetes in that trial. We simulate ten trials.

```
1 rbinom(10 , size = 50, prob = 0.09 )
```

```
[1] 5 4 5 2 5 1 7 3 4 3
```

We can also plot the probability mass function, the cumulative distribution function, and the quantile function of this binomial distribution. For this, package "`distr`" (Ruckdeschel et al. (2006)) can be used, which includes an object-oriented implementation of probability distributions. We can install the package via

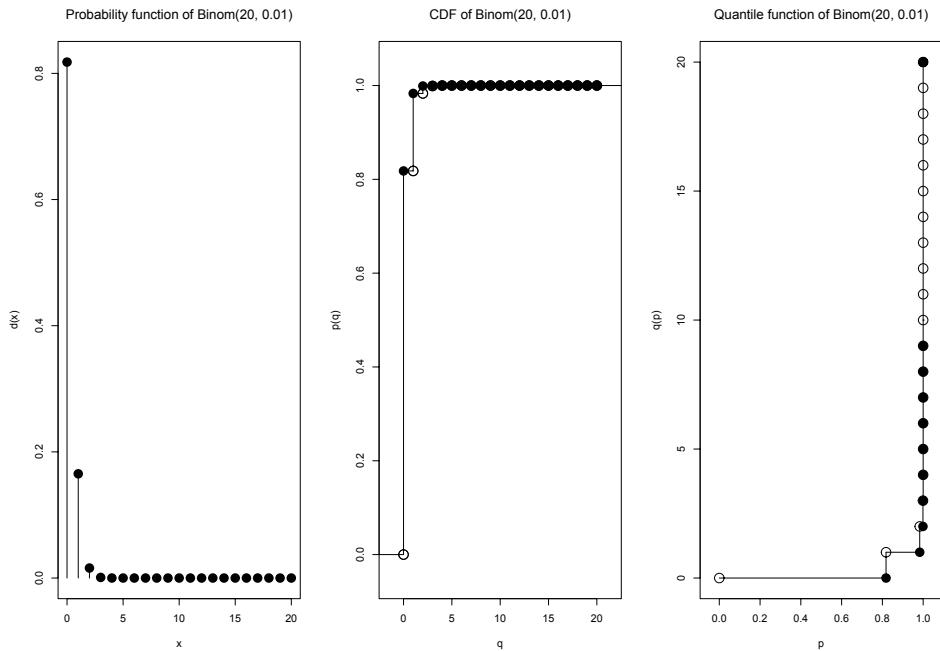
```
1 install.packages("distr")
```

or window *Packages* of RStudio (cf. Section 2.4.1). We load the package and by function `Binom` generate a random variable X with distribution $\text{Binom}(20, 0.01)$ matching our bulb example.

```
1 library(distr)
2 X <- Binom(size = 20, prob = 0.01 )
```

By means of function `plot`, we can display the probability (mass) function, the cumulative distribution function (CDF), and the quantile function of a given random variable.

```
1 plot(X)
```



Hypergeometric distribution

If we draw without instead of with replacement, it results in the so-called hypergeometric distribution. The following example is very similar to Example 4.3.

Example 4.6.

- a) We consider a box (= population) with $m + n = 500$ bulbs, where $m = 5$ are defective and randomly draw **without** replacement a sample of $k = 20$ bulbs. Let X be the random variable describing the number of defective bulbs in our sample. By means of the distribution of X , we can for instance determine how likely it is to draw no defective bulb. It holds

$$P(X = 0) = \frac{495}{500} \cdot \frac{494}{499} \cdot \dots \cdot \frac{476}{481} = 81.5\% \quad (4.16)$$

There are 20 draws, where in each draw a functioning bulb is drawn and put aside. Hence, numerator and denominator are reduced by one after each draw; i.e., the proportion of defective bulbs changes from draw to draw.

b) In 2000, the population of Andorra was about 66000 inhabitants (Wikipedia (2015a)), where about 6000 inhabitants (WHO(2015a)) had diabetes. We conduct a trial in Andorra and randomly draw **without** replacement 50 inhabitants. Let X be the random variable describing the number of persons having diabetes in our sample. How likely is it that there is at least one person in our sample having diabetes? We consider the complementary event: the sample includes no person with diabetes and obtain

$$P(X = 0) = \frac{60000}{66000} \cdot \frac{59999}{65999} \cdot \dots \cdot \frac{59951}{65951} = 0.9\% \quad (4.17)$$

Thus, the wanted probability is

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 0.009 = 99.1\% \quad (4.18)$$

We define the hypergeometric distribution.

Definition 4.7. We consider a box (urn) with $m \in \mathbb{N}$ white and $n \in \mathbb{N}$ black balls and randomly draw $k \in \mathbb{N}$ balls without replacement ($k < m+n$). The random variable X , describing the number j of white balls in the sample ($j \leq m$), has the following probability mass function

$$d(j) = P(X = j) = \frac{\binom{m}{j} \binom{n}{k-j}}{\binom{m+n}{k}} \quad (4.19)$$

“I studied English for 16 years but...
...I finally learned to speak it in just six lessons”

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

The distribution is called **hypergeometric distribution** with parameters m , n and k , abbreviated by $X \sim \text{Hyper}(m, n, k)$.

We give some additional explanations.

Remark 4.8.

- a) For large populations and samples the computation of the binomial coefficients included in the definition of the hypergeometric distribution is difficult. This is caused by the fact that factorials of large numbers have to be determined and the factorial grows exponentially.
- b) Already for populations of a moderate size and if the sample is not too large compared to the population, the difference between hypergeometric and binomial distribution is very small. It means, it only happens with a small probability that the same ball is drawn more than once.
- c) Expectation and variance of $\text{Hyper}(m, n, k)$ read

$$E(X) = k \cdot \frac{m}{m+n} \quad \text{Var}(X) = k \cdot \frac{m}{m+n} \cdot \frac{n}{m+n} \cdot \frac{m+n-k}{m+n-1} \quad (4.20)$$

The formulas show a certain analogy to the binomial distribution. The factor $\frac{m+n-k}{m+n-1}$ representing the essential difference to the binomial distribution is called **finite sample correction**. We will meet it once again in Example 5.12.

The hypergeometric distribution is abbreviated by `hyper` in R leading to functions `dhyper`, `phyper`, `qhyper`, and `rhyper`. We compute the probabilities of Example 4.6 using R.

```
1 ## a) no defective bulb
2 dhyper(0, m = 5, n = 495, k = 20)
```

```
[1] 0.8146893
```

```
1 ## b) no person with diabetes
2 dhyper(0, m=6000, n=60000, k = 50)
```

```
[1] 0.008502747
```

We can compute the probability of at least one person with diabetes by directly applying function `phyper` with argument `lower.tail = FALSE`

```
1 ## b) at least one person with diabetes
2 phyper(0, m=6000, n=60000, k=50, lower.tail = FALSE)
```

```
[1] 0.9914973
```

We simulate 10 samples of size 50 for our diabetes example.

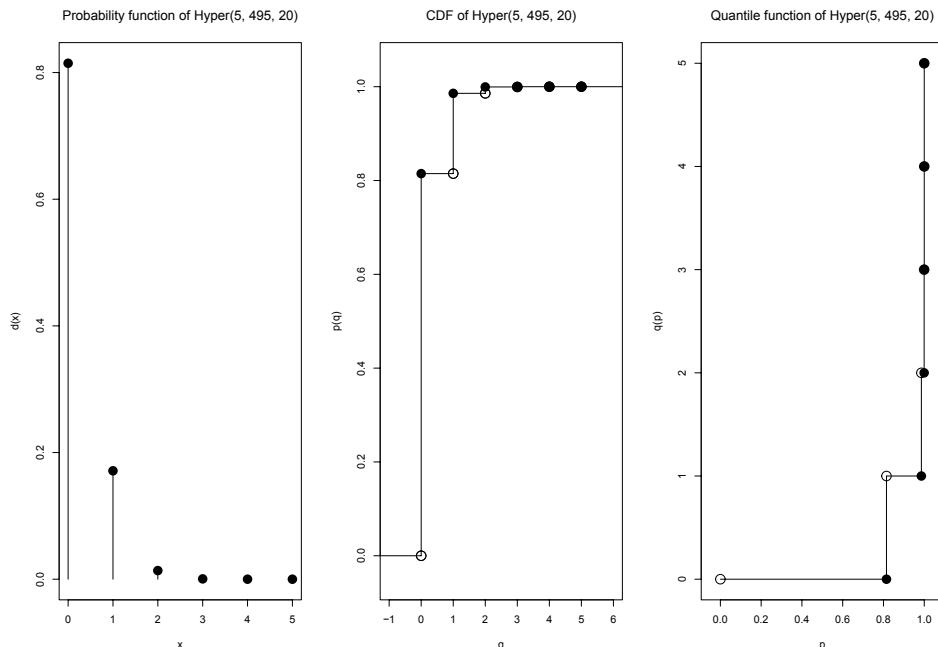
```
1 rhyper(10, m=6000, n=60000, k=50)
```

```
[1] 5 7 3 5 4 6 3 7 4 2
```

That is, every number represents the number of persons with diabetes in a random sample of size 50. We visualize the distribution Hyper(5, 495, 20) of the bulb example by means of package "distr" (Ruckdeschel et al. (2006)).

```
1 X <- Hyper(m=5, n=495, k=20)
```

```
2 plot(X)
```



As the following comparison for our bulb example shows, the hypergeometric and the binomial distribution already yield quite similar results, although the population is relatively small.

```
1 ## probability of 0, 1, 2, 3 defective bulbs
2 ## with replacement
3 dbinom(0:3, size=20, prob=0.01)
```

```
[1] 0.8179069376 0.1652337248 0.0158557615 0.0009609552
```

```
1 ## without replacement
2 dhyper(0:3, m=5, n=495, k=20)
```

```
[1] 0.8146893166 0.1711532178 0.0136348475 0.0005134461
```

With operator : we can quickly generate integer sequences; e.g.

```
1 0:3
```

```
[1] 0 1 2 3
```

```
1 8:11
```

```
[1] 8 9 10 11
```

```
1 -3:5
```

```
[1] -3 -2 -1 0 1 2 3 4 5
```



What do you want to do?

No matter what you want out of your future career, an employer with a broad range of operations in a load of countries will always be the ticket. Working within the Volvo Group means more than 100,000 friends and colleagues in more than 185 countries all over the world. We offer graduates great career opportunities – check out the Career section at our web site www.volvologroup.com. We look forward to getting to know you!

VOLVO
AB Volvo (publ)
www.volvologroup.com

VOLVO TRUCKS | RENAULT TRUCKS | MACK TRUCKS | VOLVO BUSES | VOLVO CONSTRUCTION EQUIPMENT | VOLVO PENTA | VOLVO AERO | VOLVO IT
VOLVO FINANCIAL SERVICES | VOLVO 3P | VOLVO POWERTRAIN | VOLVO PARTS | VOLVO TECHNOLOGY | VOLVO LOGISTICS | BUSINESS AREA ASIA

Download free eBooks at bookboon.com

Negative binomial distribution

Another important discrete distribution, which is in a certain way related to the binomial distribution, is the negative binomial distribution. We start with an introductory example showing possible applications of this distribution.

Example 4.9.

- a) We again consider the production of bulbs, where 1% of the bulbs is defective. Let X be the random variable that describes the number of functioning bulbs drawn (with replacement) until the first defective bulb is obtained. How likely is it that exactly the 20th bulb is the first defective bulb? That is, we first get 19 functioning bulbs leading to

$$d(19) = P(X = 19) = 0.99^{19} \cdot 0.01 = 0.8\% \quad (4.21)$$

- b) In 2014, the worldwide prevalence (disease frequency) of diabetes in adults was about 9% (WHO (2015b)). We conduct a trial and draw (with replacement) a sample of 250 persons. We need at least 20 persons with diabetes such that our trial has the required validity (power). How likely is it that we get the necessary number of diabetes patients at the latest with inclusion of the 250th person? That is, that we have to draw at most 230 persons without diabetes. The answer is, as we will see below,

$$p(230) = P(X \leq 230) = \sum_{l=0}^{230} \binom{l+20-1}{l} \cdot 0.91^l \cdot 0.09^{20} = 74.1\% \quad (4.22)$$

Thus, we will get 20 diabetes patients with a probability of about 74%.

We define the negative binomial distribution.

Definition 4.10 (Negative binomial distribution). *We consider a box (urn) with black and white balls, where the proportion of white balls is $p \in [0, 1]$. We randomly draw with replacement from the box until we have got $r \in \mathbb{N}$ white balls. Let X be the random variable describing the number $k \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$ of black balls that we obtain until we have got r white balls for the first time. The probability mass function of X is*

$$d(k) = P(X = k) = \binom{k+r-1}{k} (1-p)^k p^r \quad (4.23)$$

The distribution is called **negative binomial distribution** with parameters r and p , abbreviated by: $X \sim Nbinom(r, p)$.

We give some additional explanations.

Remark 4.11.

- a) The negative binomial distribution is a so-called **waiting time distribution**. We can apply it to specify how many unsuccessful attempts or eventless time intervals we have to wait until the required number of successes or events has occurred.
- b) The negative binomial distribution can be generalized such that parameter $r \in (0, \infty) \subset \mathbb{R}$.
- c) The negative binomial distribution is sometimes also called **Pascal distribution** or **Pólya distribution**. This mainly happens when the range of r is important. The name Pascal distribution is usually used if $r \in \mathbb{N}$ and the name Pólya distribution if $r \in (0, \infty)$. In case of $r = 1$, it is also called **geometric distribution**.
- d) Expectation and variance of $Nbinom(r, p)$ are

$$E(X) = r \frac{p}{1-p} \quad Var(X) = r \frac{p}{(1-p)^2} \quad (4.24)$$

In R, the negative binomial distribution is abbreviated by `nbinom` and the parameters are called `size` and `prob` as in case of the binomial distribution. Thus, we get functions `dnbnom`, `pnbnom`, `qnbnom`, and `rnbnom`. We compute the probabilities of Example 4.9 using R.

```
1 ## a) 20th bulb = 1st defective bulb
2 dnbnom(19, size = 1, prob = 0.01)
```

[1] 0.008261686

```
1 ## b) At most 250 persons to get 20 patients with diabetes
2 pnbnom(230, size = 20, prob = 0.09)
```

[1] 0.7407983

We can also use the quantile function in case of the diabetes example. We can for instance determine the sample size, which is needed, such that we achieve our goal of 20 diabetes patients with a given (high) probability. In such cases 90%, 95%, or even 99% are frequently used. In case of 95% certainty, we obtain

```
1 qnbnom(0.95, size = 20, prob = 0.09)
```

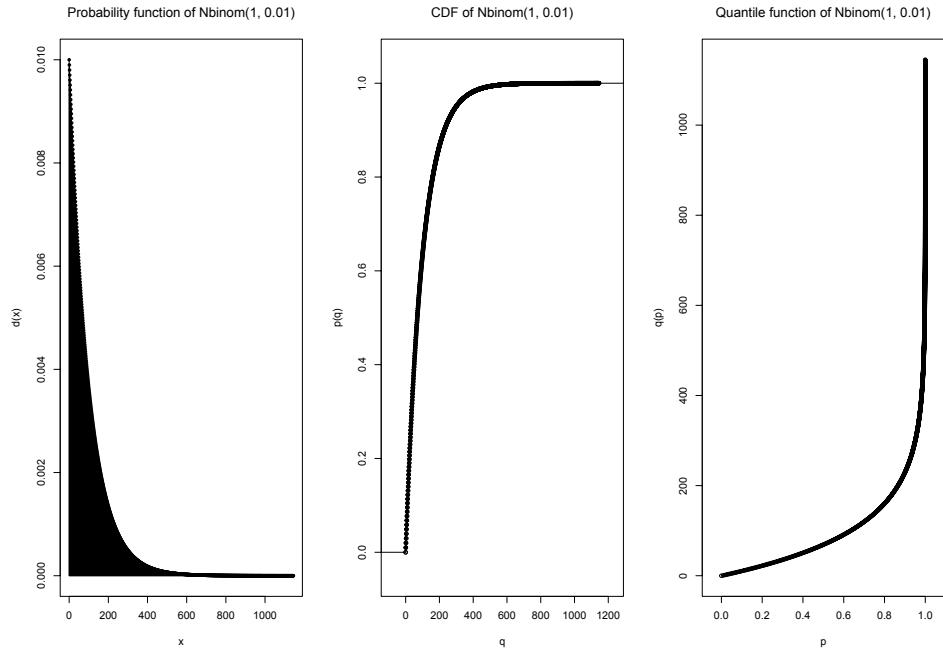
[1] 286

The number represents the number of persons without diabetes, i.e., in total we should randomly draw 306 persons. We simulate 10 diabetes trials.

```
1 rnbinom(10, size = 20, prob = 0.09)
[1] 235 204 174 183 225 165 261 117 258 134
```

Each of the numbers above states how many persons without diabetes had to be drawn to get the required number of 20 persons with diabetes. We visualize the negative binomial distribution of our bulb example by means of package "distr" (Ruckdeschel et al. (2006)).

```
1 X <- Nbinom(size = 1, prob = 0.01)
2 plot(X, cex.points = 0.75)
```



With the help of argument `cex.points` we reduce the size of the plotted points.

Poisson distribution

As last discrete distribution, we introduce the Poisson distribution, which has various applications.

Example 4.12.

- a) A conventional bulb today has an average (median) lifespan of 1000 hours. Assuming an exponential decrease of the number of functioning bulbs, we obtain

$$50\% = 0.5 = P(\text{Time till failure} > 1000h) = e^{-1000\lambda} \quad (4.25)$$

which leads to a failure rate per hour of about $\lambda = 0.0007$. We assume that we have 20 bulbs in our home that are on for 100 hours per month. Let X be the random variable describing the number of bulbs failing per month. How likely is it, that we have to change at least one bulb per month? We obtain

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-20 \cdot 100 \cdot 0.0007} = 1 - e^{-1.4} = 75.3\% \quad (4.26)$$

- b) The proportion of persons newly falling ill in a certain time period is called **incidence** or **incidence rate**. Finland has worldwide the highest incidence rate of type 1 diabetes for children up to an age of 15 years. On average, every year 55 of 100 000 children in that age newly fall ill with type 1 diabetes (Harjutsalo et al. (2013)), which corresponds to a rate of = 0.00055.

According to Wikipedia (2015c) there live about 900 000 children in that age in Finland; that is, on average we have to expect 495 new cases per year. Let X be the number of new cases per year. How likely is it, that there are more than 450 new cases in one year in Finland? As we will see below, we get

$$P(X > 450) = 1 - P(X \leq 450) = 1 - \sum_{k=0}^{450} \frac{495^k}{k!} e^{-495} = 97.8\% \quad (4.27)$$

We define the Poisson distribution.

Definition 4.13 (Poisson distribution). A random variable X follows a **Poisson distribution** with parameter $\lambda \in (0, \infty)$, if it has the following probability mass function

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k \in \mathbb{N}_0$$

abbreviated by: $X \sim \text{Pois}(\lambda)$

We give some additional explanations.

Remark 4.14.

- a) The parameter λ describes the number of events that we can expect on average in a predefined time period.
- b) The Poisson distribution has various applications and can also be used as an approximation of the binomial distribution. The approximation works well if the probability p of the event is small and the sample size n is large. In this case, we may use $\text{Pois}(np/)$ as approximation for $\text{Binom}(n, p)$. Therefore, the Poisson distribution is also called the distribution of rare events.
- c) Expectation and variance of $\text{Pois}(\lambda)$ are

$$E(X) = \lambda \quad \text{Var}(X) = \lambda \quad (4.28)$$

The Poisson distribution is abbreviated by `pois` in R leading to functions `dpois`, `ppois`, `qpois`, and `rpois`. We compute the probabilities of Example 4.12 using R.

```
1 ## a) no defective bulb
2 dpois(0, lambda = 1.4)
```

[1] 0.246597

By means of `ppois` and `lower.tail = FALSE` we obtain

```
1 ## a) at least one defective bulb  
2 ppois(0, lambda = 1.4, lower.tail = FALSE)
```

```
[1] 0.753403
```

```
1 ## b) at least 450 new cases  
2 ppois(450, lambda = 495, lower.tail = FALSE)
```

```
[1] 0.9784977
```

By applying the quantile function, we can determine how many bulbs per month we have to change at most with a high probability (here 99%).

```
1 qpois(0.99, lambda = 1.4)
```

```
[1] 5
```

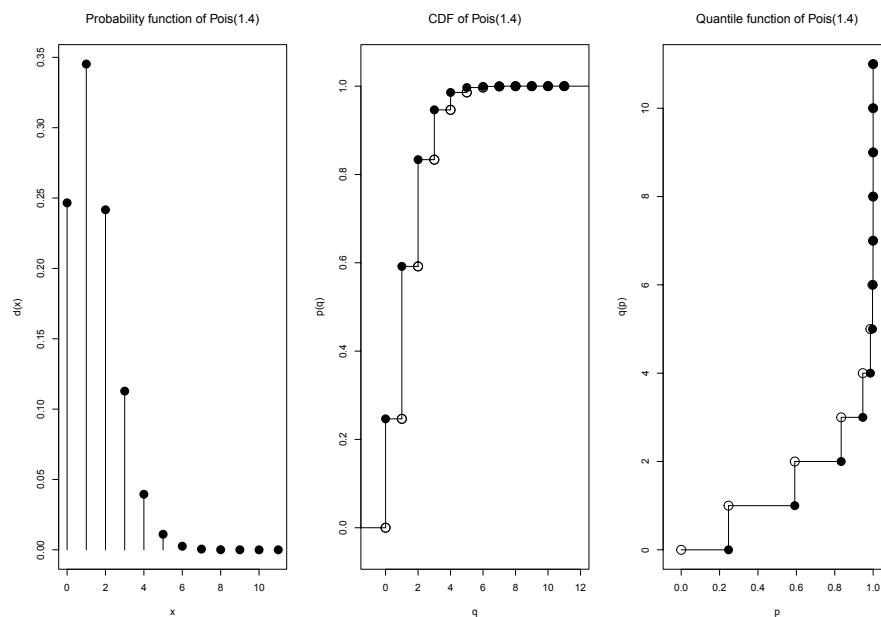


That is, a stock of five bulbs should suffice for more than one month with a high probability. We simulate the number of new cases of type 1 diabetes in Finland for ten years.

```
1 rpois(10, lambda = 495)
[1] 502 481 493 451 491 457 489 500 492 477
```

We visualize the distribution of the bulb example by means of package "distr" (Ruckdeschel et al. (2006)).

```
1 X <- Pois(lambda = 1.4)
2 plot(X)
```



4.2 Continuous Distributions

A random variable X , which may attain all values in an interval $I \subset \mathbb{R}$, is called **continuous random variable**.

Note:

This notion of continuity does not reflect a property of function X , i.e., the random variable X is not necessarily a continuous function. This notion of continuity – more precisely absolute continuity – is derived from the distribution of X . It means that the cumulative distribution function p of X is (almost everywhere) differentiable with derivative $d = p'$: respectively, p is the indefinite integral of d

$$p(x) = \int_{-\infty}^x d(t) dt \quad (4.29)$$

We may describe the **continuous probability distribution** of random variable X , or **continuous distribution** of X for short, by the so-called **probability density** or **density** d , where

$$d(x) \geq 0 \quad \text{for (almost) all } x \in \mathbb{R} \text{ and} \quad \int_{-\infty}^{\infty} d(x) dx = 1 \quad (4.30)$$

must hold. The probability $P(X \in [a, b])$ of some interval $[a, b] \in \mathbb{R}$ is given by

$$P(X \in (a, b]) = \int_a^b d(x) dx = p(b) - p(a) \quad (4.31)$$

Thus, the probability is nothing else but the area under the density curve. In particular, it follows

$$P(X = x) = 0 \quad (4.32)$$

i.e., single points possess probability 0. Consequentially, it holds

$$P(X \in (a, b)) = P(X \in (a, b]) = P(X \in [a, b)) = P(X \in [a, b]) \quad (4.33)$$

That is, it does not make any difference, if we consider open, semi-open or closed intervals. Similar to the discrete case, the quantile function in general reads

$$q(p) = \min \{x \in \mathbb{R} \mid p(x) \geq p\} \quad p \in [0, 1] \quad (4.34)$$

Every cumulative distribution function is monotonically increasing, if it is even strictly monotonically increasing, the quantile function is just the usual inverse function of the cumulative distribution function.

As in case of the computation of probabilities, one has to integrate to determine expectation and variance of continuous random variables. The expectation reads

$$E(X) = \int_{-\infty}^{\infty} x d(x) dx \quad (4.35)$$

and the variance is

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - E(X))^2 d(x) dx \quad (4.36)$$

Note:

Strictly speaking, there are no continuous random variables in practice, as all measurements that we make can only be done with restricted precision and hence, may at most produce finitely many results. Therefore, continuous random variables can be regarded as an abstract description of reality, in which the restricted precision of our measurements is ignored. Nevertheless, they are very useful and yield sufficiently precise descriptions in many practical applications.

Normal distribution

In the sequel, we will introduce some important continuous distributions. We start with the probably most important in statistics, the normal or Gaussian distribution.

Definition 4.15 (Normal distribution). A real random variable X follows a **Normal or Gaussian distribution** with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in (0, \infty)$, if it has the following density

$$d(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (4.37)$$

It is abbreviated by $X \sim \text{Norm}(\mu, \sigma^2)$.

We give some additional explanations.

Remark 4.16.

- a) The central role of the normal distribution follows from the fact that a superposition (sum) of independent factors, under quite weak assumptions can, at least approximately, be described by this distribution. This is a paraphrase of the statement of one of the most important theorems of probability theory, the **central limit theorem**.



www.sylvania.com

We do not reinvent
the wheel we reinvent
light.

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM

OSRAM
SYLVANIA

- b) In presence of a normal distribution, we can make quite precise statements about the probabilities of certain intervals using only its mean and standard deviation. It holds

$$\begin{aligned} P(X \in [\mu - \sigma, \mu + \sigma]) &= 68.3\% \\ P(X \in [\mu - 2\sigma, \mu + 2\sigma]) &= 95.4\% \\ P(X \in [\mu - 3\sigma, \mu + 3\sigma]) &= 99.7\% \end{aligned} \tag{4.38}$$

This yields the often handy and easy to remember **2σ rule**: Within a distance of 2σ around the mean (expectation) about 95% of the values are located. The 2σ rule is relatively robust and approximately holds for quite many distributions.

- c) The normal distributions also plays an important role in quality and process control. The name of one of the most famous quality management systems – Six Sigma – comes from the normal distribution. Thus, the goal of this system is an extremely low failure probability.
d) As the names of the parameters already indicate, the expectation and variance of Norm (μ, σ^2) are

$$E(X) = \mu \quad \text{Var}(X) = \sigma^2 \tag{4.39}$$

- e) If $X \sim \text{Norm} (\mu, \sigma^2)$ it holds

$$Z = \frac{X - \mu}{\sigma} \sim \text{Norm} (0, 1) \tag{4.40}$$

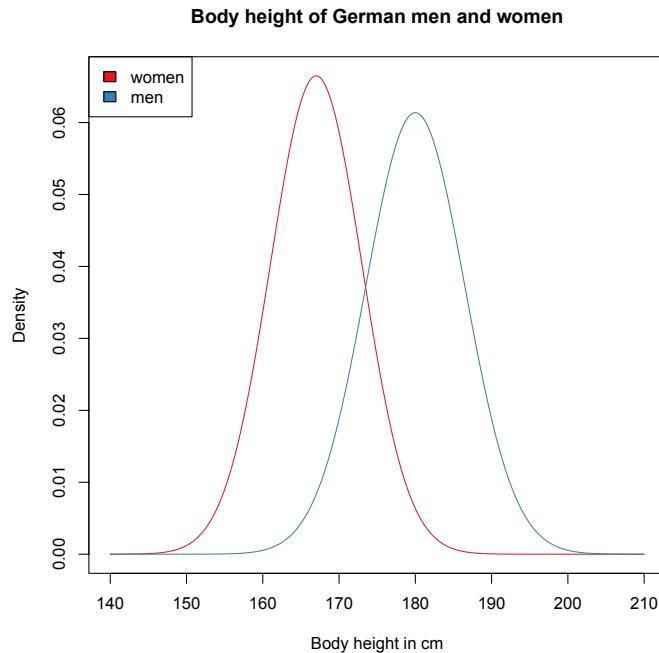
and one also calls Norm (0, 1) the **standard normal distribution**.

The normal distribution is abbreviated by `norm` in R leading to the functions `dnorm` (density), `pnorm`, `qnorm`, and `rnorm`. The names of the parameters are `mean` and `sd`. In the following example, we present two applications of the normal distribution.

Example 4.17.

- a) The body height of adults in a country can be well described by normal distributions. In case of the women in Germany, we get a mean of about 167 cm and a standard deviation of about 6.0 cm. In case of the men in Germany, the mean is about 180 cm and the standard deviation about 6.5 cm (Wikipedia (2015d)). We plot the density function of men and women using function `curve`.

```
1 curve(expr = dnorm(x, mean = 167, sd = 6.0), from = 140, to = 210, n = 501,
2   col = "#E41A1C", xlab = "Body height in cm", ylab = "Density",
3   main = "Body height of German men and women")
4 curve(expr = dnorm(x, mean = 180, sd = 6.5), from = 140, to = 210, n = 501,
5   add = TRUE, col = "#377EB8")
6 legend("topleft", legend = c("women", "men"), fill = c("#E41A1C", "#377EB8"))
```



The argument `expr` is the R expression that shall be plotted. With `from` and `to` one can specify the range of the x axis where expression `expr` is evaluated and drawn on a grid of `n` equidistant points. Finally, by using `add = TRUE` we can add further curves to an already existing plot. The proportion of women larger than 175 cm accordingly is

```
1 pnorm(175, mean = 167, sd = 6.0, lower.tail = FALSE)
```

```
[1] 0.09121122
```

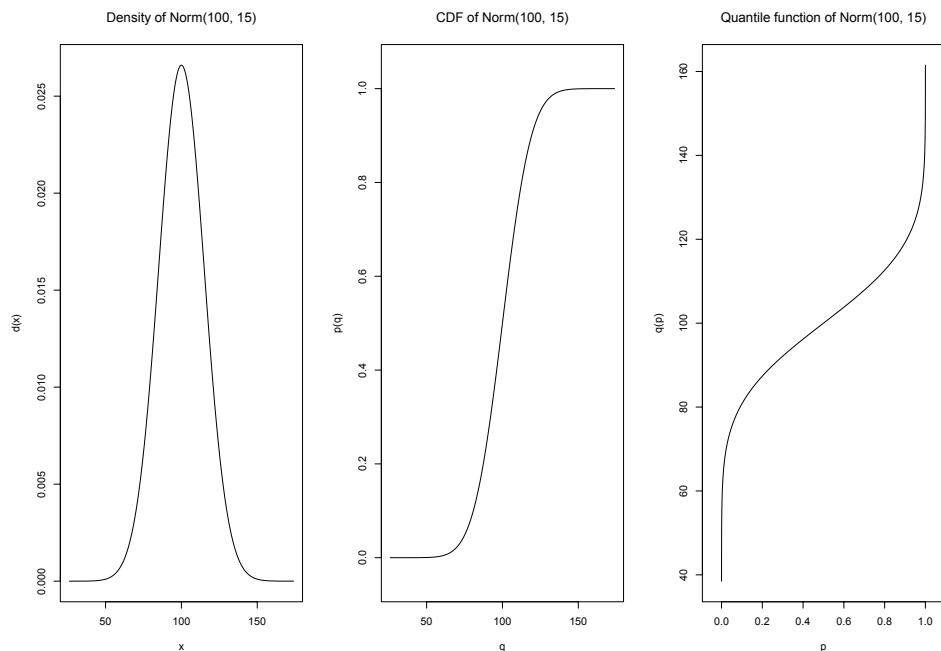
The tallest 5% of men are taller than

```
1 qnorm(0.95, mean = 180, sd = 6.5)
```

```
[1] 190.6915
```

- b) The intelligence quotient (IQ) can also very well be described by a normal distribution. The IQ scales have a mean of 100 and a standard deviation of 15 (Wikipedia (2015e)). We plot the respective normal distribution by means of package "`distr`" (Ruckdeschel et al. (2006)).

```
1 X <- Norm(mean = 100, sd = 15)
2 plot(X)
```



Thus, the 2σ rule states that about 5% of the population have an IQ score smaller than 70 or larger than 130.

CHALLENGING PERSPECTIVES

Internship opportunities

EADS unites a leading aircraft manufacturer, the world's largest helicopter supplier, a global leader in space programmes and a worldwide leader in global security solutions and systems to form Europe's largest defence and aerospace group. More than 140,000 people work at Airbus, Astrium, Cassidian and Eurocopter, in 90 locations globally, to deliver some of the industry's most exciting projects.

An **EADS internship** offers the chance to use your theoretical knowledge and apply it first-hand to real situations and assignments during your studies. Given a high level of responsibility, plenty of learning and development opportunities, and all the support you need, you will tackle interesting challenges on state-of-the-art products.

We welcome more than 5,000 interns every year across disciplines ranging from engineering, IT, procurement and finance, to strategy, customer support, marketing and sales. Positions are available in France, Germany, Spain and the UK.

To find out more and apply, visit www.jobs.eads.com. You can also find out more on our **EADS Careers Facebook page**.

AIRBUS **ASTRIUM** **CASSIDIAN** **EUROCOPTER**

EADS

Log-normal distribution

The second important is closely related with the normal distribution and is called log-normal distribution. We first give the definition.

Definition 4.18 (Log-normal distribution). A *real random variable X attaining only positive values follows a **log-normal distribution** with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in (0, \infty)$, if it has the following density*

$$d(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2} & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \quad (4.41)$$

It is abbreviated by $X \sim \text{Lnorm}(\mu, \sigma)$

We give some additional explanations.

Remark 4.19.

- a) The log-normal distribution occurs in many scientific disciplines. In particular, many biological processes happen on an exponential scale and thus many parameters in biology and medicine can be described by a log-normal distribution. That is, in a similar way as additive superpositions in the sense of the central limit theorem lead to a normal distribution, multiplicative superpositions lead to a log-normal distribution.
- b) If X is log-normal distributed, $\log(X)$ is normal distributed. In view of part (a) we can say, that a multiplicative superposition by applying the logarithm becomes an additive superposition.
- c) The parameters of the log-normal distribution are nothing else but the expectation and the variance of $\log(X)$. For the random variable X itself we get

$$E(X) = e^{\mu + \frac{\sigma^2}{2}} \quad \text{Var}(X) = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1) \quad (4.42)$$

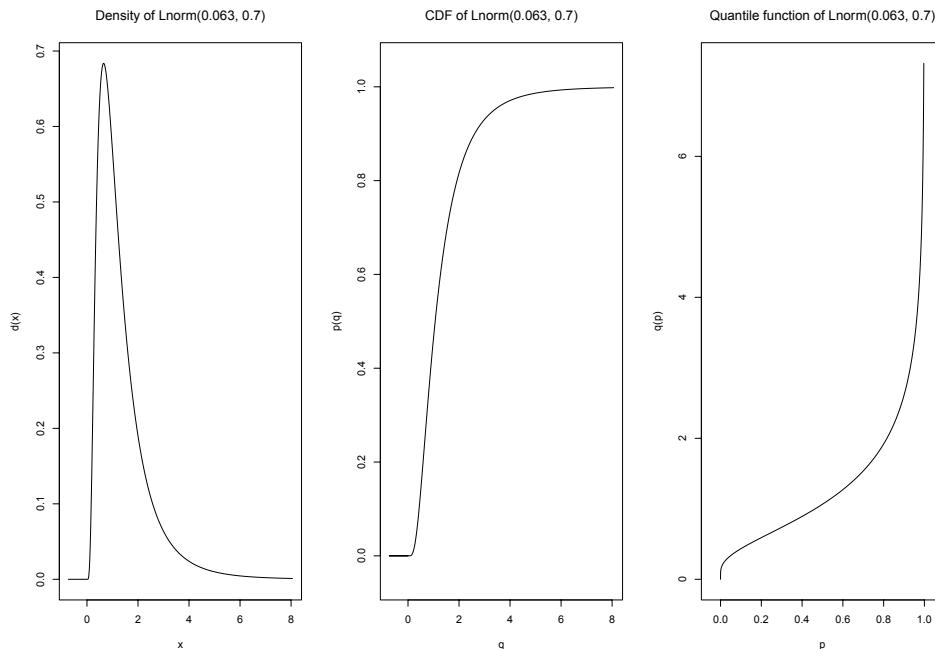
The log-normal distribution is abbreviated by `lnorm` in R. Accordingly, we obtain functions `dlnorm`, `plnorm`, `qlnorm`, and `rlnorm`, where the names of the parameters are `meanlog` and `sdlog`. We give an examples for an application of the log-normal distribution.

Example 4.20.

- a) For examining the thyroid function the concentration of thyrotropin (TSH) in the blood is analyzed. Its concentration in persons with normal thyroid function can be described by a lognormal distribution (Hamilton et al. (2008)). The declarations of the normal range vary especially with regard to the upper bound. In this example, we use a normal range of 0.27–4.2 $\mu\text{U/ml}$ for adults (Hagemann (2014)). By using the connection between log-normal and normal distribution, we can determine the distribution of TSH in persons with normal thyroid function. In addition, we use the information that the **normal range** of a parameter is always chosen such that 2.5% of the healthy persons may have lower or higher values, respectively (Wikipedia (2015f)). In case of the normal distribution, the normal range approximately corresponds to the 2σ interval.

After log-transforming, the normal range reads [-1.309,1.435]. Since the normal distribution is symmetric, the expectation must be the middle of this interval, i.e., $\mu = 0.063$. The length of the interval roughly is 4σ , more precisely it is 3.92σ . Starting with the interval length of 2.744, the division by 3.92 leads to $\sigma = 0.7$. Therefore, the distribution of log-TSH is Norm (0.063, 0.7²), thus TSH is Lnorm (0.063, 0.7) distributed. We plot the distribution of TSH by means of package "distr" (Ruckdeschel et al. (2006)).

```
1 X <- Lnorm(meanlog = 0.063, sdlog = 0.7)
2 plot(X)
```



- b) Several examples of applications of the log-normal distribution from various scientific disciplines are collected in Limpert et al. (2001) and Limpert and Stahel (2011). In particular, both articles give recommendations for handling log-normal distributed data in practice.

Gamma distribution

A very flexible distribution with many application is the so-called gamma distribution.

Definition 4.21 (Gamma distribution). A real random variable X attaining only positive values follows a **gamma distribution** with scale parameter $\sigma \in (0, \infty)$ and shape parameter $\alpha \in (0, \infty)$, if it has the following density

$$d(x) = \begin{cases} \frac{1}{\sigma^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\sigma}} & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \quad (4.43)$$

where the **gamma function** Γ is

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \quad (4.44)$$

It is abbreviated by $X \sim \text{Gamma}(\sigma, \alpha)$.

We give some additional explanations.



360°
thinking.

Deloitte.

Discover the truth at www.deloitte.ca/careers

© Deloitte & Touche LLP and affiliated entities.



Click on the ad to read more

Remark 4.22.

- a) The shape parameter makes the gamma distribution very flexible, thus it has many applications for instance in insurance mathematics, genetics or also medicine.
- b) An important special case of the gamma distribution is the exponential distribution, which is obtained for $\alpha = 1$. In addition, one usually uses the rate $\lambda = \frac{1}{\sigma}$ as parameter leading to the following density

$$d(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \quad (4.45)$$

It is abbreviated by $X \sim \text{Exp}(\lambda)$. One can consider it as the continuous counterpart of the geometric distribution, a special case of the negative binomial distribution. It describes the time between two events of a process, where the events occur continuously and independently from each other at a fixed rate. It is for instance used to estimate survival probabilities.

- c) If we simultaneously consider $k \in \mathbb{N}$ independent processes, whose events follow $\text{Exp}(\lambda)$, their sum follows a so-called **Erlang distribution**. The Erlang distribution itself is a special case of the gamma distribution, where it holds $\alpha = k$ and one usually uses the rate $\lambda = \frac{1}{\sigma}$ as second parameter as in case of the exponential distribution. Hence, the density reads

$$d(x) = \begin{cases} \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \quad (4.46)$$

The Erlang distribution can for example be used to model the time between calls in a call center, where the number of calls may for instance be described by a Poisson distribution.

- d) Another important special case of the gamma distribution is the χ^2 **distribution** with $n \in \mathbb{N}$ degrees of freedom, $\text{Chisq}(n)$ for short. It holds $\sigma = 2$ and $\alpha = \frac{n}{2}$. Thus, the density reads

$$d(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x} & \text{if } x > 0 \\ 0 & \text{sonst} \end{cases} \quad (4.47)$$

The χ^2 distribution also arises in the framework of the normal distribution as we will see later in this section.

- e) Expectation and variance of $X \sim \text{Gamma}(\sigma, \alpha)$ are

$$E(X) = \alpha\sigma \quad \text{Var}(X) = \alpha\sigma^2 \quad (4.48)$$

We introduce some applications of the gamma distribution. The gamma distribution is available in R in form of the functions `dgamma`, `pgamma`, `qgamma`, and `rgamma`, where the parameters are called `scale` and `shape`. The exponential distribution is provided by functions `dexp`, `pexp`, `qexp`, and `rexp` with parameter `rate`.

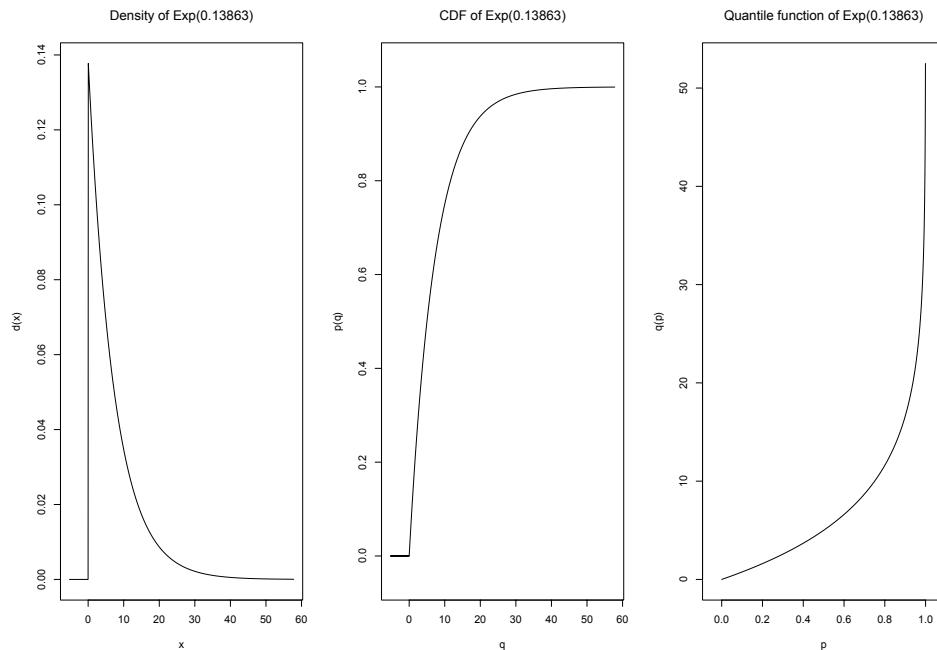
Example 4.23.

- a) A modern battery of a smart phone has a median life expectancy of five years. We use the exponential distribution to model the life expectancy, which yields

$$0.5 = P(X \leq 5\text{years}) = 1 - e^{-5\text{years} \cdot \lambda} \quad (4.49)$$

This leads to a failure rate per year of $\lambda = 0.13863$. We plot the distribution by means of package "distr" (Ruckdeschel et al. (2006)).

```
1 X <- Exp(rate = 0.13863)
2 plot(X)
```



Thus, how likely is it that the battery fails already in the first year?

```
1 pexp(1, rate = 0.13863)
```

```
[1] 0.1294499
```

That is, more than 10% of the batteries fail already in the first year. After how many years are 95% of the batteries out of order? We obtain

```
1 qexp(0.95, rate = 0.13863)
```

```
[1] 21.60955
```

That is, in the extreme case a battery may theoretically work for more than 20 years.

- b) The gamma distribution offers a way to model the hospital length of stay of a group of patients; e.g., all patients with a certain diagnosis or more precisely belonging to a certain DRG (diagnosis related group). Assuming a scale parameter of $\sigma = 5$ and a shape parameter of $\alpha = 1.8$ for a selected DRG, we obtain the following density, which we plot using function `curve`

```
1 curve(dgamma(x, scale = 5, shape = 1.8), from = 0, to = 30, n = 501,
2       xlab = "Hospital length of stay in days", ylab = "Density",
3       main = "A selected DRG")
```

SIMPLY CLEVER

ŠKODA



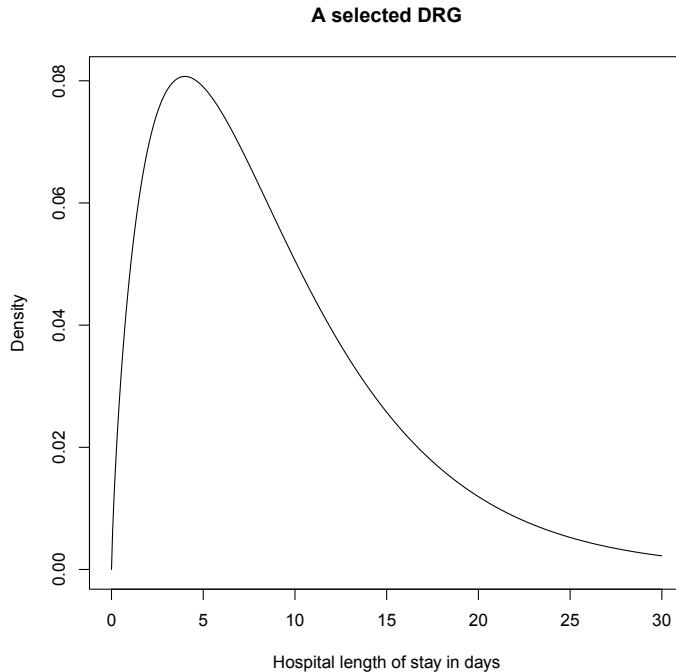
We will turn your CV into
an opportunity of a lifetime



Do you like cars? Would you like to be a part of a successful brand?
We will appreciate and reward both your enthusiasm and talent.
Send us your CV. You will be surprised where it can take you.

Send us your CV on
www.employerforlife.com





That is, most of the patients of this DRG will be discharged within a few days. But, it may happen that patients have to stay in the hospital for more than two weeks. How likely is it, that a randomly selected patient has to stay in the hospital for more than ten days? We get

```
1 pgamma(10, scale = 5, shape = 1.8, lower.tail = FALSE)
[1] 0.3472818
```

Thus, slightly more than one third of the patients have to stay for more than ten days. After how many days 99% of the patients have been discharged? We obtain

```
1 qgamma(0.99, scale = 5, shape = 1.8)
[1] 31.3043
```

That is, it happens only very rarely that a patient has to stay for more than one month.

Weibull distribution

If the failure rate changes over time, the so-called Weibull distribution offers a way to model the process. We start with its definition.

Definition 4.24 (Weibull distribution). A real random variable X attaining only positive values follows a **Weibull distribution** with scale parameter $\sigma \in (0, \infty)$ and shape parameter $\alpha \in (0, \infty)$, if it has the following density

$$d(x) = \begin{cases} \frac{\alpha}{\sigma} \left(\frac{x}{\sigma} \right)^{\alpha-1} e^{-\left(\frac{x}{\sigma}\right)^\alpha} & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \quad (4.50)$$

It is abbreviated by $X \sim \text{Weibull}(\sigma, \alpha)$

We give some additional explanations.

Remark 4.25.

- a) The Weibull distribution plays an important role in the reliability analysis of parts and components for instance in the automobile industry. In contrast to the exponential distribution, the shape parameter offers a possibility to model also aging.
- b) The Weibull distribution belongs to the class of **extreme value distributions**, more precisely it is an extreme value distribution of Typ III. By the theorem of Fisher-Tippett-Gnedenko, this distribution, under certain assumptions, arises as the maximum of independent random variables.
- c) In case $\alpha = 1$, the Weibull distribution is identical to the exponential distribution.
- d) Expectation and variance are

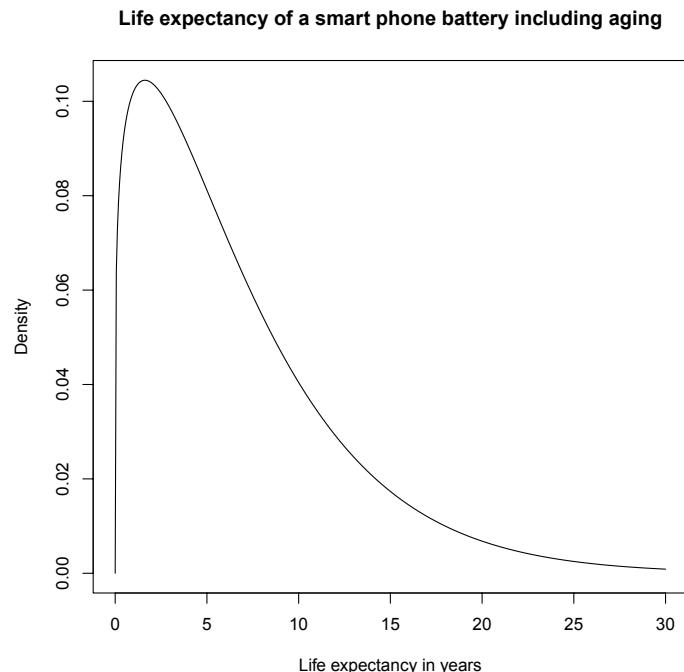
$$E(X) = \sigma \Gamma \left(1 + \frac{1}{\alpha} \right) \quad \text{Var}(X) = \sigma^2 \left(\Gamma \left(1 + \frac{2}{\alpha} \right) - \Gamma \left(1 + \frac{1}{\alpha} \right)^2 \right) \quad (4.51)$$

We introduce some applications.

Example 4.26.

- a) We again consider the battery of a modern smart phone with a failure rate of 0.13863 as in Example 4.23 (a), i.e. $\sigma = \frac{1}{0.13863}$. In addition, we assume that its aging can be described by a shape parameter of $\alpha = 1.2$. We plot the distribution using function curve

```
1 curve(dweibull(x, scale = 1/0.13863, shape = 1.2), from = 0, to = 30, n = 501,
2   xlab = "Life expectancy in years", ylab = "Density",
3   main = "Life expectancy of a smart phone battery including aging")
```



Linköping University –
innovative, highly ranked,
European

Interested in Computer Science? Kick-start your career
with an English-taught master's degree.

→ [Click here!](#)



There are less defective batteries in the first year than in case of the exponential distribution

```
1 pweibull(1, scale = 1/0.13863, shape = 1.2)
```

```
[1] 0.08914775
```

However, it takes less time until 95% of the batteries are out of order

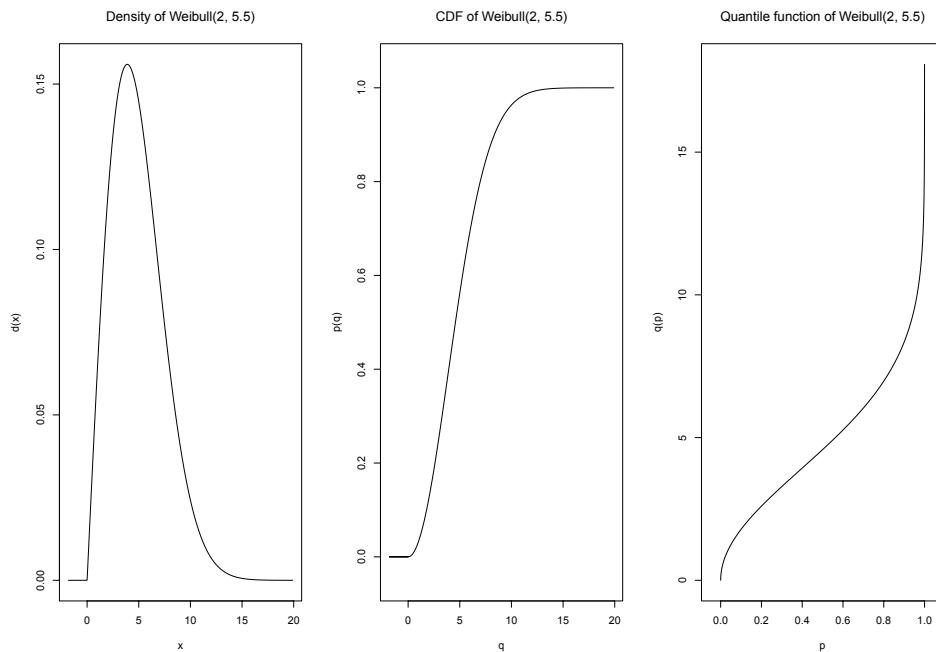
```
1 qweibull(0.95, scale = 1/0.13863, shape = 1.2)
```

```
[1] 17.99818
```

i.e., only about 18 years.

- b) The Weibull distribution is also used to model wind speed. We assume that the maximum wind speeds (in $\frac{m}{s}$) per day at a selected place may be described by a Weibull distribution with $\sigma = 5.5$ and $\alpha = 2$. We plot the distribution by means of package "distr" (Ruckdeschel et al. (2006)).

```
1 X <- Weibull(scale = 5.5, shape = 2)
2 plot(X)
```



We get as median wind speed

```
1 qweibull(0.5, scale = 5.5, shape = 2)
```

```
[1] 4.57905
```

which is a gentle breeze. How likely is at least a strong breeze, i.e., a wind speed of at least $11 \frac{m}{s}$? We obtain

```
1 pweibull(11, scale = 5.5, shape = 2, lower.tail = FALSE)
```

```
[1] 0.01831564
```

That is, it happens only in about 2% of the days.

χ^2 , t and F distribution

Finally, we introduce some continuous distributions that arise in the context of the normal distribution and play an important role in inferential statistics. We first give the definitions.

Definition 4.27.

- a) A real random variable X attaining only positive values follows a **χ^2 distribution** with $n \in \mathbb{N}$ degrees of freedom, if it has the following density

$$d(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{1}{2}x} x^{(\frac{n}{2}-1)} & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \quad (4.52)$$

It is abbreviated by $X \sim \text{Chisq}(n)$.

- b) A real random variable X follows a **t distribution** with $n \in \mathbb{N}$ degrees of freedom, if it has the following density

$$d(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2}) \sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (4.53)$$

It is abbreviated by $X \sim t(n)$

- c) A real random variable X attaining only positive values follows an **F distribution** with $m \in \mathbb{N}$ and $n \in \mathbb{N}$ degrees of freedom, if it has the following density

$$d(x) = \begin{cases} \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} n^{\frac{n}{2}} m^{\frac{m}{2}} \frac{x^{\frac{n}{2}-1}}{(m+nx)^{\frac{n+m}{2}}} & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \quad (4.54)$$

It is abbreviated by $X \sim F(m, n)$.

We give some additional explanations.

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com



Month 16

I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements





Download free eBooks at bookboon.com

134

 **Click on the ad to read more**

Remark 4.28.

- a) The χ^2 distributions arises as the sum of the square of n independent standard normal random variables. In inferential statistics, the distribution for instance occurs in connection with estimating the variance.
- b) Let Z be some standard normal random variable and Y an independent $\text{Chisq}(n)$ distributed random variable. Then, it holds

$$\sqrt{n} \frac{Z}{\sqrt{Y}} \sim t(n) \quad (4.55)$$

The distribution arises in inferential statistics for example by considering standardized arithmetic means.

- c) Let $X \sim \text{Chisq}(m)$ and $Y \sim \text{Chisq}(n)$ be some independent random variables. Then, it holds

$$\frac{n \cdot X}{m \cdot Y} \sim F(m, n) \quad (4.56)$$

The distribution arises in inferential statistics for instance by investigating the ratio of two variances.

Note:

Of course, there are many more probability distributions, which can be used as models for various applications. In particular, these basic distributions can be applied for constructing more complex models such as regression models.

Table 4.1 includes important notions from statistics and their counterparts in probability theory.

Statistics	Probability theory
attribute/variable	random variable
levels	possible values of a random variable
relative frequency	probability
frequency distribution	probability mass function
density estimation	(probability) density
empirical cumulative distribution function	cumulative distribution function
(sample) quantile	quantile
arithmetic mean	expectation
(sample) variance	variance

Table 4.1: Notions from statistics and their counterparts in probability theory.

There are also counterparts to (sample) correlation and covariance in probability theory. For their definition one has to consider the common distribution of two random variables, which goes beyond this introductory book.

4.3 Exercises

Please always describe and briefly explain your results.

1. Plot the distribution $\text{Binom}(20, p)$ for $p \in \{0.1, 0.2, \dots, 0.9\}$ by means of package "distr" (Ruckdeschel et al. (2006)).
2. Determine expectation and variance of $\text{Binom}(2, p)$ without using the explicit formulas for expectation and variance.
3. People with blood group 0-negative are universal donors, where about 7% of the humans have this blood type. Let us assume you conduct a trial, in which 20 persons are randomly selected. How likely is it that there are at least three universal donors in the sample? Use the binomial distribution.
4. In a certain hospital the median birth rate is 1.8 births per hour. How many delivery rooms does the hospital need, such that each birth is in a delivery room with 95% probability? Use the Poisson distribution.
5. An oil company conducts a geological study in a certain region where it is drilled for oil at randomly selected positions. Let us assume that the probability of finding oil in the selected region is 20%. How likely is it that the company has to drill at least five times until the first oil find? How often the company has to drill to find oil twice with 99% certainty? Apply the negative binomial distribution.
6. Plot the distribution $\text{Gamma}(1, \alpha)$ for $\alpha \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$ by means of package "distr" (Ruckdeschel et al. (2006)). The function to generate gamma distributed random variables is called `Gammad`.
7. Determine expectation and variance of $\text{Exp}(1)$ without using the explicit formulas for expectation and variance.
8. The expected birth weight of healthy boys is $\mu = 3.35$ kg with a standard deviation of $\sigma = 0.43$ kg. How likely is it that a healthy boy with a birth weight of less than 3 kg is born? What is the normal range of the birth weight of boys? Apply the normal distribution.

9. You want to investigate the impact of gamma rays and conduct an animal experiment with mice. The mice are exposed to a radiation of 2.4 Gray. The survival time of the mice in weeks can be described by a gamma distribution with parameters $\sigma = 15$ and $\alpha = 8$. How likely is it that a randomly chosen mouse lives between 50 and 100 weeks? How many weeks does it take until 95% of the mice have died?
10. The median life time of a common bulb today is 1000 hours. We assume that we can describe the life time by a Weibull distribution with $\sigma = 1250$ and $\alpha = 1.8$. How likely is it that a bulb is defective already in the first 100 hours? After how many hours are 99% of the bulbs defective?

**"I studied English for 16 years but...
...I finally learned to speak it in just six lessons"**

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

Download free eBooks at bookboon.com

137

Click on the ad to read more

5 Estimation

The chapter is about estimating parameters of simple parametric models. It covers the following topics:

- Issues of inferential statistics
- Importance of estimation in the framework of inferential statistics
- Parametric probability models
- Point estimator, estimator
- Unbiasedness, efficiency, consistency
- Maximum likelihood estimator (abbreviated: ML estimator)
- Quantile-quantile plot (abbreviated: qq plot)
- Minimum distance estimator (abbreviated: MD estimator)
- Kolmogorov(-Smirnov)-MD estimator (abbreviated: KS-MD estimator)
- Cramér-von-Mises-MD estimator (abbreviated: CvM-MD estimator)
- Interval estimator, confidence interval
- Confidence interval for arithmetic mean and standard deviation
- Exact confidence intervals, asymptotic confidence intervals
- Confidence intervals for ML estimators
- Continuity correction, finite-sample correction
- Confidence intervals for median and MAD
- Confidence intervals for CvM-MD estimator

The R code of this chapter is included in file `Estimation.R`, which can be downloaded from my website (link: www.stamats.de/RCodeEN.zip). For experimenting with your own R code, it is advisable to generate your own R script as explained at the beginning of Chapter 2.

5.1 Introduction

This introduction provides a brief example to make clear, which questions we can address applying inferential statistics.

We consider a coin and for simplicity label the sides with 0 and 1, where we exclude the possibility that after tossing the coin might land on its edge. We are interested in the question:

Is it a fair coin?

Here, fair means that both sides of the coin occur with equal probability. We can describe the coin toss using the Bernoulli distribution $\text{Bernoulli}(p)$, where p is the probability that side 1 is tossed. By means of this probability model, we can state the question more precisely and obtain:

Is the probability of side 1 equal to 50%, abbreviated: $p = 0.5$?

How can we address this question? We could test the coin by a very detailed materials analysis. However, this surely would be very costly and only possible with an appropriate technical equipment. Certainly, a random experiment is faster and simpler: we toss the coin several times and record the results. The results of this random experiment are our sample, which is the basis for our decision by means of statistical procedures.

Before we conduct this random experiment, there are some things to clarify:

I: How often should we toss the coin to get a most reliable result?

II: How do we summarize the results such that we may infer the actual probability p of side 1 in a most optimal way?

III: Is the observed count of side 1 in the range of the expected frequency of a fair coin or is it too small or too large?

The answers of inferential statistics to these questions are:

Ad I: We can perform a so-called sample size calculation using a confidence interval (see Section 5.3) or a statistical test (see Chapter 6). With these procedures, we can determine the number of replications in such a way that we can decide, if the coin is fair with a given certainty.

Ad II: By means of point estimators (see Section 5.2) we can summarize the observed values. In case of the coin, the observed relative frequency of side 1 can be compared with the theoretical value ($p = 0.5$).

Ad III: We can again use confidence intervals or statistical tests to correctly answer this question with a given high certainty. For instance, if $p = 0.5$ is covered by the computed confidence interval, we consider the coin as a fair coin.

Note:

Statistics is not able to absolutely answer a question. The possibility of a wrong decision can never be excluded. Some scientists even believe that most of the published research results are wrong (Ioannidis (2005)). This criticism may be approached with a proper methodology, sufficiently large trials, a careful application of statistical procedures, and a cautious interpretation of the results.

5.2 Point Estimation

In this section, we want to determine the unknown parameters of simple parametric models. This procedure is called estimation, more precisely we are looking for point estimators of the unknown parameters. We first define the notions parametric model and point estimator.

Definition 5.1 (Parametric model, point estimator).

- a) A **parametric model** is a set $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ of probability distributions, where the elements of \mathcal{P} are uniquely identifiable by their parameter $\theta \in \Theta \subset \mathbb{R}^k$ ($k \in \mathbb{N}$). This is also called a **parametric family**.
- b) Let $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ be a parametric family of probability distributions, where $\Theta \subset \mathbb{R}^k$ ($k \in \mathbb{N}$) is the set of all possible parameters. Furthermore, let x_1, \dots, x_n be a **representative sample** of size $n \in \mathbb{N}$ from some element $P_\theta \in \mathcal{P}$ (θ unknown). Then, a **point estimator** or **estimator** S_n is a random variable

$$S_n : \mathbb{R}^n \rightarrow \Theta, (x_1, \dots, x_n) \mapsto S_n(x_1, \dots, x_n) =: \hat{\theta} \quad (5.1)$$

where $\hat{\theta}$ is the **point estimation** or **estimation** of θ .

We give some additional explanations.



The advertisement features a stylized Earth graphic with blue oceans and green continents, overlaid with white clouds. A horizontal line with arrows points from the left side of the Earth towards the right, ending at a small orange dot on the continent. The text on the right side reads:

In the past four years we have drilled
89,000 km
 That's more than **twice** around the world.

Who are we?
 We are the world's largest oilfield services company¹. Working globally—often in remote and challenging locations—we invent, design, engineer, and apply technology to help our customers find and produce oil and gas safely.

Who are we looking for?
 Every year, we need thousands of graduates to begin dynamic careers in the following domains:
 ■ Engineering, Research and Operations
 ■ Geoscience and Petrotechnical
 ■ Commercial and Business

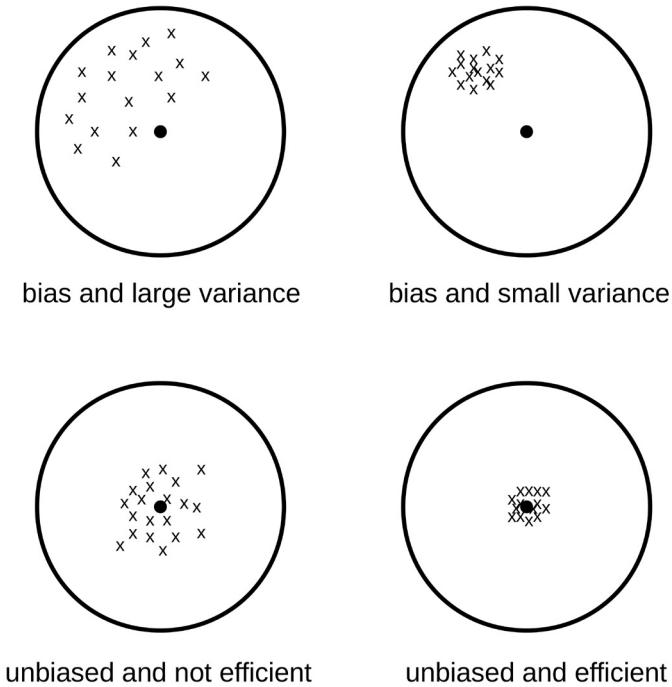
What will you be?

Schlumberger

¹Based on Fortune 500 ranking 2011. Copyright © 2015 Schlumberger. All rights reserved.

Remark 5.2.

- a) The notion *parametric family* implies that the elements of the set may be identified by their parameters. Formally, there is a function that maps a given θ to a certain P_θ and the mapping is unique.
- b) The observations of a representative sample correspond to realizations of independent random variables X_1, \dots, X_n , where it holds $X_i \sim P_\theta$ ($i = 1, \dots, n$). Therefore, the random variables are also called *independent and identical distributed (iid)*.
- c) A point estimator S_n is a random variable, i.e., a random function. Consequentially, an estimator has a certain distribution that depends on the unknown distribution P_θ . The quality of an estimator is usually assessed by $E(S_n)$ und $\text{Var}(S_n)$. If $E(S_n) = \theta$, the estimator is called **bias-free** or **unbiased**. It means that the estimator in average estimates the true parameter. If $\text{Var}(S_n)$ is additionally minimal, the estimator is called **efficient**. That is, there is no unbiased estimator that is able to estimate θ more accurately. Instead of unbiasedness, one often has to be satisfied with the so-called **consistency**, which means that the estimator with increasing sample size more and more approaches (in a probability theoretic sense) the true (unknown) parameter; i.e., $\lim_{n \rightarrow \infty} S_n = \theta$ (in a probability theoretic sense). Figure 5.1 illustrates the notions unbiased and efficient, where the center of the circle corresponds to the true (unknown) parameter.

**Figure 5.1:** Illustration of unbiased and efficient.

In the following example, we introduce some unbiased and efficient estimators.

Example 5.3.

- a) We consider the probability model $\{\text{Bernoulli}(p) \mid p \in (0, 1)\}$. Then, the relative frequency is an unbiased and efficient estimator of the unknown probability p .
- b) Let the probability model $\{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}\}$ be given, where $\sigma^2 \in (0, \infty)$ is known. Then, the arithmetic mean is an unbiased and efficient estimator of the unknown expectation μ .
- c) The situation becomes somewhat more complicated in case of the model $\{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma \in (0, \infty)\}$. The sample variance is a possible estimator for the unknown variance σ^2 , but it is not unbiased. The bias is $-\frac{1}{n}\sigma^2$. We obtain an unbiased estimator by using the standardization $\ln \frac{1}{n-1}$; i.e.

$$\tilde{S}_n(x_1, \dots, x_n) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \text{AM}(x_1, \dots, x_n))^2 \quad (5.2)$$

Therefore, avoiding a bias is the reason why one usually uses $\frac{1}{n-1}$ instead of $\frac{1}{n}$ for computing the empirical variance. Regarding the accuracy of the estimation, the variance of the (true) sample variance is smaller than the variance of \tilde{S}_n .

We use our ICU dataset und want to estimate the prevalence (disease frequency) of liver failure on the ICU. Please, import the dataset as described in Section 2.3, if you have not done this already. There you also find more information about the data. In contrast to descriptive statistics, it is now necessary for the validity of the results that the 500 ICU patients were randomly and representatively selected from the ICU population. We compute the relative frequency as described in Section 2.4.1.

```
1 ## unbiased and efficient
2 table(ICUData$liver.failure)/nrow(ICUData)
```

0	1
0.96	0.04

That is, 4% of the randomly selected ICU patients had a liver failure. We now regard this as an estimate for all ICU patients and later we will further ensure the result. Therefore, a possible model for the prevalence of liver failure on the ICU is $\text{Bernoulli}(0.04)$.

The analysis in Section 2.5.1 suggests that the maximum body temperature of ICU patients – except for strongly undercooled (hypothermic) patients such as patient 398 – is quite well described by a normal distribution. We estimate expectation and variance.

```
1 ## unbiased and efficient
2 mean(ICUData$temperature [-398])
```

```
[1] 37.72044
```

```
1 ## unbiased
2 sd(ICUData$temperature [-398])
```

```
[1] 1.173187
```

The results are identical to Section 2.5.1. However, we do not longer use these values for describing the sample, but as parameters of a probability model, which describes the underlying population.

Note:

For interpreting the result and inferring to the ICU population, it is of crucial importance, whether we want to include strongly undercooled patients such as patient 398. If this is not the case, we can use $\text{Norm}(37.7, 1.2^2)$ as a model for the maximum body temperature. Otherwise, we have to understand that we can not describe the maximum body temperature by a normal distribution as such an extreme temperature as 9.1°C is practically impossible.

A circular graphic titled "WHILE YOU WERE SLEEPING..." featuring a collage of international business scenes. The segments include: "INDIAN NUCLEAR CORP", "RUSSIAN ENTREPRENEUR", "LUXURY GOODS MARKET LONDON", "CHINA PLAYS TROUBLE", "CHINESE PREMIUM FURNITURE", "AFRICA HELPS OUTSOURCE", "AFRICA FIRM", and "INDIAN RIVER". Below the graphic is the URL www.fuqua.duke.edu/whileyouweresleeping.

DUKE
THE FUQUA
SCHOOL
OF BUSINESS

We apply the estimated model and compute the probability that the maximum body temperature is less than 10°C. We get

```
1 pnorm(10, mean = 37.7, sd = 1.2)
```

```
[1] 3.404114e-118
```

Next, we will address the question how to find optimal or at least good estimators. This is also called **estimator construction**. The probably most frequently applied principle is maximum likelihood, which is defined as follows.

Definition 5.4 (Maximum likelihood estimator). Let $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ ($k \in \mathbb{N}$), be some probability model with probability mass function or density d_θ . Furthermore, let x_1, \dots, x_n be realizations of independent and P_θ distributed random variables X_1, \dots, X_n . Then, the **likelihood function** is

$$L(\theta) = \prod_{i=1}^n d_\theta(x_i) \quad (5.3)$$

and the **maximum likelihood estimator** (abbreviated: **ML estimator**) for θ is the position of the maximum of $L(\theta)$.

We give some additional explanations.

Remark 5.5.

- a) In case of the ML estimator, θ is chosen such that the observed data has the maximum possible probability in the assumed probability model.
- b) The ML construction principle is generally applicable and usually leads to an (asymptotically) unbiased and efficient estimator. However, there are also probability models, where it is not applicable.
- c) In simple cases, the ML estimator can be determined by direct analytical calculations. The numerical computation of the likelihood function is numerically difficult in practice (product of many small numbers) and usually the so-called **log-likelihood function** is used

$$l(\theta) = \ln(L(\theta)) = \sum_{i=1}^n \ln(d_\theta(x_i)) \quad (5.4)$$

where the position of the maximum is identical to $L(\theta)$. This simple “trick” clearly simplifies the numerical computations and leads to more stable results.

- d) There are several R packages that include functions to compute ML estimators. For simple probability models one can for example apply the packages "stats4" (R Core Team (2015a)), "MASS" (Venables and Ripley (2002)), "fitdistrplus" (Delignette-Muller and Dutang (2015)), or "distrMod" (Kohl and Ruckdeschel (2010)).

We present some examples of ML estimators.

Example 5.6.

- a) In case of the simple Bernoulli model $\mathcal{P} = \{\text{Bernoulli}(p) \mid p \in (0, 1)\}$, the ML estimator can explicitly be determined via the first derivative of the log-likelihood function. The likelihood function reads

$$L(p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \quad (5.5)$$

and thus the log-likelihood function is

$$l(p) = \sum_{i=1}^n \ln [p^{x_i} (1-p)^{1-x_i}] = \sum_{i=1}^n [x_i \ln(p) + (1-x_i) \ln(1-p)] \quad (5.6)$$

We calculate the derivative of the log-likelihood function and obtain

$$\begin{aligned} l'(p) &= \frac{d}{dp} l(p) = \sum_{i=1}^n \left[\frac{x_i}{p} - \frac{1-x_i}{1-p} \right] = \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \left[n - \sum_{i=1}^n x_i \right] \\ &= -\frac{n}{1-p} + \frac{(1-p)+p}{p(1-p)} \sum_{i=1}^n x_i \\ &= -\frac{n}{1-p} + \frac{1}{p(1-p)} \sum_{i=1}^n x_i \end{aligned} \quad (5.7)$$

Setting the first derivative equal to zero ($l'(p) = 0$) yields

$$\frac{n}{1-p} = \frac{1}{p(1-p)} \sum_{i=1}^n x_i \quad \Leftrightarrow \quad p = \frac{1}{n} \sum_{i=1}^n x_i \quad (5.8)$$

As x_i can only take the values 0 and 1, the arithmetic mean of the x_i is nothing else but the relative frequency of 1.

- b) Normal distribution model: The ML estimator for the expectation is the arithmetic mean, for the variance it is the sample variance (i.e. standardization $\frac{1}{n}$).
c) Poisson model: The ML estimator is the arithmetic mean.
d) Exponential model: The ML estimator is the inverse of the arithmetic mean.

Instead of the functions `mean` and `sd`, we apply function `fitdistr` of package "MASS" (Venables and Ripley (2002)) to determine the ML estimator of the maximum body temperature. We need not to install package "MASS", since it belongs to the group of recommended packages and thus is included in the standard installation of R. We use the normal distribution model and exclude patient 398.

```
1 library(MASS)
2 fitdistr(ICUData$temperature[-398], densfun = "normal")
```

mean	sd
37.72044088	1.17201119
(0.05246643)	(0.03709937)

As the ML estimator for the variance includes the standardization $\frac{1}{n}$, the result slightly differs from the result of function `sd`. In addition to the estimates, we get some additional output showing the **standard errors** of the estimates; see Section 5.3 for more details.

Alternatively, we use package "distrMod" (Kohl and Ruckdeschel (2010)), which is derived from package "distr" (Ruckdeschel et al. (2006)). We can install the package either with

```
1 install.packages("distrMod")
```

Excellent Economics and Business programmes at:


/
**university of
groningen**


**“The perfect start
of a successful,
international career.”**

www.rug.nl/feb/education

CLICK HERE
to discover why both socially
and academically the University
of Groningen is one of the best
places for a student to be



or the window *Packages* of RStudio (cf. Section 2.4.1). We load the package, where the estimation proceeds in two steps. First, we define the probability model and then we estimate the parameters of the generated model by means of function `MLEstimator`.

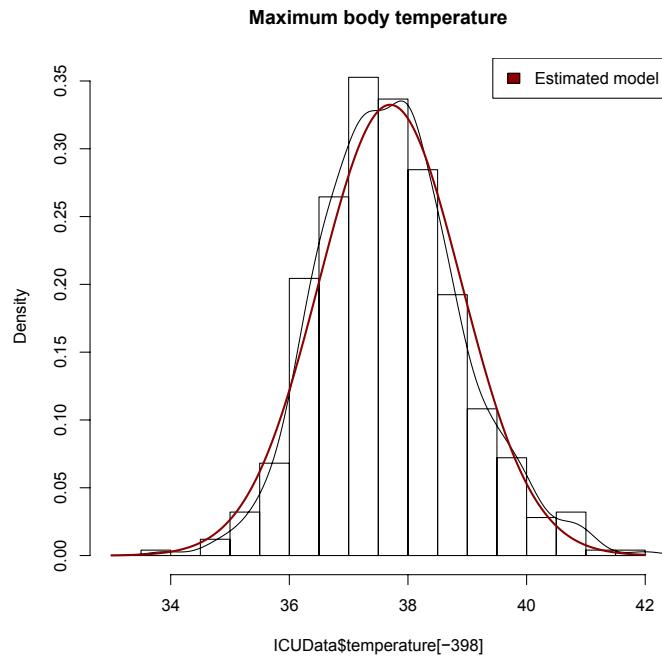
```
1 library(distrMod)
2 ## Define probability model
3 model <- NormLocationScaleFamily()
4 ## Estimate parameters by ML
5 MLEstimator(ICUData$temperature[-398], model)
```

```
Evaluations of Maximum likelihood estimate:
-----
An object of class "Estimate"
generated by call
  MLEstimator(x = ICUData$temperature[-398], ParamFamily = model)
samplesize: 499
estimate:
  mean           sd
37.72044088   1.17201119
( 0.05246643) ( 0.03709937)
asymptotic (co)variance (multiplied with samplesize):
  mean           sd
mean 1.37361 0.0000000
sd   0.00000 0.6868051
Criterion:
negative log-likelihood
787.2522
```

The normal distribution model is called `NormLocationScaleFamily`, because it is more generally a **location and scale model**, since expectation (location parameter) as well as variance (dispersion or scale parameter) must be estimated. The result for the ML estimate is identical to the result of `fitdistr`. The abstract approach of package "distrMod" (Kohl and Ruckdeschel (2010)) enables the computation of several additional values, which can for instance be used to compute confidence intervals (see Section 5.3).

We plot the data (without patient 398) and the estimated model by means of a histogram in combination with a density plot.

```
1 hist(ICUData$temperature[-398], breaks = seq(from = 33, to = 42, by = 0.5),
2       main = "Maximum body temperature", ylab = "Density", freq = FALSE)
3 lines(density(ICUData$temperature[-398]))
4 curve(dnorm(x, mean = 37.7, sd = 1.2), col = "darkred", from = 33, to = 42,
5        n = 501, add = TRUE, lwd = 2)
6 legend("topright", fill = "darkred", legend = "Estimated model")
```

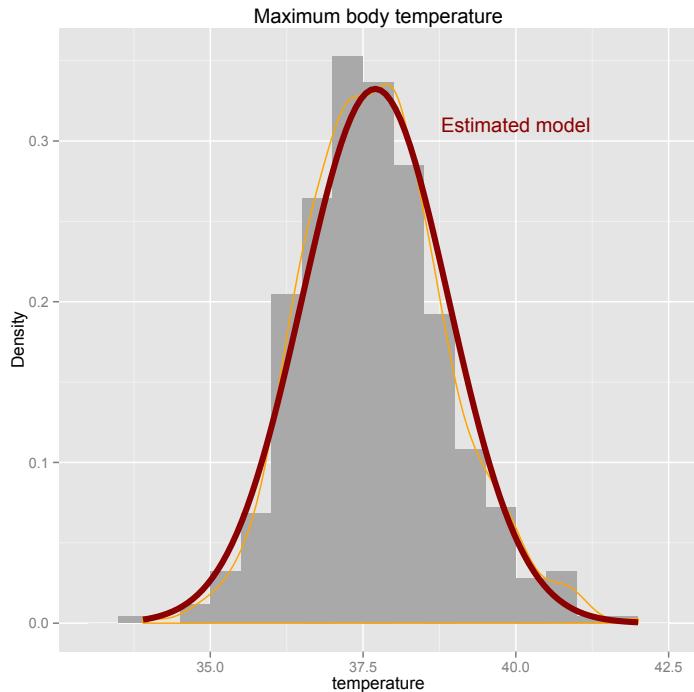


Argument `lwd` controls the thickness of the lines, where the default value is 1 and values larger than 1 lead to thicker lines. We repeat the plot applying the functions of package "ggplot2" (Wickham (2009)). Beside the functions `ggplot`, `geom_histogram`, and `geom_density`, we need function `stat_function`, which can be used to add the graph of a function to a plot.

```

1 ggplot(ICUData[-398,], aes(x=temperature)) +
2   geom_histogram(aes(y=..density..), binwidth = 0.5, right = TRUE,
3                 fill = "darkgrey") +
4   geom_density(color = "orange") + ylab("Density") +
5   stat_function(fun = dnorm, args = list(mean = 37.7, sd = 1.2),
6                 color = "darkred", lwd = 2) +
7   annotate("text", x = 40, y = 0.31, col = "darkred",
8           label = "Estimated model") +
9   ggtitle("Maximum body temperature")

```



With function `annotate` we additionally label the graph.

.....

www.alcatel-lucent.com/careers

Alcatel-Lucent 

What if
you could
build your
future and
create the
future?

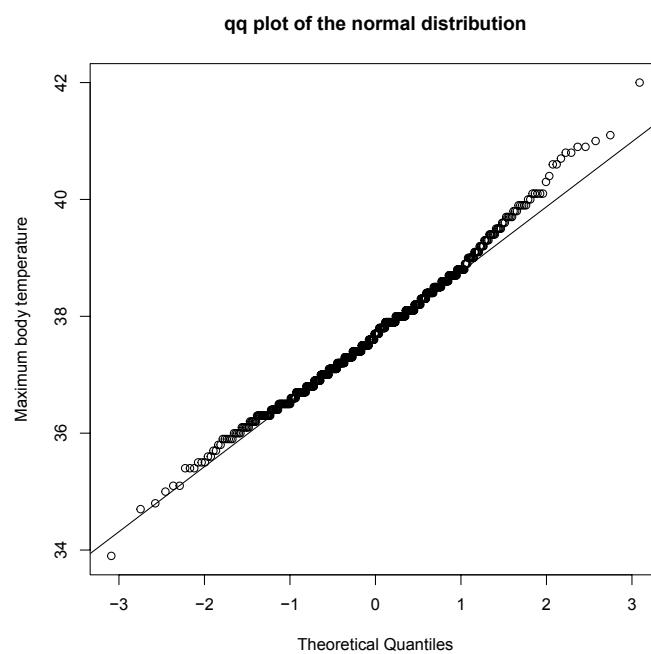


One generation's transformation is the next's status quo.
In the near future, people may soon think it's strange that
devices ever had to be "plugged in." To obtain that status, there
needs to be "The Shift".



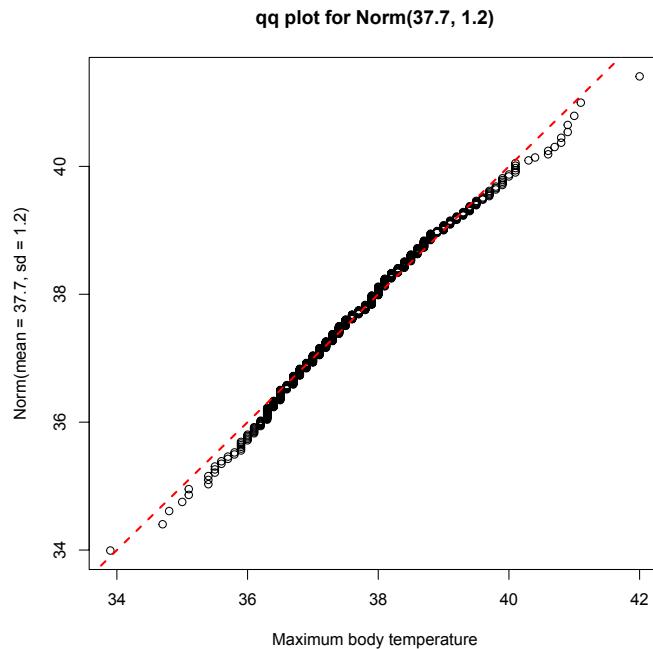
In a similar way, we could compare the empirical cumulative distribution function with the cumulative distribution function of the model, what we will not do here. Instead, we will introduce a new kind of plot, which is frequently applied for such comparisons, the so-called **quantile-quantile plot** (qq plot for short). In this plot, the empirical and theoretical quantiles are compared. The closer the points are to the straight line, the better the theoretical model explains the observations. In case of the normal distribution, we can use R functions `qqnorm` and `qqline`.

```
1 qqnorm(ICUData$temperature[-398], main = "qq plot of the normal distribution",
2        ylab = "Maximum body temperature")
3 qqline(ICUData$temperature[-398])
```



With the help of this plot we can generally check, if the data may stem from a normal distribution. In the current situation, we see that slightly more high temperatures were observed as we would expect in case of the normal distribution. If we want to compare our data with a concrete distribution, we could apply function `qqplot` instead of `qqnorm`. However, the call is somewhat cumbersome. It is clearly simpler to apply function `qqplot` of package "distr" (Ruckdeschel et al. (2006)).

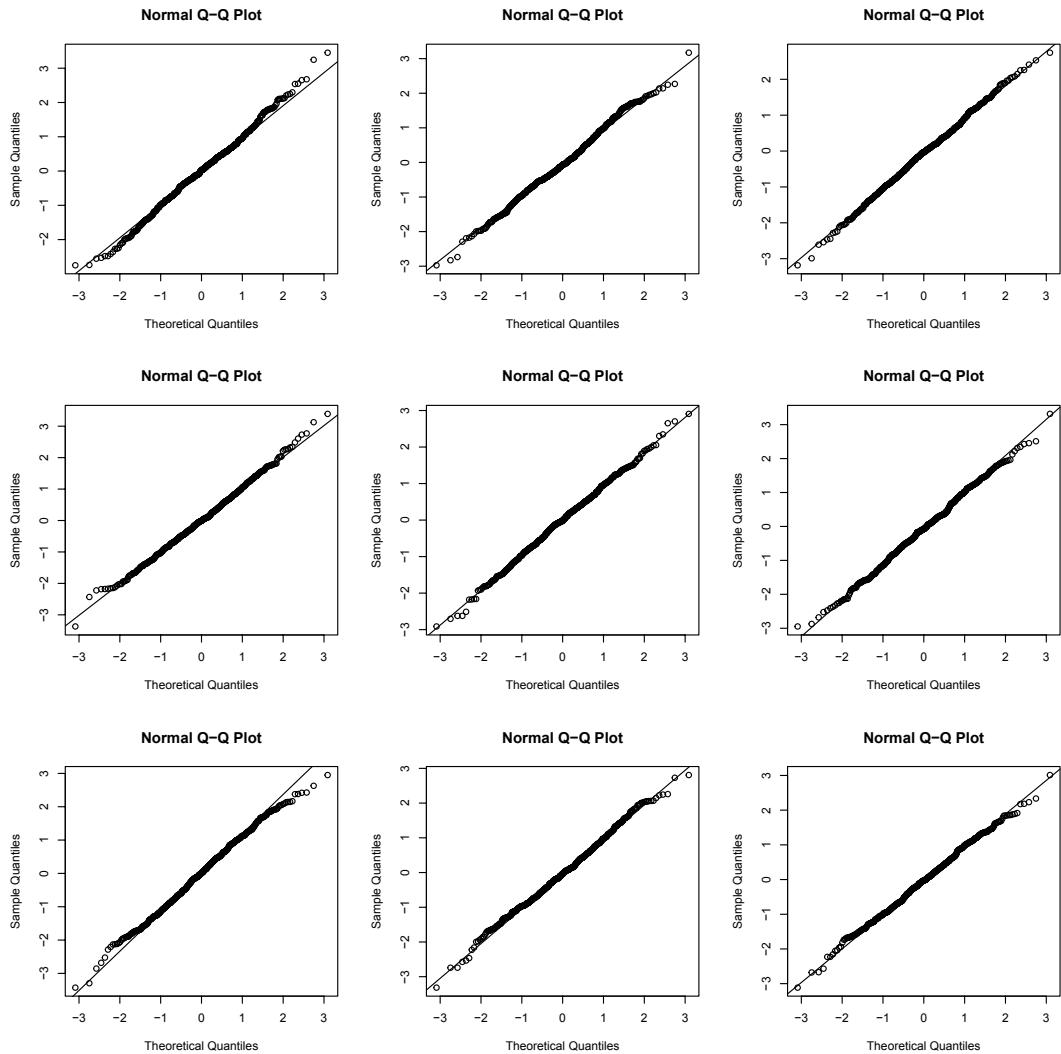
```
1 qqplot(ICUData$temperature[-398], Norm(mean = 37.7, sd = 1.2),
2        xlab = "Maximum body temperature",
3        main = "qq plot for Norm(37.7, 1.2)")
```



In contrast to the default plot in R, the x and y axis are interchanged. Our data seem to be in good agreement with the estimated model, but there are also some deviations. We compare our plot with some qq plots of normally distributed data, to be able to better judge the result. In the sequel, we generate standard normal data via function `rnorm` and generate a qq plot by means of functions `qqnorm` and `qqline`. To get a better impression of the variations between samples, we repeat it nine times. For this, we use a so-called `for` loop. Furthermore, we apply function `par`, which can be used to change various graphical parameters, to adapt the graphic device such that all plots are shown in one figure. With argument `mfrow` a figure can be divided into a certain number of rows and columns. In our situation, we choose three rows and three columns, which will be filled row-wise.

```

1 par(mfrow=c(3,3))
2 for(i in 1:9){
3   x <- rnorm(499)
4   qqnorm(x)
5   qqline(x)
6 }
```



Thus, there are also certain deviations of the straight line in case of normally distributed data. This confirms our first impression that the maximum body temperature of ICU patients (without strongly hypothermic patients) is quite well described by a normal distribution.

In case of the normal distribution, we can also apply the median and the appropriately standardized MAD as consistent estimators of mean and standard deviation. As we have already seen in Section 2.4.1, it is not necessary to remove patient 398.

```
1 median(ICUData$temperature)
```

```
[1] 37.7
```

```
1 mad(ICUData$temperature)
```

```
[1] 1.18608
```

The results are very similar to the ML estimates.

Note:

Median and MAD yield consistent estimates of the theoretical median and MAD under very general assumptions. In particular, it is not necessary to assume a certain parametric model. Therefore, they are also called non-parametric estimators. Because of this general property and their additional robustness, these estimators are useful for many applications.

Another estimating principle, which works well for simple probability models, is the so-called minimum distance estimation.

Definition 5.7 (Minimum-distance estimator). *Let $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ ($k \in \mathbb{N}$), be some probability model. Furthermore, let x_1, \dots, x_n be realizations of independent and P_θ distributed random variables X_1, \dots, X_n , and \hat{F}_n their empirical distribution. Then, we consider*

$$D(\theta) = \text{dist}(P_\theta, \hat{F}_n) \quad (5.9)$$

where *dist* represents a distance between distributions. The **minimum-distance estimator** (abbreviated: MD estimator) for θ is the position of the minimum of $D(\theta)$.

We give some additional explanations.



Maastricht University *Leading in Learning!*

Join the best at the Maastricht University School of Business and Economics!



Top master's programmes

- 33rd place Financial Times worldwide ranking: MSc International Business
- 1st place: MSc International Business
- 1st place: MSc Financial Economics
- 2nd place: MSc Management of Learning
- 2nd place: MSc Economics
- 2nd place: MSc Econometrics and Operations Research
- 2nd place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

Visit us and find out why we are the best!
Master's Open Day: 22 February 2014



Maastricht University is the best specialist university in the Netherlands (Elsevier)

www.mastersopenday.nl

Remark 5.8.

- a) MD estimators are usually consistent estimators.
- b) In the sequel, we will determine the Cramér-von-Mises-MD estimator (CvM-MD estimator for short) and the Kolmogorov(-Smirnov) MD estimator (KS-MD estimator for short). The definitions of the corresponding distances are based on the cumulative distribution functions. The **Cramér-von-Mises distance** reads

$$dist_{CvM}(P_\theta, \hat{F}_n) = \int |P_\theta(x) - \hat{F}_n(x)|^2 Q(dx) \quad (5.10)$$

where usually P_θ or \hat{F}_n is chosen for the distribution Q . In case of \hat{F}_n , the integral becomes a sum. The **Kolmogorov(-Smirnov) distance** is

$$dist_{KS}(P_\theta, \hat{F}_n) = \max_{x \in \mathbb{R}} |P_\theta(x) - \hat{F}_n(x)| \quad (5.11)$$

Both MD estimator are very robust against outliers and certain model deviations.

We apply function `MDEstimator` of package "distrMod" (Kohl and Ruckdeschel (2010)) for computing the MD estimators. We again consider the maximum body temperature of our ICU patients and first compute the CvM-MD estimator. As in case of the ML estimator, we proceed in two steps: First we define the model and then compute the estimator. Without patient 398 we get

```
1 model <- NormLocationScaleFamily()
2 MDEstimator(ICUData$temperature[-398], model, distance = CvMDist)
```

```
Evaluations of Minimum CvM distance estimate:
-----
An object of class "Estimate"
generated by call
  MDEstimator(x = ICUData$temperature[-398], ParamFamily = model,
               distance = CvMDist)
samplesize: 499
estimate:
  mean      sd
37.675207 1.136109
Criterion:
  CvM distance
  0.01305277
```

The result is very similar to the ML estimator. We repeat the estimation and this time apply the KS-MD estimator.

```
1 MDEstimator(ICUData$temperature[-398], model, distance = KolmogorovDist)
```

```
Evaluations of Minimum Kolmogorov distance estimate:
-----
An object of class "Estimate"
generated by call
  MDEstimator(x = ICUData$temperature[-398], ParamFamily = model,
               distance = KolmogorovDist)
samplesize: 499
estimate:
  mean      sd
37.679362 1.141183
Criterion:
Kolmogorov distance
  0.03156417
```

Again, we obtain a very similar result. We repeat the estimation and this time do not omit patient 398. The results show the robustness of the MD estimators in contrast to the ML estimator. To make the difference easier to recognize, we reduce the printed output of the functions to the minimum by means of function `distrModOptions` and argument `show.details = "minimal"`.

```
1 ## change amount of printed output
2 distrModOptions(show.details = "minimal")
3 ## ML estimator
4 MLEstimator(ICUData$temperature, model)
```

```
Evaluations of Maximum likelihood estimate:
-----
  mean      sd
37.66320000 1.73373751
( 0.07753510) ( 0.05482559)
```

```
1 ## CvM-MD estimator
2 MDEstimator(ICUData$temperature, model, distance = CvMDist)
```

```
Evaluations of Minimum CvM distance estimate:
-----
estimate:
  mean      sd
37.67177 1.13942
```

```
1 ## KS-MD estimator
2 MDEstimator(ICUData$temperature, model, distance = KolmogorovDist)
```

```
Evaluations of Minimum Kolmogorov distance estimate:
-----
estimate:
mean      sd
37.676420 1.143365
```

```
1 ## reset printed output to default
2 distrModOptions(show.details = "maximal")
```

Thus, in case of the MD estimators the results remain almost unchanged, whereas in case of the ML estimator especially the estimate of the standard deviation clearly increases.

Note:

There are several other important classes of estimators such as generalized ML estimators (M estimator for short), asymptotically linear estimators (AL estimator for short) or rank based estimators (R estimator for short). A very important construction principle especially for complex models is the least-squares estimation (LS estimation for short) introduced by Gauß and Legendre. In this case, the model is estimated by minimizing the sum of the quadratic deviations of the observations from the model.

Empowering People. Improving Business.

BI Norwegian Business School is one of Europe's largest business schools welcoming more than 20,000 students. Our programmes provide a stimulating and multi-cultural learning environment with an international outlook ultimately providing students with professional skills to meet the increasing needs of businesses.

BI offers four different two-year, full-time Master of Science (MSc) programmes that are taught entirely in English and have been designed to provide professional skills to meet the increasing need of businesses. The MSc programmes provide a stimulating and multi-cultural learning environment to give you the best platform to launch into your career.

- MSc in Business
- MSc in Financial Economics
- MSc in Strategic Marketing Management
- MSc in Leadership and Organisational Psychology

www.bi.edu/master

BI NORWEGIAN BUSINESS SCHOOL

EFMD
EQUIS
ACCREDITED

5.3 Confidence Intervals

In the previous section, we have learned about several estimating procedures and we now know that we should use unbiased (or at least consistent) and efficient estimators. However, these are only theoretical properties, which in practice can not tell us, how close our point estimator actually is to the wanted unknown parameter. A possibility to further safeguard the point estimator, are so-called confidence intervals.

Definition 5.9 (Confidence interval). *Let $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ ($k \in \mathbb{N}$), be some probability model. Furthermore, let x_1, \dots, x_n be realizations of independent and P_θ distributed random variables X_1, \dots, X_n . Then, the **interval estimator***

$$\hat{I}(x_1, \dots, x_n) = [S_u(x_1, \dots, x_n), S_o(x_1, \dots, x_n)] \quad (5.12)$$

is called a $(1 - \alpha)$ -confidence interval, if

$$P(\theta \in \hat{I}) \geq 1 - \alpha$$

for $\alpha \in (0, 1)$. Here, S_u and S_o are estimators for the lower and upper bound of the interval.

We give some additional explanations.

Remark 5.10.

- a) *The definition also allows for one-sided confidence intervals. In this case, one boundary of the interval is free and only S_u or S_o is needed.*
- b) *It is said: A confidence interval covers the true unknown parameter with a probability of $1 - \alpha$. This should express, that in 95% of the cases, in which the data is used to compute confidence intervals, these intervals include the true unknown parameter. The statement, that the true unknown parameter lies in the computed confidence interval with 95% probability strictly speaking is wrong. Because after determining the confidence interval, the true unknown parameter either lies in the interval or not.*
- c) *We take a more detailed look at the components of a confidence interval. More concretely, they usually are of the following form*

$$\hat{I}(x_1, \dots, x_n) = [S_n(x_1, \dots, x_n) - k_1 \sigma_{S_n}, S_n(x_1, \dots, x_n) + k_2 \sigma_{S_n}] \quad (5.13)$$

The components are:

- *A point estimator S_n of the true unknown parameter θ .*
- *The standard deviation of the point estimator σ_{S_n} .*
- *Two constants $k_1, k_2 \in (0, \infty)$ usually depending on α , n and the distribution of S_n .*

Moreover, the following notions are used:

Confidence level: the chosen coverage probability $1 - \alpha$ for the true unknown parameter θ .

Basis: point estimator of the true unknown parameter θ , often the center of the interval.

Confidence bounds: lower and upper bound of the interval.

Maximum estimate error: maximum distance between point estimator and the confidence bounds.

We give some examples of confidence intervals.

Example 5.11.

- a) We consider the normal distribution model $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}\}$, where we assume $\sigma^2 \in (0, \infty)$ to be known. As we have learned in Section 5.2, the arithmetic mean is an unbiased and efficient estimator of μ . We assume that the observations x_1, \dots, x_n are realizations of independent and identical distributed random variables X_1, \dots, X_n with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ ($i = 1, \dots, n$) and obtain

$$\text{AM}(x_1, \dots, x_n) \sim \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right) \quad (5.14)$$

It follows, $\sigma_{S_n} = \frac{1}{\sqrt{n}}\sigma$, which is also called the **standard error** (SE) of the arithmetic mean (SEM). Because of the symmetry of the normal distribution, we get $k_1 = k_2$ and have to choose $k_1 = k_2 = z_{1-\alpha/2}$, the $(1-\alpha/2)$ quantile of the standard normal distribution. Consequentially, the $(1-\alpha)$ confidence interval reads

$$\text{AM}(x_1, \dots, x_n) \mp z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (5.15)$$

In practice, in most cases σ is also unknown and must be estimated, too. As we want to capture the true unknown value of μ , the unbiased sample variance \tilde{S} (i.e. standardization $\frac{1}{n-1}$) is an appropriate candidate for the estimation, where

$$\frac{(n-1)\tilde{S}}{\sigma} \sim \text{Chisq}(n-1) \quad (5.16)$$

That is, the additional estimation of σ leads us away from the normal distribution towards the t distribution with $n-1$ degrees of freedom (see also Remark 4.28 (b)). Thus, we get as confidence interval

$$\text{AM}(x_1, \dots, x_n) \mp t_{n-1; 1-\alpha/2} \frac{\sqrt{\tilde{S}(x_1, \dots, x_n)}}{\sqrt{n}} \quad (5.17)$$

where $t_{n-1;1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the t distribution with $n - 1$ degrees of freedom.

- b) If we conversely consider the probability model $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) \mid \sigma \in (0, \infty)\}$, where we more realistically assume μ to be unknown, we obtain the following asymmetric $(1-\alpha)$ confidence interval for σ^2

$$\left[\frac{(n-1)\tilde{S}}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)\tilde{S}}{\chi^2_{n-1;\alpha/2}} \right] \quad (5.18)$$

Here, $\chi^2_{n-1;1-\alpha/2}$ and $\chi^2_{n-1;\alpha/2}$ are the $(1-\alpha/2)$ and the $\alpha/2$ quantile of the χ^2 distribution with $n - 1$ degrees of freedom, respectively.

Note:

The above confidence intervals are not only of interest for the normal distribution model, but may also be used as approximations for other probability models. The reason for it is the central limit theorem, which states that the distribution of the arithmetic mean of quite arbitrary independent and identical distributed random variables converges towards a normal distribution.

Need help with your dissertation?

Get in-depth feedback & advice from experts in your topic area. Find out what you can do to improve the quality of your dissertation!

Get Help Now



Go to www.helpmyassignment.co.uk for more info

Helpmyassignment

In addition to the point estimates for the maximum body temperature of our ICU patients (cf. Section 5.2), we will now determine 95% confidence intervals (i.e. $\alpha = 0.05$). Omitting patient 398, this leads to the following confidence bounds for the mean μ .

```
1 ## arithmetic mean
2 AM <- mean(ICUData$temperature[-398])
3 AM
```

```
[1] 37.72044
```

```
1 ## standard deviation
2 SD <- sd(ICUData$temperature[-398])
3 SD
```

```
[1] 1.173187
```

```
1 ## alpha
2 alpha <- 0.05
3 alpha
```

```
[1] 0.05
```

```
1 ## sample size
2 n <- nrow(ICUData)-1
3 n
```

```
[1] 499
```

```
1 ## lower confidence bound
2 AM - qt(1-alpha/2, df = n-1)*SD/sqrt(n)
```

```
[1] 37.61725
```

```
1 ## upper confidence bound
2 AM + qt(1-alpha/2, df = n-1)*SD/sqrt(n)
```

```
[1] 37.82363
```

The reported interval should be chosen in dependence of the accuracy of the temperature measurement, e.g. [37.61, 37.83] or [37.60, 37.85] or [37.6, 37.9] might be appropriate. Each interval covers the true unknown mean with at least 95% probability. The confidence interval of the arithmetic mean can somewhat simpler also be computed by means of function `t.test`. This function can be used for computing t tests, which will be introduced in Chapter 6. However, at this point we only take a look at the confidence interval (`conf.int`) and ignore the remaining results.

```
1 t.test(ICUData$temperature[-398])$conf.int
```

```
[1] 37.61725 37.82363
attr(,"conf.level")
[1] 0.95
```

As the sample size is quite large in our example, we could, in sense of the central limit theorem, use the quantile of the standard normal distribution instead of the quantile of the t-distribution with 498 degrees of freedom. We compare the two quantiles

```
1 qt(1-alpha/2, df = n-1)
```

```
[1] 1.964739
```

```
1 qnorm(1-alpha/2)
```

```
[1] 1.959964
```

and get a difference of less than 0.005.

Consequently, the confidence bounds of the approximative interval are very similar.

```
1 ## lower confidence bound
2 AM - qnorm(1-alpha/2)*SD/sqrt(n)
```

```
[1] 37.61751
```

```
1 ## upper confidence bound
2 AM + qnorm(1-alpha/2)*SD/sqrt(n)
```

```
[1] 37.82338
```

The differences are probably beyond measurement accuracy and are irrelevant for practical applications. Based on ML estimators, we can determine a similar approximative confidence interval. We apply function `fitdistr` of package "MASS" (Venables and Ripley (2002)) combined with function `confint`.

```

1 library(MASS)
2 ## ML estimator
3 ML <- fitdistr(ICUData$temperature[-398], densfun = "normal")
4 ## approximative confidence interval
5 confint(ML)
```

```

 2.5 %    97.5 %
mean 37.617609 37.823273
sd    1.099298  1.244725
```

That is, we get an approximative confidence interval for the mean μ as well as for the standard deviation σ . We can also determine these intervals by means of function `MLEstimator` of package "distrMod" (Kohl and Ruckdeschel (2010)) as well as function `confint`.

Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations.

Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF

Download free eBooks at bookboon.com

```

1 library(distrMod)
2 ## model
3 model <- NormLocationScaleFamily()
4 ## ML estimator
5 ML2 <- MLEstimator(ICUData$temperature[-398], model)
6 ## approximative confidence interval
7 confint(ML2)

A[n] asymptotic (CLT-based) confidence interval:
      2.5 %    97.5 %
mean 37.617609 37.823273
sd   1.099298 1.244725
Type of estimator: Maximum likelihood estimate
samplesize: 499
Call by which estimate was produced:
MLEstimator(x = ICUData$temperature[-398], ParamFamily = model)

```

We compare the approximative confidence interval of the standard deviation, which is symmetric around the ML estimator of the standard deviation, with the asymmetric interval that uses the χ^2 distribution.

```

1 ## lower confidence bound
2 sqrt((n-1)*SD^2/qchisq(1-alpha/2, df = n-1))

```

```
[1] 1.104636
```

```

1 ## upper confidence bound
2 sqrt((n-1)*SD^2/qchisq(alpha/2, df = n-1))

```

```
[1] 1.250879
```

The confidence interval is slightly different, but the differences are only in the range of permilles.

Note:

If the sample size n is not too small, one can use the approximative confidence intervals emerging from the central limit theorem. Figure 5.2 shows the ratio between the 95% quantile of the t distribution with increasing degrees of freedom and the 95% quantile of the standard normal distribution. From a sample size of about 25 onwards, the difference between the quantiles and thus between the maximum estimate errors is below 5%.

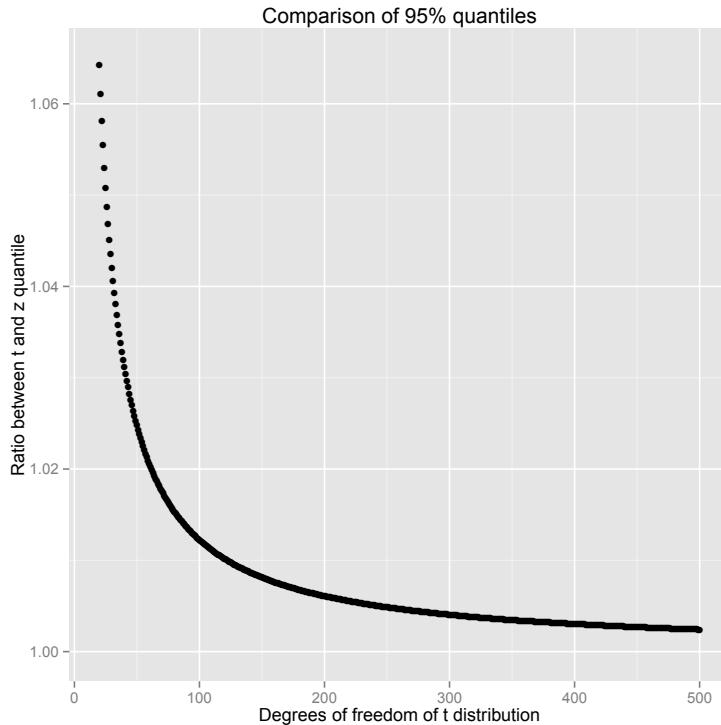


Figure 5.2: Ratio between 95% quantiles of t and standard normal distribution.

In the following example we discuss the Bernoulli model.

Example 5.12. We consider the probability model $\mathcal{P} = \{\text{Bernoulli}(p) \mid p \in (0, 1)\}$. As we have learned in Section 5.2, the relative frequency \hat{p} is the ML estimator of p and is unbiased and efficient. As the Bernoulli distribution is a discrete distribution, the distribution of \hat{p} is also discrete and quantiles of discrete distribution are not necessarily unique. Consequentially, there is a whole series of proposals for “exact” confidence intervals for the probability p ; for example the Clopper-Pearson or the Agresti-Coull interval. I omit the explicit specification of the formulas.

An application of the central limit theorem yields the following approximative confidence interval for p

$$\left(\hat{p} \pm \frac{1}{2n} \right) \mp z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (5.19)$$

The correction term $\frac{1}{2n}$ is called **continuity correction** and improves the approximation. Furthermore, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. To make the asymptotic interval applicable, one should verify that $n\hat{p} > 5$ and $n(1 - \hat{p}) > 5$; i.e., the more \hat{p} approaches 0 or 1, the larger the sample size has to be.

If we consider drawing without replacement and the underlying population is small having $N \in \mathbb{N}$ members, it is recommended, to apply the following slightly modified confidence interval

$$\left(\hat{p} \pm \frac{1}{2n} \right) \mp z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \frac{N-n}{N-1}} \quad (5.20)$$

The additional factor $\frac{N-n}{N-1}$, which we have already met in Remark 4.8 (c), is called **finite-sample correction** and represents the difference between drawing with and without replacement.

We consider the prevalence of liver failure on the ICU and additionally safeguard the estimation by a confidence interval. There are several packages including functions for computing “exact” confidence intervals. We will apply function `binomCI` of package "MKmisc" (Kohl (2015)). We can install the package using the following R Code

```
1 install.packages ("MKmisc")
```

or via window *Packages* of RStudio (cf. Section 2.4.1). We load the package and compute the Clopper-Pearson and the Agresti-Coull interval. For this, we need the number of patients with liver failure as well as the total number of patients.

**“I studied English for 16 years but...
...I finally learned to speak it in just six lessons”**

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

```

1 library(MKmisc)
2 ## absolute frequency of liver failure
3 table(ICUData$liver.failure)

```

0	1
480	20

```

1 ## Clopper-Pearson interval
2 binomCI(x = 20, n = 500, method = "clopper-peerson")

```

```

$estimate
[1] 0.04

$CI
[1] 0.02460131 0.06110261
attr(,"confidence level")
[1] 0.95

```

```

1 ## Agresti-Coull interval
2 binomCI(x = 20, n = 500, method = "agresti-coull")

```

```

$estimate
[1] 0.0435072

$CI
[1] 0.02569479 0.06131960
attr(,"confidence level")
[1] 0.95

```

We get minor differences between the intervals, in particular, the Agresti-Coull interval is not based on the relative frequency. The approximative interval reads

```

1 ## relative frequency
2 p <- 20/500
3 ## sample size
4 n <- 500
5 ## alpha
6 alpha <- 0.05
7 ## lower confidence bound
8 p + 1/(2*n) - qnorm(1-alpha/2)*sqrt(p*(1-p)/n)

```

[1] 0.02382374

```
1 ## upper confidence bound
2 p - 1/(2*n) + qnorm(1-alpha/2)*sqrt(p*(1-p)/n)
```

```
[1] 0.05617626
```

Moreover, we can again compute an asymptotic confidence interval by means of function `MLEstimator` of package "distrMod" (Kohl and Ruckdeschel (2010)) and function `confint`. To get a better overview, we additionally reduce the output to the minimum.

```
1 ## minimum output
2 distrModOptions(show.details = "minimal")
3 ## Bernoulli model
4 model <- BinomFamily(size = 1)
5 ## ML estimator
6 MLp <- MLEstimator(ICUData$liver.failure, model)
7 MLp
```

```
Evaluations of Maximum likelihood estimate:
-----
0.040000000
(0.008763561)
```

```
1 ## confidence interval
2 confint(MLp)
```

```
A[n] asymptotic (CLT-based) confidence interval:
      2.5 %    97.5 %
[1,] 0.02282374 0.05717626
```

```
1 ## reset output to default
2 distrModOptions(show.details = "maximal")
```

The result corresponds to the asymptotic confidence interval above without continuity correction. By applying function `BinomFamily` with argument `size = 1`, we can generate a Bernoulli model. Roughly summarized, we can assume a prevalence of liver failure on the ICU in the range from 2.2% to 6.1% with relatively high certainty.

Note:

As there is often more than one way to determine a confidence interval of a certain parameter, it is recommended to specify not only the interval but also the type of the interval in practice. Only by doing this, a reader can reproduce the analysis and its results.

A very interesting option to describe the location and scale of data are median and MAD. On the one hand, both estimators are very robust, on the other hand, it is not necessary to assume a specific parametric family.

Example 5.13. Let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ be the increasingly sorted observations. Then, the $(1-\alpha)$ confidence interval of the median reads

$$[x_{(k)}, x_{(n-k+1)}] \quad (5.21)$$

where $k \in \mathbb{N}$ has to be determined, such that the following inequality holds

$$1 - 2 \sum_{i=1}^{k-1} \binom{n}{i} 0.5^n \geq 1 - \alpha \quad (5.22)$$

This approach can be transferred to the MAD by considering it as the median of $|x_1 - M|, \dots, |x_n - M|$ with $M = \text{median}(x_1, \dots, x_n)$. In case of the normal distribution the MAD is usually standardized by 1.4826 to yield a consistent estimator of the standard deviation.



What do you want to do?

No matter what you want out of your future career, an employer with a broad range of operations in a load of countries will always be the ticket. Working within the Volvo Group means more than 100,000 friends and colleagues in more than 185 countries all over the world. We offer graduates great career opportunities – check out the Career section at our web site www.volvologroup.com. We look forward to getting to know you!

VOLVO
AB Volvo (publ)
www.volvologroup.com

VOLVO TRUCKS | RENAULT TRUCKS | MACK TRUCKS | VOLVO BUSES | VOLVO CONSTRUCTION EQUIPMENT | VOLVO PENTA | VOLVO AERO | VOLVO IT
VOLVO FINANCIAL SERVICES | VOLVO 3P | VOLVO POWERTRAIN | VOLVO PARTS | VOLVO TECHNOLOGY | VOLVO LOGISTICS | BUSINESS AREA ASIA

Download free eBooks at bookboon.com

We consider the maximum body temperature of our ICU patients and determine 95% confidence intervals for median and MAD. For this, we apply function `medianCI` of package "MKmisc" (Kohl (2015)). In case of the MAD, we choose the version that is standardized with 1.4826 to get a confidence interval for the standard deviation.

```
1 ## exact confidence interval for the median
2 medianCI(ICUData$temperature)
```

```
$call
medianCI(x = ICUData$temperature)

$estimate
[1] 37.7

$CI
[,1] [,2]
[1,] 37.4 37.8
[2,] 37.5 37.9
attr("(exact) confidence level")
[1] 0.9500548
```

```
1 ## exact confidence interval for the MAD
2 M <- median(ICUData$temperature)
3 medianCI(1.4826 * abs(ICUData$temperature - M))
```

```
$call
medianCI(x = 1.4826 * abs(ICUData$temperature - M))

$estimate
[1] 1.18608

$CI
[,1] [,2]
[1,] 1.03782 1.18608
[2,] 1.03782 1.33434
attr("(exact) confidence level")
[1] 0.9500548
```

In both cases, we get two possible intervals, which is one of the disadvantages of these exact confidence intervals. Since the sample size in our example is quite large, we may instead turn to the asymptotic confidence interval. We obtain

```

1 ## asymptotic confidence interval for the median
2 medianCI(ICUData$temperature , method = "asymptotic")

$call
medianCI(x = ICUData$temperature , method = "asymptotic")

$estimate
[1] 37.7

$CI
[1] 37.5 37.8
attr(,"(asymptotic) confidence level")
[1] 0.95

```

```

1 ## asymptotic confidence interval for the MAD
2 medianCI(1.4826*abs(ICUData$temperature - M) , method = "asymptotic")

$call
medianCI(x = 1.4826 * abs(ICUData$temperature - M) , method = "asymptotic")

$estimate
[1] 1.18608

$CI
[1] 1.03782 1.33434
attr(,"(asymptotic) confidence level")
[1] 0.95

```

The results are very similar to the exact intervals. Overall, the intervals are somewhat longer than in case of the arithmetic mean and the (sample) standard deviation. This is the price we have to pay for these non-parameteric estimating procedures and their robustness.

In the sequel, we take a look at the MD estimators. Here, the computation of confidence intervals is rather difficult, as the distribution of these estimators is quite hard to determine. In case of the CvM-MD estimator, we can compute an asymptotic confidence interval by means of the function `MDEstimator` of package "distrMod" (Kohl and Ruckdeschel (2010)) and function `confint`. First, we again consider the maximum body temperature of our ICU patients.

```

1 ## minimum output
2 distrModOptions("show.details" = "minimal")
3 ## model
4 model <- NormLocationScaleFamily()
5 ## CvM-MD estimator incl. variance
6 MD <- MDEstimator(ICUData$temperature, model,
7                     asvar.fct = distrMod:::CvMMDcovariance)
8 ## 95% confidence interval
9 confint(MD)

```

```

A[n] asymptotic (CLT-based) confidence interval:
      2.5 %    97.5 %
mean 37.539403 37.813437
sd   1.055498  1.231232

```

The confidence intervals are slightly longer than in case of the ML estimator, but we did not have to exclude patient 398 due to the robustness of the CvM-MD estimator.

In a similar fashion, we can also compute the confidence interval for the prevalence of liver failure on the ICU.

```

1 ## model
2 model <- BinomFamily(size = 1)
3 ## CvM-MD estimator incl. variance
4 MDp <- MDEstimator(ICUData$liver.failure, model,
5                      asvar.fct = distrMod:::CvMMDCovariance)
6 ## 95% confidence interval
7 confint(MDp)

A[n] asymptotic (CLT-based) confidence interval:
      2.5 %    97.5 %
prob 0.02281891 0.05716894

```

```

1 ## reset to default output
2 distrModOptions("show.details" = "maximal")

```

The results are almost identical to the ML estimator.

Note:

Beside the introduced options, there are many more possibilities to compute confidence intervals in R. In particular, confidence intervals are usually determined during the computation of statistical tests, which will be introduced in Chapter 6.

Finally, we demonstrate by the following simple example how confidence intervals may be used for sample size calculations (cf. Section 5.1).

Example 5.14. We consider the question how many persons polling institutes should ask in opinion polls to get reliable prognoses. Assuming a large population as in case of national elections, we can confidently neglect the finite-sample correction and can apply the asymptotic confidence interval given in Example 5.12. As we are interested in the deviation from the estimated value, i.e. the maximum estimate error, we have to take a closer look at the following expression

$$z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (5.23)$$

Apparently, the estimate error varies with the confidence level $1 - \alpha$, the estimated probability \hat{p} and the sample size n . We assume a 95% confidence interval; that is, we get for $z_{1-\alpha/2} = z_{0.975}$

```
1 qnorm(0.975)
```

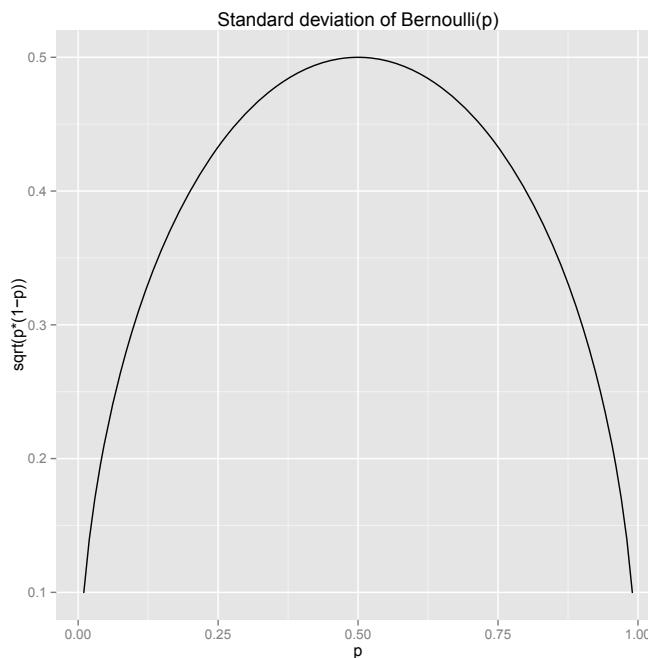
```
[1] 1.959964
```

Next, we take a closer look at the standard deviation $\sqrt{p(1-p)}$ of the Bernoulli distribution.

```

1 ## values of p
2 p <- seq(from = 0.01, to = 0.99, length = 100)
3 ## standard deviation
4 SD <- sqrt(p*(1-p))
5 ## plot
6 qplot(p, SD, ylab = "sqrt(p*(1-p))", xlab = "p", geom = "line",
7       main = "Standard deviation of Bernoulli(p)")

```



As we see, the standard deviation is maximal for $p = 0.5$ and symmetrically decreases, if we move away from this value in either direction. Thus, in case of $p = 0.5$, the maximum estimate error of the 95% confidence interval is

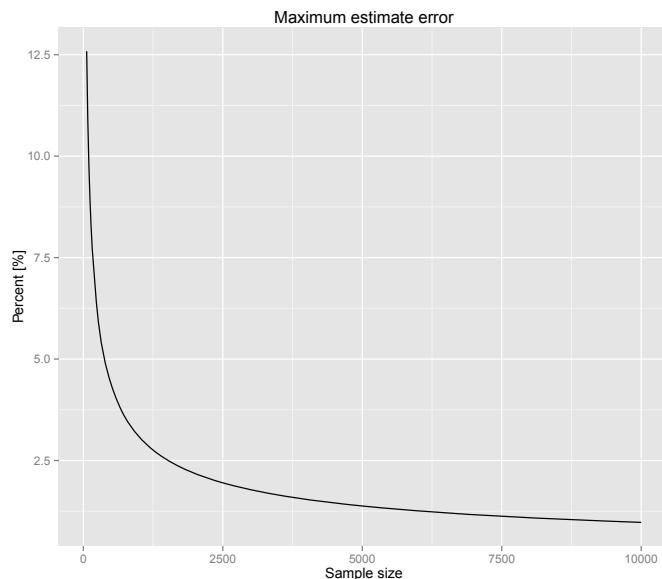
$$1.96 \times \frac{0.5}{\sqrt{n}} = \frac{0.975}{\sqrt{n}} = \frac{97.5}{\sqrt{n}} \% \quad (5.24)$$

We plot the maximum estimate error as a function of the sample size.

```

1 ## sample size
2 n <- seq(60, 10000, by = 20)
3 ## maximum estimate error
4 maxFehler <- 97.5 / sqrt(n)
5 ## plot
6 qplot(n, maxFehler, geom = "line", xlab = "Sample size",
7       ylab = "Percent [%]", main = "Maximum estimate error")

```



Usually, 1000 persons are interviewed in an opinion poll. In this case, the maximum estimate error is at most about 3%. In case of important elections, pollsters interview up to 50000 persons leading to a maximum estimate error of less than 0.5% in the worst case. These considerations and calculations are not only important for pollsters, but play an important role in other fields such as epidemiology or medical statistics, too. Here it for instance is about estimating prevalences of diseases or success rates of treatments.



5.4 Exercises

Always briefly describe and explain the results. Use the ICU dataset for exercises 5–8 and always select appropriate functions for the computations.

1. Construct a dataset consisting of exactly five positive numbers, such that the median is equal to 5 and the arithmetic mean is equal to 7. In a second step, modify the dataset, such that the median is unchanged, but the arithmetic mean is larger than the third quartile.
2. How must a dataset look like, such that the standard deviation is equal to 0? In which situation is the standard deviation maximal? Use simple datasets to think about the questions.
3. One can study bone resorption by means of TRAP (tartrate resistant acid phosphatase), which can be measured in one's blood. In a trial of 31 young women, the arithmetic mean of TRAP was equal to 13.2 U/l (Units per liter). Assume a normal distribution model for TRAP, where the standard deviation is known to be $\sigma = 6.5$ U/l. Specify a 95% confidence interval for the mean μ of the women, who are represented by the trial. How does the interval change, if the standard deviation is not known, but was estimated as 6.5 U/l by means of the sample standard deviation (standardization $\frac{1}{n-1}$)?
4. Assume there are 6 successes in 20 tries. May you use the approximative confidence interval for the probability of success p ? Compare the Clopper-Pearson and the asymptotic confidence interval including the continuity correction. How does the interval change, if we assume drawing without replacement from a population of size $N = 1000$?
5. Estimate the probability that an ICU patient is male. Compute the ML and the CvM-MD estimator for the Bernoulli model $\mathcal{P} = \{\text{Bernoulli}(p) \mid p \in (0, 1)\}$. Determine the corresponding asymptotic confidence intervals. Compare the asymptotic confidence interval with the Clopper-Pearson interval.
6. Assume that one can describe the logarithmized bilirubin values of ICU patients by a normal distribution. Compute the ML estimator and compare the result with median and MAD. Determine also the respective confidence intervals. Plot the data by means of a histogram and add the two normal densities for the estimated parameters. In addition, verify by a qq plot, whether the assumption of a normal distribution for the logarithmized bilirubin values seems justifiable.
7. Assume that the length of stay (LOS) of ICU patients can be described by a gamma distribution. Compute the ML and the CvM-MD estimator as well as their asymptotic confidence intervals. Plot the data by means of a histogram and add the two gamma densities for the estimated parameters. In addition, verify by a qq plot, whether the assumption of a gamma distribution seems plausible.

8. Investigate the age of ICU patients and assume that you can describe it by a Weibull distribution. Determine the ML and the KS-MD estimator and give the asymptotic confidence interval of the ML estimator. Plot the data by means of a histogram and add the two Weibull densities for the estimated parameters. In addition, verify by a qq plot, whether it seems plausible to assume a Weibull distribution for age.
9. You want to study the probability of trash in a production process and for this purpose draw a representative sample of the produced parts. You estimate the unknown probability of trash and determine the corresponding 95% confidence interval. You repeat this procedure every month for five months, where each month you draw a new independent sample. Then, the probability that all five intervals cover the true unknown parameter is smaller than 95%. How large is this probability exactly? How likely is it, that at least four of the five confidence intervals cover the true unknown parameter?



www.sylvania.com

**We do not reinvent
the wheel we reinvent
light.**

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM

OSRAM
SYLVANIA



Click on the ad to read more

6 Statistical Tests

In this chapter we introduce statistical tests. In detail, it covers the following topics:

- Hypotheses
- Test decisions, power, sensitivity, type I and type II error
- Order for the correct conduct of a test
- t test: one-sample, paired, two-sample, Welch
- Wilcoxon signed rank test, Wilcoxon-Mann-Whitney U test
- F test, Ansari-Bradley test
- One-sample binomial test (exact and asymptotic)
- Fisher's exact test, χ^2 test
- Testing correlations (Pearson, Spearman, Kendall)
- One-way ANOVA, Kruskal-Wallis test
- Post hoc tests
- Pairwise t tests and Wilcoxon-Mann-Whitney U tests incl. Bonferroni-Holm corrections
- Testing for normality: Shapiro-Wilk test, Lilliefors (Kolmogorov-Smirnov) test, Cramér-von Mises test, Shapiro-Francia test

The R code of this chapter is included in file `Tests.R`, which can be downloaded from my website (link: www.stamats.de/RCodeEN.zip). You should use an additional R script to experiment with your own R code. Generating a new R script is described at the beginning of Chapter 2.

6.1 Introduction

Empirical investigations and studies usually start with a new idea, a conjecture about a certain often open problem. This conjecture is usually postulated on the basis of empirical observations and/or subjectspecific theoretical considerations. It facilitates the verification of an assumption, if it may be formulated in a precise and quantifiable way. In this case, one speaks of a **hypothesis**. First, one should collect all available information about the problem and elaborate the theoretical background to verify whether the hypothesis is generally plausible. Frequently, one hereby already realizes that the hypothesis can not be true, which saves work (and money).

In many fields, direct proofs of hypotheses are not possible and they can not be verified directly by a single experiment. At this point, statistics comes into play. We collect representative and relevant data for the problem and subject the data to a statistical analysis, where the results can be ensured by so-called **statistical tests**. In the following example, the general approach is described by means of a dice game.

Example 6.1. We consider a dice game, where it is important to throw “6”. After some time of playing, we realize that our dice only rarely gives “6”. Therefore, we conjecture

the frequency of “6” is too small

or more generally,

the frequency of “6” is incorrect.

In particular, this implies that not all sides of the dice occur with identical probability; that is, the dice is not fair. The precise and quantifiable formulation of the conjecture leads to the hypothesis:

The probability p of “6” is not equal to $\frac{1}{6}$; abbreviated: $p \neq \frac{1}{6}$

In general, an answer by means of statistical tests is only possible, if there are mutually exclusive cases. For the dice either

our hypothesis is true, i.e. $p \neq \frac{1}{6}$

or

our hypothesis is not true, i.e. $p = \frac{1}{6}$

In the present case, one collects information (evidence) for the hypothesis by throwing the dice n times and by counting the number of “6”. The open questions we can answer by means of statistical tests are:

1. How often should we throw the dice?
2. How many “6” do we need to decide in favor or against the hypothesis?

As the previous example shows, the origin of statistical tests are two mutually exclusive hypotheses. These are usually denoted as follows:

Null hypothesis H_0 : Hypothesis that shall be *falsified*.

Alternative (hypothesis) H_1 : Hypothesis that shall be *confirmed* (research hypothesis).

We transfer this notion to our dice example.

Example 6.2. We again consider the dice game, where it is important to throw “6”. Here, we obtain

Null hypothesis $H_0: p = \frac{1}{6}$ versus Alternative $H_1: p \neq \frac{1}{6}$

Since the alternative includes the cases $p < \frac{1}{6}$ and $p > \frac{1}{6}$, it is also called a **two-sided** hypothesis. Of course, also the **one-sided** cases

- $H_0: p = \frac{1}{6}$ versus $H_1: p < \frac{1}{6}$
- $H_0: p \geq \frac{1}{6}$ versus $H_1: p < \frac{1}{6}$

would be possible.

Note:

The decision whether to consider a one-sided or two-sided alternative, must always be made before conducting the test. In medicine, for instance, one-sided alternatives emerge only rarely. In fact, often an improvement is solely of interest, but a worsening would have far reaching consequences, thus for ethical reasons and for the safety of the patients a two-sided alternative has to be chosen.

In the framework of inferential statistics we assume (representative) samples of larger populations. All values we compute, depend on the concrete sample and are subject to uncontrollable random variations. In view of the decisions that are made based on statistical results, one has to conclude that wrong decisions can never completely avoided. It is inevitable, that we make a wrong decision with a (hopefully small) positive probability. If we transfer this situation to statistical testing, we get the situation shown in Table 6.1.

	H_0 is true	H_1 is true
Decision for H_0	correct decision $1 - \alpha$ (sensitivity)	type II error β
Decision for H_1	type I error α (signifikance level)	correct decision $1 - \beta$ (power, specificity)

Table 6.1: Decision situation in case of statistical tests.

Thus, the possible wrong decisions are:

Type I error: Probability of rejecting H_0 although it is true.

Type II error: Probability of not rejection H_0 although it is false.

We describe the errors and their consequences in more detail by means of an example.

Example 6.3. We consider the following situation in medicine: There is an effective and safe therapy, that is in use for many years – a so-called **gold standard**. Now, somebody is convinced, that his new therapeutic approach is even more effective.

In this case, it would be a type **I** error, if one decides against the gold standard and in favor of the new therapy, although the new approach is not better or perhaps even worse. As a consequence, the patients are withheld from a more effective therapy and in cases, where the therapy has adverse effects, it would even harm patients.

In contrast, a type **II** error would be that one keeps the gold standard, although the new approach is actually better. That is, one has missed a chance for an improvement. However, the patients still get an effective and safe therapy.

In this medical application, the type **I** error would be the more serious wrong decision.

We briefly summarize the essential facts about the two errors, where we start with the type I error.

CHALLENGING PERSPECTIVES

Internship opportunities

EADS unites a leading aircraft manufacturer, the world's largest helicopter supplier, a global leader in space programmes and a worldwide leader in global security solutions and systems to form Europe's largest defence and aerospace group. More than 140,000 people work at Airbus, Astrium, Cassidian and Eurocopter, in 90 locations globally, to deliver some of the industry's most exciting projects.

An **EADS internship** offers the chance to use your theoretical knowledge and apply it first-hand to real situations and assignments during your studies. Given a high level of responsibility, plenty of learning and development opportunities, and all the support you need, you will tackle interesting challenges on state-of-the-art products.

We welcome more than 5,000 interns every year across disciplines ranging from engineering, IT, procurement and finance, to strategy, customer support, marketing and sales. Positions are available in France, Germany, Spain and the UK.

To find out more and apply, visit www.jobs.eads.com. You can also find out more on our **EADS Careers Facebook page**.

AIRBUS **ASTRIUM** **CASSIDIAN** **EUROCOPTER**

EADS

Type I error:

- It is inevitable, but controllable.
- The error probability must be set *before* conducting the test!
- α forms the basis for determining the acceptance respectively, rejection region of H_0 .
- In principle, α may be arbitrarily chosen. The standard choice is $\alpha = 0.05$, sometimes also $\alpha = 0.01$ or smaller is used, but very (very) rarely > 0.05 .
- Dependent on the acceptance of H_1 is also called statistically significant ($\alpha = 0.05$), statistically very significant ($\alpha = 0.01$), or statistically extremely or highly significant ($\alpha = 0.001$).

Type II error:

- It is difficult to determine/estimate.
- In general, it holds: The larger α , the smaller β ; that is, a small α and a small β are two competing aims.
- Furthermore, it holds: The larger the sample size n , the smaller is β . In practice, this is the only way to control β and implies the need for a detailed sample size calculation and power analysis.
- However, for sample size calculations a certain prior knowledge about the effect size, the variation of the applied estimators, the type I error and the intended power is required.
- Standard assumptions for sample size calculations are $\alpha = 0.05, 0.01$ and $1 - \beta = 0.8, 0.9$.

The following list contains the necessary steps for conducting a statistical test. In the framework of a clinical trial, one strictly has to follow the given order, as it ensures that nobody can influence the result of the test after the start of the trial.

1. Definition of the hypotheses H_0 and H_1 (one-/two-sided?)
2. Fixing of the type I error (significance level)
3. Selection of an appropriate test T
4. Sample size calculation and power analysis
5. Determination of rejection (K_α) and acceptance (\bar{K}_α) region of H_0
6. Conduct of the experiments and generation of relevant data x_1, \dots, x_n
7. Calculation of the test statistics $t = T(x_1, \dots, x_n)$
8. Decision for H_1 ($t \in K_\alpha$) or H_0 ($t \in \bar{K}_\alpha$)

In practice, the test decision is usually based on the so-called **p value**. It is the (conditional) probability $p = P(T \in K_\alpha | H_0)$ that the value of the test statistics is in the rejection region k_α of H_0 under the assumption that H_0 is true. If p is small, it is unlikely that the data stems from the null hypothesis and thus, one decides for the alternative. More precisely, one decides as follows:

If $p \leq \alpha$: rejection von H_0

If $p > \alpha$: acceptance of H_0 , i.e. rejection of H_1

Note:

The p value is *not* the probability of H_0 . This probability does not exist, because either H_0 is true or false. The p value is also *not* the probability of rejecting the null hypothesis H_0 , although it is true. Furthermore, it is of crucial importance to realize, that statistical significance is not identical to relevance. In case of a very large sample, also smallest difference may be significant without leading to any consequences. Consequentially, it is important, to keep an eye on the effect size besides significance.

In the following example, we show how a statistical test has to be conducted in practice by using the two-sample t test, probably the most frequently applied statistical test.

Example 6.4. Let us assume a (well-defined) population including two (well characterized and disjoint) groups, that we want to compare. We are interested in the expectation (location parameter) of a certain attribute X . We additionally assume that the attribute is normally distributed (at least approximately) and that it has identical variances for both groups; that is, it holds for group I: $X_1 \sim \text{Norm}(\mu_1, \sigma^2)$ and for group II: $X_2 \sim \text{Norm}(\mu_2, \sigma^2)$. We conduct steps 1-8, as listed above, to compare the expectations of the two groups by means of a statistical test:

1. We consider the following hypotheses

$$H_0: \mu_1 = \mu_2 \text{ versus } H_1: \mu_1 \neq \mu_2$$

That is, the alternative is two-sided.

2. We choose the standard type I error (significance level): $\alpha = 0.05$
3. Since we assume a normal distribution for both groups and want to estimate the mean, where the variance is also unknown and has to be estimated, it leads to a t distribution. Consequentially, we select the two-sample t test. Let x_1, \dots, x_{n_1} be the observations of group I and y_1, \dots, y_{n_2} the observations of group II, then the test statistics reads

$$T(x_1, \dots, x_{n_1}; y_1, \dots, y_{n_2}) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\text{AM}(x_1, \dots, x_{n_1}) - \text{AM}(y_1, \dots, y_{n_2})}{\text{SD}(x_1, \dots, x_{n_1}; y_1, \dots, y_{n_2})} \quad (6.1)$$

where

$$\text{SD}(x_1, \dots, x_{n_1}; y_1, \dots, y_{n_2}) = \sqrt{\frac{(n_1 - 1)\tilde{S}(x_1, \dots, x_{n_1}) + (n_2 - 1)\tilde{S}(y_1, \dots, y_{n_2})}{n_1 + n_2 - 2}} \quad (6.1)$$

and \tilde{S} is the sample variance with standardization $\frac{1}{n-1}$

4. For sample size calculation and power analysis we additionally need the (expected) effect size $\delta = |\mu_1 - \mu_2|$, the (expected) variance σ^2 and the wanted power $1 - \beta$.

The influence of the effect size on the sample size is displayed in Figure 6.1, where we consider the standard setup $\beta = 0.2$ and assume $\sigma = 1$ without restriction. The computations were performed by applying function `power.t.test`. As we see, the required sample size clearly decreases with increasing effect size; that is, the larger the effect, the smaller the sample size or we can also put it the other way round: with very large samples we may even verify small (irrelevant) effects. Figure 6.2 shows the dependence of the sample size on the variance, where we assumed an effect size of 1 without restriction. The computations were again performed by means of function `power.t.test`. Thus, the larger the variance, the larger the sample size has to be chosen. In particular, the (expected) ratio $\frac{\delta}{\sigma}$ is of crucial importance, which is also called the (expected) **standardized effect**.

360°
thinking.

Deloitte.

Discover the truth at www.deloitte.ca/careers

© Deloitte & Touche LLP and affiliated entities.

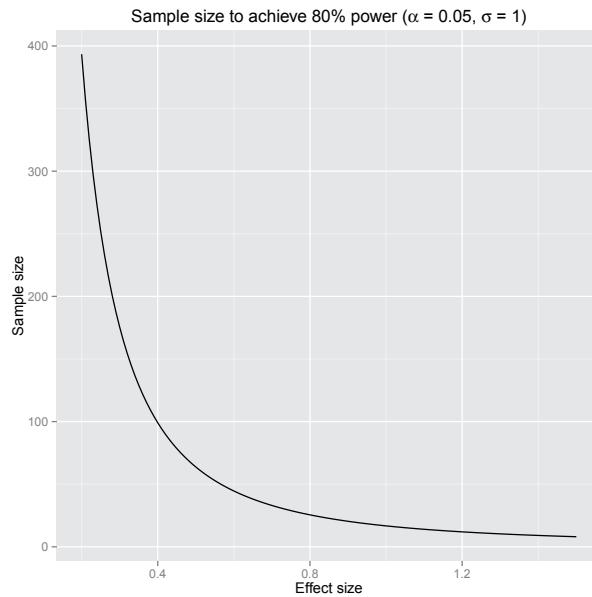


Figure 6.1: Sample size dependent on effect size.

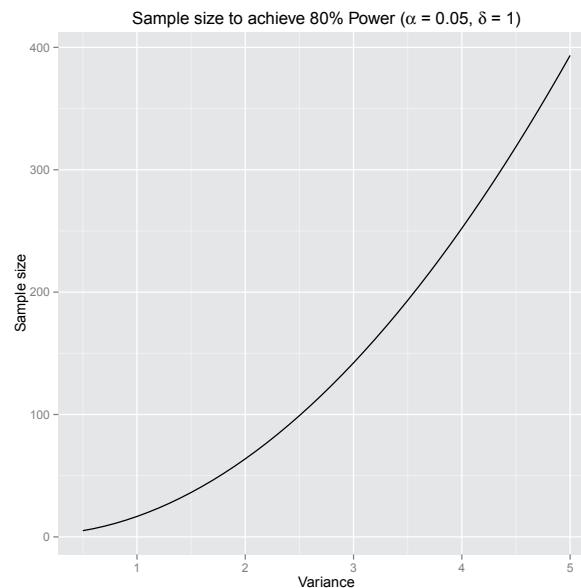


Figure 6.2: Sample size dependent on variance.

5. If we assume the null hypothesis H_0 is true, the test statistic T follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom. This fact we can use to determine the acceptance region \bar{K}_α of H_0 . Because of the symmetry of the situation, we obtain $\bar{K}_\alpha = [-c, c]$, where it must hold

$$P(-c \leq T \leq c | H_0) = 1 - \alpha \quad (6.3)$$

i.e. c is the $(1 - \alpha/2)$ quantile of the $t_{n_1+n_2-2}$ distribution. c is also called **critical value** of the test. Under the assumption $n_1 = n_2 = 20$, we get

```
1 qt(0.975 , df = 38)
```

```
[1] 2.024394
```

6. We conduct the experiment and generate random numbers by means of function `rnorm`. More precisely, we use $X_1 \sim \text{Norm}(0.5; 1)$ for group I and $X_2 \sim \text{Norm}(1.5; 1)$ for group II. As $\sigma = 1$, this corresponds to a standardized effect of $\frac{\delta}{\sigma} = \delta = 1$. Under these assumptions a sample size of 17 per group is sufficient to verify the difference with a power of 80% and a type I error of 5%. We use a sample size of $n = 20$ increasing the power to about 87%.

```
1 ## random numbers for demonstration
2 X1 <- rnorm(n = 20, mean = 0.5, sd = 1)
3 X2 <- rnorm(n = 20, mean = 1.5, sd = 1)
```

7. We compute the test statistic by means of function `t.test`.

```
1 t.test(X1, X2, var.equal = TRUE)$statistic
```

```
t
-5.824201
```

We have to compare this value with the critical value. If it is smaller, one decides for H_0 otherwise for H_1 .

8. Alternatively, we can also determine the p value; that is, the probability that the computed value or a more extreme value of the test statistic occurs under the assumption that H_0 is true. Applying function `t.test`, we obtain

```
1 t.test(X1, X2, var.equal = TRUE)$p.value
```

```
[1] 9.920463e-07
```

The complete output of function `t.test` reads

```
1 t.test(X1, X2, var.equal = TRUE)
```

```
Two Sample t-test

data: X1 and X2
t = -5.8242, df = 38, p-value = 9.92e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.464156 -1.192993
sample estimates:
mean of x   mean of y
-0.02931314 1.79926093
```

The printed 95% confidence interval is an interval for $\mu_1 - \mu_2$ and thus represents the expected effect. One can also use it for the test decision. In the present case, if 0 is inside of the interval, we decide for H_0 otherwise for H_1 .

There are several functions in R that can be used for sample size calculations, in most cases the functions start with `power.`, e.g. `power.t.test` or `power.prop.test`. Moreover, there are various contributed packages providing functions for various tests and models.

SIMPLY CLEVER

ŠKODA



We will turn your CV into
an opportunity of a lifetime



Do you like cars? Would you like to be a part of a successful brand?
We will appreciate and reward both your enthusiasm and talent.
Send us your CV. You will be surprised where it can take you.

Send us your CV on
www.employerforlife.com



Note:

It is common practice, to check the assumptions of statistical tests in pre-tests. This includes the verification of distributional assumptions, especially the normal distribution, or the assumption of equal variances (homogeneity of variances). Beside the methodological problem, that several hypotheses are verified using only one dataset, the pre-tests often have a small power. Therefore, in case of small sample sizes, deviations are only detected with a small probability, whereas in case of large sample sizes, small and for the envisaged test irrelevant deviations are reported. Rasch et al. (2011) show, using t tests as an example, that the practice of pre-tests does not pay off.

6.2 Examples

We start with probably the most frequently applied test, the t test and its variants.

Example 6.5.

- a) In the simplest case, there is a single sample, whose values are realizations of independent and Norm (μ , σ^2) distributed random variables. One studies the unknown location parameter μ , where the variance σ^2 is unknown and thus also has to be estimated. Possible null hypotheses are for instance $\mu = \mu_0$ or $\mu \leq \mu_0$, where $\mu_0 \in \mathbb{R}$ is known and must be specified before performing the test. The corresponding test is called **one-sample t test** and can be computed by function `t.test`.
- b) The basis are two so-called paired samples. This is for instance the case, if we measure a certain attribute from a person at two different time points. That is, we get pairs of values (x_i, y_i) , which we consider as realizations of independent and identical distributed pairs of random variables $(X_i; Y_i)$ ($i = 1, \dots, n, n \in \mathbb{N}$). Furthermore, it holds $D_i = X_i - Y_i \sim \mathcal{N}(\mu, \sigma^2)$, where we are interested in the unknown location parameter μ and the variance σ^2 is unknown. Possible null hypotheses are for example $\mu = \mu_0$ or $\mu \leq \mu_0$ for some given $\mu_0 \in \mathbb{R}$. The test is called **paired t test** and can be computed by function `t.test` using argument `paired = TRUE`. The paired t test is identical to the one-sample t test for the differences of the pairs.
- c) There are two (independent) samples of size n_1 and n_2 from $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ where one is interested in the location parameter and the variance σ^2 (identical for both samples!) is unknown. This so-called (classical) **two-sample t test** is in detail discussed in Example 6.4. It can be applied by means of function `t.test` using argument `var.equal = TRUE`.
- d) The situation is similar as in part (c). In contrast, the two groups may now have different variances. This leads to the so-called **Welch t test**, which can also be computed by function `t.test`.

We use our ICU dataset, which is in more detail explained in Section 2.3. As we have seen in the previous sections, the maximum body temperature of ICU patients can be well described by a normal distribution. We investigate the hypothesis, whether the average ICU patient has an increased body temperature, i.e.

$$H_0: \mu \leq 37.5 \text{ versus } H_1: \mu > 37.5$$

We apply the one-sample t test, where we omit patient 398. We can specify the one-sided alternative by argument `alternative = 'greater'` of function `t.test`. By argument `mu = 37.5`, we define the value, which we want to use for comparison.

```
1 t.test(ICUData$temperature[-398], mu = 37.5, alternative = "greater")
```

```
One Sample t-test

data: ICUData$temperature[-398]
t = 4.1973, df = 498, p-value = 1.599e-05
alternative hypothesis: true mean is greater than 37.5
95 percent confidence interval:
37.63389      Inf
sample estimates:
mean of x
37.72044
```

Based on a significance level of 5%, we can assume that in mean ICU patients have an increased body temperature during their stay on the ICU.

The p value clearly increases, if we add patient 398, but the test still favors the alternative.

```
1 t.test(ICUData$temperature, mu = 37.5, alternative = "greater")
```

```
One Sample t-test

data: ICUData$temperature
t = 2.1027, df = 499, p-value = 0.01799
alternative hypothesis: true mean is greater than 37.5
95 percent confidence interval:
37.5353      Inf
sample estimates:
mean of x
37.6632
```

In the second step, we investigate, whether the maximum body temperature of females (μ_1) and males (μ_2) is significantly different. As we consider two independent groups and as we may assume a normal distribution for both groups, we can apply the two-sample t test. It is an open question, if we may assume equal variances or not. We compute the variances of females and males.

```
1 ## females
2 sd(ICUData$temperature[ICUData$sex == "female"])
```

[1] 1.258335

```
1 ## males
2 sd(ICUData$temperature[ICUData$sex == "male"])
```

[1] 1.946092

The results of both groups are clearly different. However, we did not take care of the male patient 398. Hence, we recompute the standard deviations of males, where we omit patient 398.

```
1 ## males without patient 398
2 sd(ICUData$temperature[-398][ICUData$sex[-398] == "male"])
```

[1] 1.123373



Linköping University –
innovative, highly ranked,
European

Interested in Computer Science? Kick-start your career
with an English-taught master's degree.

[→ Click here!](#)

LiU LINKÖPING
UNIVERSITY



Again, the value of this patient has a strong impact on the result. We will once include patient 398 and once omit patient 398 in our computations. In both cases, we choose the more conservative approach and apply the Welch t test, where the separation of the sexes is done by means of the formula $\text{temperature} \sim \text{sex}$.

```
1 ## with patient 398
2 t.test(temperature ~ sex, data = ICUData)
```

```
Welch Two Sample t-test

data: temperature by sex
t = -0.37638, df = 481.71, p-value = 0.7068
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.3368620 0.2285543
sample estimates:
mean in group female mean in group male
37.62800           37.68215
```

```
1 ## without patient 398
2 t.test(temperature ~ sex, data = ICUData[-398,])
```

```
Welch Two Sample t-test

data: temperature by sex
t = -1.2514, df = 323.73, p-value = 0.2117
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.36618695 0.08144621
sample estimates:
mean in group female mean in group male
37.62800           37.77037
```

That is, the body temperature of females is somewhat lower in mean. However, the test (with and without patient 398) supports the hypothesis that this is only a random variation. Consequentially, we can/must assume that the means of the maximum body temperature of females and males are not different (null hypothesis).

If we can not assume a normal distribution and if the sample size is small to moderate, we should apply different tests.

Example 6.6.

- a) In case of a single sample or two paired samples, we can apply the **Wilcoxon signed rank test** an alternative to the t test. Strictly speaking, the test is only applicable in case of continuous and symmetric distributions. However, in practice, it is also applied in case of discrete and asymmetric distributions. In **R**, it is implemented in function `wilcox.test`.
- b) The counterpart to the two-sample t test is the **Wilcoxon-Mann-Whitney U test**. Strictly speaking, it is applicable in case of two continuous distributions of the same shape. This implies that also the variances of the two distributions should be equal, as in case of the classical t test. However, empirical results show, that a minor violation of this assumption does not influence the Wilcoxon-Mann-Whitney U test. As the test is based on ranks, it is also applied in case of ordinal data. The test is also implemented in function `wilcox.test`.

First, we investigate, whether the average of the maximum body temperature of the ICU patients is increased. We apply function `wilcox.test` and compare the results with and without patient 398. As the confidence interval is not automatically computed, we additionally use argument `conf.int = TRUE`

```
1 ## with patient 398
2 wilcox.test(ICUData$temperature, mu = 37.5, alternative = "greater",
3             conf.int = TRUE)
```

```
Wilcoxon signed rank test with continuity correction

data: ICUData$temperature
V = 69173, p-value = 0.0002349
alternative hypothesis: true location is greater than 37.5
95 percent confidence interval:
 37.59999      Inf
sample estimates:
(pseudo)median
 37.69991
```

```

1 ## without patient 398
2 wilcox.test(ICUData$temperature[-398], mu = 37.5, alternative = "greater",
3             conf.int = TRUE)

```

```

Wilcoxon signed rank test with continuity correction

data: ICUData$temperature [-398]
V = 69173, p-value = 0.0001671
alternative hypothesis: true location is greater than 37.5
95 percent confidence interval:
37.60003      Inf
sample estimates:
(pseudo)median
37.69997

```

As in case of the one-sample t test, we get a significant results and thus must assume that in average the body temperature is increased. However, the result is less influenced by patient 398.

In the second step, we again compare females and males, now using the Wilcoxon-Mann-Whitney U test. For this, we can again apply function `wilcox.test`.

I joined MITAS because
I wanted **real responsibility**



The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com

Month 16
I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements





```
1 ## with patient 398
2 wilcox.test(temperature ~ sex, data = ICUData, conf.int = TRUE)
```

```
Wilcoxon rank sum test with continuity correction

data: temperature by sex
W = 26388, p-value = 0.1833
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
-0.39997835 0.09992453
sample estimates:
difference in location
-0.1000091
```

```
1 ## without patient 398
2 wilcox.test(temperature ~ sex, data = ICUData[-398,], conf.int = TRUE)
```

```
Wilcoxon rank sum test with continuity correction

data: temperature by sex
W = 26212, p-value = 0.1643
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
-0.39995318 0.09997909
sample estimates:
difference in location
-0.199997
```

The results are again in agreement with the t test, where the influence of patient 398 is clearly smaller.

In the following example we are not interested in the mean values but the variances of two independent groups.

Example 6.7.

- We consider two independent groups and are interested in an attribute, which in both groups is normal distributed. In contrast to Example 6.5 we are interested in the variances and not the means. We consider the ratio of the two variances. As noted in Remark 4.28, this leads to an F distribution and the test is called **F test**. We can compute the test by means of function `var.test`.
- A counterpart to the F test based on ranks and thus, not requiring the assumption of normal distributions, is the **Ansari-Bradley test**. It can be applied via function `ansari.test`.

As we have seen above, the variances of the maximum body temperature are different for females and males. We investigate, whether this is a random variation or not. We perform the test with and without patient 398.

```
1 ## mit Patient 398
2 var.test(temperature ~ sex, data = ICUData)
```

```
F test to compare two variances

data: temperature by sex
F = 0.41809, num df = 174, denom df = 324, p-value = 6.066e-10
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.3236546 0.5457195
sample estimates:
ratio of variances
0.4180863
```

```
1 ## ohne Patient 398
2 var.test(temperature ~ sex, data = ICUData[-398,])
```

```
F test to compare two variances

data: temperature by sex
F = 1.2547, num df = 174, denom df = 323, p-value = 0.08265
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.9711595 1.6379524
sample estimates:
ratio of variances
1.254713
```

In case of the variance, the extreme value of patient 398 is clearly more influential than in case of the means. It depends, whether we include patient 398 or not, if we get a significant difference or not. We want to investigate, whether this is also true for the Ansari-Bradley test and for this purpose apply function `ansari.test`.

```
1 ## with patient 398
2 ansari.test(temperature ~ sex, data = ICUData)
```

```
Ansari-Bradley test

data: temperature by sex
AB = 20904, p-value = 0.1688
alternative hypothesis: true ratio of scales is not equal to 1
```

```
1 ## without patient 398
2 ansari.test(temperature ~ sex, data = ICUData[-398,])
```

```
Ansari-Bradley test

data: temperature by sex
AB = 20808, p-value = 0.1477
alternative hypothesis: true ratio of scales is not equal to 1
```

The results show that this is not the case and the influence of patient 398 is clearly smaller. It is confirmed that we can assume equal variances.

**“I studied English for 16 years but...
...I finally learned to speak it in just six lessons”**

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



In the following example, we consider the probability of success assuming a Bernoulli model.

Example 6.8.

- a) There are observations of a binary attribute (values: 0 and 1) and we want to investigate the probability of 1. For the comparison of the relative frequencies with a given value, we can apply the **one-sample binomial test** implemented in function `binom.test`. As in case of the confidence intervals (cf. Example 5.12), we can also use a normal approximation. The corresponding test is provided by function `prop.test`.
- b) If we want to compare two groups with respect to a binary attribute, we can use a 2×2 contingency table (cf. Table 6.2). If we want to find out, whether the distribution of the binary attribute is different for the two groups, it leads to hypergeometric distributions and **Fisher's exact test**. The test is implemented in function `fisher.test`. The asymptotic version is a χ^2 test computable by means of function `chisq.test`. In this case, the cell counts should not be too small. Depending on the reference, the minimum cell count should be somewhere between 1 and 5.

	A	Not A	Sum
B	a	b	a + b
Not B	c	d	c + d
Sum	a + c	b + d	a + b + c + d

Table 6.2: Example of a 2×2 contingency table.

Both functions can also be applied to $r \times s$ contingency tables. In case of large values of r and s , Fisher's exact test is computationally demanding. If the cell counts are not too small, one can instead apply the χ^2 test.

We again use our ICU dataset and investigate the prevalence of liver failure. We want to find out, whether we may assume a prevalence of less than 5%; i.e.

$$H_0 : p \geq 0.05 \text{ versus } H_1 : p < 0.05$$

We apply the exact as well as the asymptotic test, i.e. functions `binom.test` and `prop.test`. The alternative we specify by argument `alternative = 'less'`.

```
1 ## exact test
2 binom.test(20, 500, p = 0.05, alternative = "less")
```

```
Exact binomial test

data: 20 and 500
number of successes = 20, number of trials = 500, p-value = 0.1789
alternative hypothesis: true probability of success is less than 0.05
95 percent confidence interval:
0.00000000 0.05759556
sample estimates:
probability of success
0.04
```

```
1 ## asymptotic test
2 prop.test(20, 500, p = 0.05, alternative = "less")
```

```
1-sample proportions test with continuity correction

data: 20 out of 500, null probability 0.05
X-squared = 0.85263, df = 1, p-value = 0.1779
alternative hypothesis: true p is less than 0.05
95 percent confidence interval:
0.00000000 0.05822552
sample estimates:
p
0.04
```

Both tests yield the same result; that is, we can not be sure that the prevalence is smaller than 5%.

In the next step, we compare the prevalences of females and males. We again apply the exact as well as the asymptotic test provided by functions `fisher.test` and `chisq.test`.

```
1 ## 2x2 contingency table
2 kont.table <- table(ICUData$liver.failure, ICUData$sex)
3 kont.table
```

	female	male
0	168	312
1	7	13

```
1 ## Fisher's exact test
2 fisher.test(kont.table)
```

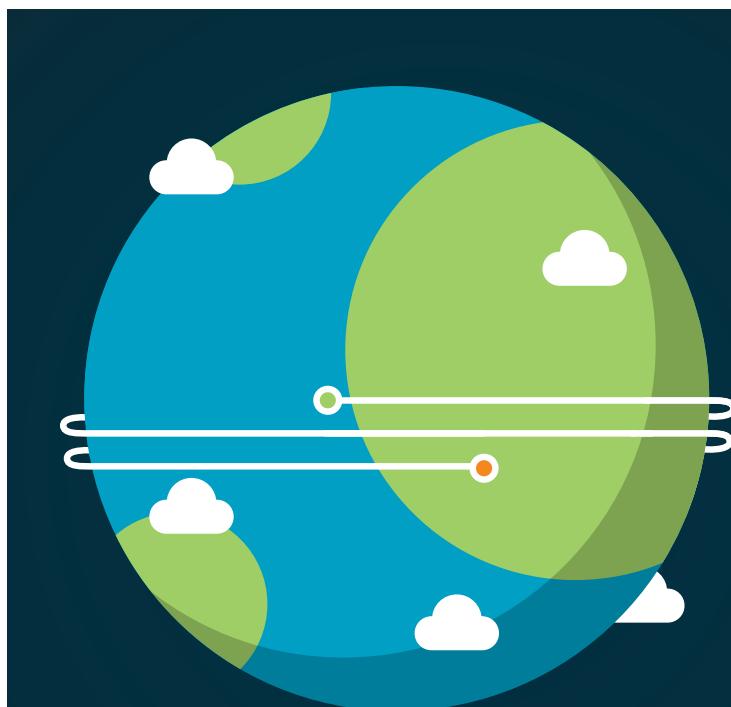
```
Fisher's Exact Test for Count Data

data: kont.table
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.3625184 3.0194493
sample estimates:
odds ratio
1
```

```
1 ## chi^2 test
2 chisq.test(kont.table)
```

```
Pearson's Chi-squared test

data: kont.table
X-squared = 0, df = 1, p-value = 1
```



In the past four years we have drilled
89,000 km
That's more than **twice** around the world.

Who are we?
We are the world's largest oilfield services company¹. Working globally—often in remote and challenging locations—we invent, design, engineer, and apply technology to help our customers find and produce oil and gas safely.

Who are we looking for?
Every year, we need thousands of graduates to begin dynamic careers in the following domains:
■ **Engineering, Research and Operations**
■ **Geoscience and Petrotechnical**
■ **Commercial and Business**

What will you be?

Schlumberger

¹Based on Fortune 500 ranking 2011. Copyright © 2015 Schlumberger. All rights reserved.

Again, both tests yield the same result. Based on the relatively low number of cases with liver failure, we have to assume that the differences are nothing else but random variations; i.e., females and males are equally affected.

Another important test investigates the correlation between two attributes.

Example 6.9. We assume two normally distributed attributes having a linear relationship. That is, we can investigate the strength of the relationship by means of Pearson's correlation ρ . The respective test statistic follows a t distribution with $n - 2$ degrees of freedom. The null hypothesis reads $\rho = 0$. If we can not assume a normal distribution and/or assume a more general monotone relationship, the correlations of Spearman and Kendall are appropriate alternatives. All three test can be computed by function `cor.test`.

We investigate, whether the correlation between maximum body temperature and maximum heart rate is significantly different from 0. We apply function `cor.test` and investigate Pearson's, Spearman's, and Kendall's correlation, where we omit patient 398 in case of Pearson's correlation.

```
1 ## Pearson without patient 398
2 cor.test(ICUData$temperature[-398], ICUData$heart.rate[-398])
```

```
Pearson's product-moment correlation

data: ICUData$temperature[-398] and ICUData$heart.rate[-398]
t = 6.9546, df = 497, p-value = 1.118e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.2156624 0.3757592
sample estimates:
cor
0.2978033
```

```
1 ## Spearman
2 cor.test(ICUData$temperature, ICUData$heart.rate, method = "spearman")
```

```
Spearman's rank correlation rho

data: ICUData$temperature and ICUData$heart.rate
S = 15292000, p-value = 1.521e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.2659957
```

```
1 ## Kendall  
2 cor.test(ICUData$temperature, ICUData$heart.rate, method = "kendall")
```

```
Kendall's rank correlation tau

data: ICUData$temperature and ICUData$heart.rate
z = 6.0032, p-value = 1.935e-09
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.1826903
```

All three cases yield a significant correlation. Unfortunately, we can only test whether the correlation is significantly different from 0 respectively positive or negative, but for instance not, whether the correlation is significantly larger than a given value.

Of course, it happens quite frequently that more than two groups have to be compared.



Click on the ad to read more

Example 6.10.

- a) We consider k groups and compare the groups with respect to some normally distributed attribute. The goal is to find out, whether the means are significantly different. This is called a **one-way ANOVA**, where ANOVA stands for “ANalysis Of Variance”. As in case of the two-sample t Tests, it is called a classical one-way ANOVA, if the variances are assumed equal and a Welch one-way ANOVA, if they are assumed to be different. Both types are implemented in function `oneway.test`.
- b) The rank based counterpart of the one-way ANOVA is the **Kruskal-Wallis test**, sometimes also called non-parametric one-way ANOVA. By adding non-parametric, it is indicated that no parametric model is assumed, in particular, no normal distribution model is required. The test is provided by R function `kruskal.test`.

We investigate the maximum body temperature of our ICU patients with respect to the outcome. In case of the one-way ANOVA, we compare the results with and without patient 398 as well as with and without assuming equal variances. That is, we four times have to apply function `oneway.test`.

```
1 ## with patient 398, classical
2 oneway.test(temperature ~ outcome, data = ICUData, var.equal = TRUE)
```

```
One-way analysis of means

data: temperature and outcome
F = 1.9572, num df = 3, denom df = 496, p-value = 0.1195
```

```
1 ## with patient 398, Welch
2 oneway.test(temperature ~ outcome, data = ICUData)
```

```
One-way analysis of means (not assuming equal variances)

data: temperature and outcome
F = 4.5435, num df = 3.00, denom df = 182.28, p-value = 0.004262
```

```
1 ## without patient 398, classical
2 oneway.test(temperature ~ outcome, data = ICUData[-398,], var.equal = TRUE)
```

```
One-way analysis of means

data: temperature and outcome
F = 5.2203, num df = 3, denom df = 495, p-value = 0.001481
```

```
1 ## without patient 398, Welch
2 oneway.test(temperature ~ outcome, data = ICUData[-398,])
```

```
One-way analysis of means (not assuming equal variances)

data: temperature and outcome
F = 5.3574, num df = 3.00, denom df = 189.68, p-value = 0.001458
```

Assuming equal variances and at the same time including patient 398, leads to no significant difference between the groups. In the three other cases, the means are significantly different; that is, the influence of a single outlier is largest in case of the classical one-way ANOVA.

We apply the Kruskal-Wallis test by means of function `kruskal.test`, where we compute the results with and without patient 398.

```
1 ## with patient 398
2 kruskal.test(temperature ~ outcome, data = ICUData)
```

```
Kruskal-Wallis rank sum test

data: temperature by outcome
Kruskal-Wallis chi-squared = 15.259, df = 3, p-value = 0.001608
```

```
1 ## without patient 398
2 kruskal.test(temperature ~ outcome, data = ICUData[-398,])
```

```
Kruskal-Wallis rank sum test

data: temperature by outcome
Kruskal-Wallis chi-squared = 15.869, df = 3, p-value = 0.001206
```

Once again, the results confirm that single outliers have only a minor effect on rank based procedures. In summary, we can conclude that the means are in some way significantly different. As there are more than two groups, the tests do not answer the questions: which groups and in which way the groups are different as well as how large the effects are.

Example 6.11. We assume that the means or location parameters of more than two groups are significantly different, where we have applied a one-way ANOVA or the Kruskal-Wallis test. In this situation, one usually computes so-called **post hoc tests** in a second step and in addition generates plots of the data. Using this approach, the differences between the groups can be made clear. In case of the one-way ANOVA, there are several possible post hoc tests. I think the most natural choice are pairwise t tests, which can easily be computed by function `pairwise.t.test`. Accordingly, pairwise Wilcoxon-Mann-Whitney U tests are the most natural choice in case of the Kruskal-Wallis test and are computable by means of function `pairwise.wilcox.test`. As several tests are simultaneously performed, one should additionally adjust the significance level or the p values to keep control over the type I error. This is also known as **multiple testing**. In R, the correction of Bonferroni-Holm is applied by default.

We conduct a pairwise comparison of the outcome groups with respect to their maximum body temperature and apply function `pairwise.t.test` as well as function `pairwise.wilcox.test`. In case of the t tests, we assume different variances (Welch t test) and omit patient 398.

Excellent Economics and Business programmes at:


/ university of
groningen




**“The perfect start
of a successful,
international career.”**

[CLICK HERE](http://www.rug.nl/feb/education)
 to discover why both socially
 and academically the University
 of Groningen is one of the best
 places for a student to be

www.rug.nl/feb/education



```

1 ## pairwise Welch t tests without patient 398
2 pairwise.t.test(ICUData$temperature[-398], ICUData$outcome[-398],
3   pool.sd = FALSE)

```

```

Pairwise comparisons using t tests with non-pooled SD

data: ICUData$temperature[-398] and ICUData$outcome[-398]

            died    home   other hospital
home          0.0410 -      -
other hospital 1.0000 0.0459 -
secondary care/rehab 1.0000 0.0036 1.0000

P value adjustment method: holm

```

```

1 ## pairwise Wilcoxon-Mann-Whitney U tests
2 pairwise.wilcox.test(ICUData$temperature, ICUData$outcome)

```

```

Pairwise comparisons using Wilcoxon rank sum test

data: ICUData$temperature and ICUData$outcome

            died    home   other hospital
home          0.0534 -      -
other hospital 1.0000 0.0534 -
secondary care/rehab 1.0000 0.0024 1.0000

P value adjustment method: holm

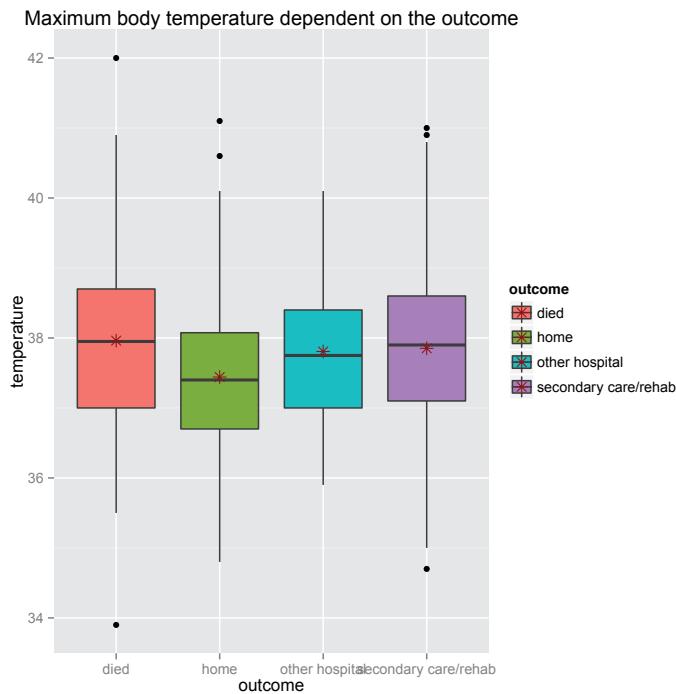
```

That is, it is mainly group “home” that differs from the other groups. We plot the outcome groups by means of box-and-whisker plots, where we add stars representing the arithmetic means. We use function `ggplot` of package "ggplot2" (Wickham (2009)) and omit patient 398, who belongs to group “died”.

```

1 ggplot(data=ICUData[-398,], aes(x=outcome, y=temperature, fill=outcome)) +
2   geom_boxplot() +
3   stat_summary(fun.y=mean, colour="darkred", geom="point", shape=8, size = 3) +
4   ggtitle("Maximum body temperature dependent on the outcome")

```



That is, the maximum body temperature of group “home” is in mean smaller than in case of the other groups.

Note:

If we compare two or more independent groups (samples), the power of the comparison is provided by the smallest group. Consequentially, the sizes of the groups in studies should be as similar as possible. In case of equal sizes, the design is called balanced.

In the last example, we will introduce distribution tests, where we will restrict our considerations to testing of normality.

Example 6.12. Let P some unknown distribution. We want to answer the question, whether P is a normal distribution, i.e. $P \in \mathcal{P} = \{\text{Norm}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma \in (0, \infty)\}$. Thus, we consider the following hypotheses

$$H_0 : P \in \mathcal{P} \text{ versus } H_1 : P \notin \mathcal{P}$$

There are several tests for this situation. The reason for it is the alternative, which is very large and can not be covered by a single test. In particular, in case of small to moderate samples, one should rather use plots such as qq plots to verify the normal distribution. In case of large or very large samples, one can often argue with the central limit theorem and thus assume an approximative normal distribution. Available tests of normality in R are: Shapiro-Wilk test, Kolmogorov-Smirnov test resp. Lilliefors test, Anderson-Darling test, Cramér-von Mises test, Shapiro-Frankia test, Jarque-Bera test, D'Agostino test, etc.

Beside the base function `shapiro.test`, we apply functions `LillieTest`, `CramerVonMisesTest` and `ShapiroFranciaTest` of package “`DescTools`” (Signorell et mult. al. (2015)) to find out, whether the maximum body temperature follows a normal distribution. Since many tests are strongly influenced by outliers, we omit patient 398.

```
1 library(DescTools)
2 ## Shapiro-Wilk test
3 shapiro.test(ICUData$temperature[-398])
```

```
Shapiro-Wilk normality test

data: ICUData$temperature[-398]
W = 0.99189, p-value = 0.008013
```

```
1 ## Lilliefors (Kolmogorov-Smirnov) test
2 LillieTest(ICUData$temperature[-398])
```

```
Lilliefors (Kolmogorov-Smirnov) normality test

data: ICUData$temperature[-398]
D = 0.048498, p-value = 0.007001
```

..... Alcatel-Lucent 

www.alcatel-lucent.com/careers

What if you could build your future and create the future?

One generation's transformation is the next's status quo.
In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".

```
1 ## Cramer-von Mises test
2 CramerVonMisesTest(ICUData$temperature [-398])
```

```
Cramer-von Mises normality test

data: ICUData$temperature [-398]
W = 0.16686, p-value = 0.01426
```

```
1 ## Shapiro-Francia test
2 ShapiroFranciaTest(ICUData$temperature [-398])
```

```
Shapiro-Francia normality test

data: ICUData$temperature [-398]
W = 0.99131, p-value = 0.006149
```

In the present case, all tests reject the normal distribution, which is probably caused by the relatively large sample. As the deviation is quite small, as our analysis in Section 5.2 shows, we can neglect these results in view of the central limit theorem. This is also confirmed by the fact, that t test, one-way ANOVA, Wilcoxon-Mann-Whitney U test and Kruskal-Wallis test yield comparable results, if we omit patient 398.

Note:

In most cases, distribution tests are not useful (Rasch et al. (2011)). In particular, one should avoid to apply Kolmogorov-Smirnov Tests (Schoder et al. (2006); Ghasemi and Zahediasl (2012)).

6.3 Exercises

Always describe and briefly explain the results. In case of exercises 2–8, use the ICU dataset and choose appropriate functions for the computations.

1. In a randomized and controlled trial, a new treatment to avoid the communication of HIV was studied. There were no significant differences between the new treatment and a control group. The ratio between newly occurring infections was 1:0, where the 95% confidence interval was [0:63; 1:58]. Based on this result, can you be sure that the new treatment has no effect? In this context, what could be the meaning of “Absence of Evidence Is Not Evidence of Absence”?
2. Verify the conjecture that more males than females are treated on ICUs. Formulate the hypotheses and decide between them by applying an exact as well as an asymptotic test.

3. Investigate, whether males die more frequently on the ICU than females. Compute an exact as well as an asymptotic test to check this. As starting point, use the 2×2 contingency table generated by

```
table(ICUData$sex, ICUData$outcome == "died")
```

4. Assume a normal distribution for the logarithmized bilirubin values and compare the mean log-concentrations of bilirubin for the ICU patients with and without liver failure applying t tests. Do you think the classical or the Welch t test is more appropriate? In a second step, apply the Wilcoxon-Mann-Whitney U test and compute the test once with and once without taking the logarithm of the bilirubin values. Compare the two results? What do you detect? What is the reason for it?
5. Compare the average length of stay of females and males by applying the Wilcoxon-Mann-Whitney U test.
6. Apply an appropriate test to investigate, whether there is a significant correlation between age and the SAPS II score. Which coefficient of correlation seems to be appropriate for you in this situation and why?
7. Assume that the maximum heart rate of ICU patients can be described by a normal distribution. Apply the Welch one-way ANOVA to find out, whether the means of the outcome groups are significantly different. If you get a significant result, study the differences in more detail by means of post hoc tests and a plot.
8. Consider the SAPS II scores of the ICU patients and compare the averages of the surgery groups. Apply the Kruskal-Wallis test. If there are significant differences, study the differences in more detail by means of post hoc tests and a plot.
9. Use the chem dataset of package "MASS" (Venables and Ripley (2002)), which can be loaded by the following R code

```
library(MASS)
data(chem)
```

Use the Shapiro-Wilk test as well as the Lilliefors (Kolmogorov-Smirnov), the Cramer-von Mises and the Shapiro-Francia test of package "DescTools" (Signorell et al. (2015)) to verify, whether the data follows a normal distribution. What do you observe? In addition, use a qq plot to check the assumption of normality, which can be generated by functions `qqnorm` and `qqline`. Do you think the plot confirms the tests? Repeat the tests of normality, but this time omit observation 17. Interpret the results.

Software versions

For generating this book the following software versions have been used:

- R version 3.2.1 Patched (2015-06-28 r68602), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=de_DE.UTF-8, LC_NUMERIC=C, LC_TIME=de_DE.UTF-8, LC_COLLATE=de_DE.UTF-8, LC_MONETARY=de_DE.UTF-8, LC_MESSAGES=de_DE.UTF-8, LC_PAPER=de_DE.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=de_DE.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, methods, stats, stats4, utils
- Other packages: DescTools 0.99.11, distr 2.5.3, distrEx 2.5, distrMod 2.5.3, ggplot2 1.0.1, knitr 1.10.5, manipulate 1.0.1, MASS 7.3-41, MKmisc 0.99, RandVar 0.9.2, RColorBrewer 1.1-2, fsmisc 1.0-27, startupmsg 0.9, SleaveListingUtils 0.6.2
- Loaded via a namespace (and not attached): boot 1.3-16, colorspace 1.2-6, DEoptimR 1.0-2, digest 0.6.8, evaluate 0.7, foreign 0.8-64, formatR 1.2, grid 3.2.1, gtable 0.1.2, labeling 0.3, magrittr 1.5, munsell 0.4.2, mvtnorm 1.0-2, plyr 1.8.3, proto 0.3-10, Rcpp 0.11.6, reshape2 1.4.1, robustbase 0.92-4, scales 0.2.5, stringi 0.4-1, stringr 1.0.0, tools 3.2.1



Maastricht University *Leading in Learning!*

**Join the best at
the Maastricht University
School of Business and
Economics!**



Top master's programmes

- 33rd place Financial Times worldwide ranking: MSc International Business
- 1st place: MSc International Business
- 1st place: MSc Financial Economics
- 2nd place: MSc Management of Learning
- 2nd place: MSc Economics
- 2nd place: MSc Econometrics and Operations Research
- 2nd place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

Visit us and find out why we are the best!
Master's Open Day: 22 February 2014

www.mastersopenday.nl

Bibliography

Box, G. and Draper, N. (1987). *Empirical model-building and response surfaces*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley. (Cited on page 19)

Chambers, J. (2000). Stages in the Evolution of S. <http://ect.bell-labs.com/sl/S/history.html> [Last access 29.08.2015]. (Cited on pages 10 and 11)

Chambers, J. (2008). *Software for Data Analysis: Programming with R*. Springer. (Cited on pages 10 and 11)

Dalgaard, P. (2010). [R] R 2.11.0 is released. <https://stat.ethz.ch/pipermail/r-help/2010-April/236141.html> [Last access 29.08.2015]. (Cited on page 11)

Dalgaard, P. (2013). [R] R 3.0.0 is released. <https://stat.ethz.ch/pipermail/r-help/2013-April/350751.html> [Last access 29.08.2015]. (Cited on page 11)

Dalgaard, P. (2015). R 3.2.1 liftoff. <https://stat.ethz.ch/pipermail/r-announce/2015/000586.html> [Last access 29.08.2015]. (Cited on page 11)

Delignette-Muller, M.L. and Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34. (Cited on page 145)

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *GenomeBiology*, 5:R80. <http://www.bioconductor.org>. (Cited on page 13)

Ghasemi, A. and Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for nonstatisticians. *Int J Endocrinol Metab*, 10(2):486–489. (Cited on page 212)

Globus, A. (1994). Principles of information display for visualization practitioners. http://www2.cs.uregina.ca/~rbm/cs100/notes/spreadsheets/tufte_paper.html [Last access 29.08.2015]. (Cited on page 91)

Grosjean, P. (2011). R GUI Projects Overview. http://www.sciviews.org/_rgui/projects/overview.html [Last access 29.08.2015]. (Cited on page 15)

Grosjean, P. (2012). IDE/Script Editors. http://www.sciviews.org/_rgui/projects/Editors.html [Last access 29.08.2015]. (Cited on page 14)

Hagemann, O. (2014). TSH-basal. <http://www.laborlexikon.de/Lexikon/Infoframe/t/TSH-basal.htm> [Last access 29.08.2015]. (Cited on page 124)

Hamilton, T.E., Davis, S., Onstad, L., and Kopecky, K.J. (2008). Thyrotropin levels in a population with no clinical, autoantibody, or ultrasonographic evidence of thyroid disease: implications for the diagnosis of subclinical hypothyroidism. *J. Clin. Endocrinol. Metab.*, 93(4):1224–1230. (Cited on page 124)

Harjutsalo, V., Sund, R., Knip, M., and Groop, P. (2013). Incidence of type 1 diabetes in finland. *JAMA*, 310(4):427–428. (Cited on page 115)

Harrower, M. and Brewer, C.A. (2003). Colorbrewer.org: An online tool for selecting color schemes for maps. *The Cartographic Journal*, 40(1):27–37. (Cited on pages 80 and 92)

Hornik, K. (2008). *The Past, Present, and Future of the R Project*. <http://www.statistik.uni-dortmund.de/useR-2008/slides/Hornik.pdf> [Last access 29.08.2015]. (Cited on pages 11 and 11)

Hornik, K. (2015). R FAQ. <http://cran.r-project.org/doc/manuals/R-FAQ.html> [Last access 29.08.2015]. (Cited on page 11)

Iacus, S.M., Urbanek, S., Goedman, R.J., and Ripley, B. (2015). R for Mac OS X FAQ. <http://cran.r-project.org/bin/macosx/RMacOSX-FAQ.html> [Last access 29.08.2015]. (Cited on page 13)

Ihaka, R. (1997). R-beta: New R Version for Unix. <https://stat.ethz.ch/pipermail/r-help/1997-December/001929.html> [Last access 29.08.2015]. (Cited on page 11)

Ihaka, R. (1998). R: Past and Future History. Technical report, Statistics Department, The University of Auckland. <https://www.stat.auckland.ac.nz/~ihaka/downloads/Interface98.pdf> [Last access 29.08.2015]. (Cited on page 11)

Ihaka, R. (2003). Colour for Presentation Graphics. <http://www.stat.auckland.ac.nz/~ihaka/downloads/DSC-Color-Slides.pdf> [Last access 29.08.2015]. (Cited on page 79)

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Med.*, 2(8):e124. (Cited on page 139)

Kohl, M. (2015). *MKmisc: Miscellaneous functions from M.* Kohl. R package version 0.99. (Cited on pages 166 and 145)

Kohl, M. and Ruckdeschel, P. (2010). R package distrMod: S4 classes and methods for probability models. *Journal of Statistical Software*, 35(10):1–27. (Cited on pages 145, 146, 147, 154, 163, 168 and 172)

Limpert, E. and Stahel,W.A. (2011). Problems with using the normal distribution—and ways to improve quality and efficiency of data analysis. *PLoS ONE*, 6(7):e21403. (Cited on page 126)

Limpert, E., Stahel, W.A., and Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51:341–352. (Cited on page 126)

Muenchen, R.A. (2015). The Popularity of Data Analysis Software. <http://r4stats.com/articles/popularity/> [Last access 29.08.2015]. (Cited on page 12)

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes.* R package version 1.1–2. (Cited on pages 80, 83 and 97)

Plummer, M. (2015). Index of /bin/linux/redhat. <http://cran.at.r-project.org/bin/linux/redhat/> [Last access 29.08.2015]. (Cited on page 13)

Empowering People. Improving Business.

BI Norwegian Business School is one of Europe's largest business schools welcoming more than 20,000 students. Our programmes provide a stimulating and multi-cultural learning environment with an international outlook ultimately providing students with professional skills to meet the increasing needs of businesses.

BI offers four different two-year, full-time Master of Science (MSc) programmes that are taught entirely in English and have been designed to provide professional skills to meet the increasing need of businesses. The MSc programmes provide a stimulating and multi-cultural learning environment to give you the best platform to launch into your career.

- MSc in Business
- MSc in Financial Economics
- MSc in Strategic Marketing Management
- MSc in Leadership and Organisational Psychology

www.bi.edu/master

BI NORWEGIAN BUSINESS SCHOOL

EFMD
EQUIS
ACCREDITED

R Core Team (2015a). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org>. (Cited on pages 9, 10, 88 and 145)

R Core Team (2015b). *R Data Import/Export*. <http://cran.r-project.org/doc/manuals/r-release/R-data.pdf>. (Cited on page 22)

R Core Team (2015c). R Developer Page. <https://developer.r-project.org/> [Last access 29.08.2015]. (Cited on page 11)

R Core Team (2015d). *R Installation and Administration*. R Foundation for Statistical Computing, Vienna, Austria. <http://cran.r-project.org/doc/manuals/R-admin.pdf>. (Cited on page 13)

Ranke, J. (2015). Index of /bin/linux/debian. <http://cran.r-project.org/bin/linux/debian/> [Last access 29.08.2015]. (Cited on page 13)

Rasch, D., Kubinger, K.D., and Moder, K. (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Stat. Papers*, 52:219–231. (Cited on pages 189 and 212)

Ripley, B.D. and Murdoch, D.J. (2015). R for Windows FAQ. <http://cran.r-project.org/bin/windows/base/rw-FAQ.html> [Last access 29.08.2015]. (Cited on page 13)

Ruckdeschel, P., Kohl, M., Stabla, T., and Camphausen, F. (2006). S4 Classes for Distributions. *R News*, 6(2):2–6. http://CRAN.R-project.org/doc/Rnews/Rnews_2006-2.pdf. (Cited on pages 105, 108, 113, 117, 122, 124, 127, 132, 136, 146 and 150)

Rutter, M. (2015). Index of /bin/linux/ubuntu. <http://cran.at.r-project.org/bin/linux/ubuntu/> [Last access 29.08.2015]. (Cited on page 13)

Schoder, V., Himmelmann, A., and Wilhelm, K.P. (2006). Preliminary testing for normality: some statistical aspects of a common concept. *Clin. Exp. Dermatol.*, 31(6):757–761. (Cited on page 212)

Signorell et mult. al. (2015). *DescTools: Tools for Descriptive Statistics*. R package version 0.99.11. (Cited on pages 46, 47, 54, 59, 61, 62, 210 and 213)

Steuer, D. (2015). Index of /bin/linux/suse. <http://cran.at.r-project.org/bin/linux/suse/> [Last access 29.08.2015]. (Cited on page 13)

The Linux Foundation (2015). Linux Foundation Announces R Consortium to Support Millions of Users Around the World. <http://www.linuxfoundation.org/news-media/announcements/2015/06/linux-foundation-announces-r-consortium-support-millions-users> [Last access 29.08.2015]. (Cited on page 12)

The R Foundation (2015a). Contributors. <http://www.r-project.org/contributors.html> [Last access 29.08.2015]. (Cited on page 12)

The R Foundation (2015b). The R Foundation. <http://www.r-project.org/foundation/> [Last access 29.08.2015]. (Cited on page 12)

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0. (Cited on pages 144, 146, 163 and 213)

Wald, A. (1980). *A method of estimating plane vulnerability based on damage of survivors*. Center for Naval Analyses, crc 432 edition. (Cited on page 19)

WHO (2015a). Country and regional data on diabetes – WHO European Region. http://www.who.int/diabetes/facts/world_figures/en/index4.html [Last access 29.08.2015]. (Cited on page 106)

WHO (2015b). Diabetes. <http://www.who.int/mediacentre/factsheets/fs312/en/> [Last access 29.08.2015]. (Cited on pages 101 and 111)

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer New York. <http://had.co.nz/ggplot2/book>. (Cited on pages 32, 32, 41, 43, 51, 67, 69, 72, 77, 78, 85, 92, 93, 96, 97, 147 and 208)

Wikipedia (2015a). Andorra–Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/w/index.php?title=Andorra&oldid=678553931> [Last access 30.08.2015]. (Cited on page 106)

Wikipedia (2015b). Bilirubin – Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/w/index.php?title=Bilirubin&oldid=670695366> [Last access 08.08.2015]. (Cited on page 28)

Wikipedia (2015c). Demographics of Finland – Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Demographics_of_Finland&oldid=672169947 [Last access 30.08.2015]. (Cited on page 115)

Wikipedia (2015d). Human height – Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Human_height&oldid=678344194 [Last access 30.08.2015]. (Cited on page 121)

Wikipedia (2015e). Intelligence quotient – Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Intelligence_quotient&oldid=678589920 [Last access 30.08.2015]. (Cited on page 122)

Wikipedia (2015f). Reference range – Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Reference_range&oldid=676363791 [Last access 30.08.2015]. (Cited on page 124)

Wikipedia (2015g). SAPS II – Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=SAPS_II&oldid=645022232 [Last access 22.06.2015]. (Cited on page 28)

Xie, Y. (2013). animation: An R package for creating animations and demonstrating statistical methods. *Journal of Statistical Software*, 53(1):1–27. (Cited on page 89)

Xie, Y. (2015). *knitr: A general-purpose package for dynamic report generation in R*. R package version 1.10.5, <http://CRAN.R-project.org/package=knitr>. (Cited on page v)

Zeileis, A., Hornik, K., and Murrell, P. (2009). Escaping RGBland: Selecting Colors for Statistical Graphics. *Computational Statistics & Data Analysis*, 53:3259–3270. (Cited on page 80)

Need help with your dissertation?

Get in-depth feedback & advice from experts in your topic area. Find out what you can do to improve the quality of your dissertation!

Get Help Now



Go to www.helpmyassignment.co.uk for more info

Helpmyassignment

Index

Symbols

: 10, 19, 91, 110, 179, 182, 196, 209
 [61, 72, 95, 124
 <- 26
 == 95, 208
 2×2 contingency table 196, 208
 2σ rule 120, 122
 \$ 29
?barplot 31
-quantile 18, 35, 36, 41, 42, 78, 98, 99, 103, 104, 105,
 112, 116, 118, 135, 138, 150, 158, 159, 161, 163,
 164, 185
 φ -coefficient 18, 47, 48
 χ^2 distribution 99, 126, 133, 159, 163
 χ^2 test 177, 196

A

abs 38
 Absolute continuity 117, 164, 167, 175
 Absolute frequency 18, 29, 65, 111, 135, 139, 142, 145,
 164, 166, 178
 cross table 44, 45, 46
 Access operator 26, 95, 110
 negative index 26, 95, 110
 Access operator \$, 26, 95, 110
 add = TRUE 121
 aes 33
 Alpha blending 51, 96
 alternative = 'greater' 188
 Alternative (hypothesis) 45, 177, 178, 179, 182, 184,
 187, 190, 199
 alternative = 'less' 196
 annotate 149
 Ansari-Bradley test 177, 193, 194
 ansari.test 193, 194
 Arithmetic mean 18, 53

confidence interval 18, 53
 efficient 18, 53
 logarithmized observations 18, 53

Download free eBooks at bookboon.com

unbiased 18, 53

Assignment 26, 96
 Assignment operator 26, 95, 110
 Attribute 21, 37, 135, 182, 187, 193, 196, 201
 categorical 26
 metric 26
 qualitative 26
 quantitative 26

B

bar chart 34, 65, 78, 83, 85, 90, 97
 bar plot 31, 32, 33, 34
 barplot 30, 31, 48, 78, 97
 base 12, 13, 17, 33, 55, 88, 206, 209, 213
 Base packages 209
 Bernoulli distribution 98, 100, 138, 164, 172
 beside = TRUE 49
 bilirubin 28, 55, 59, 175, 208
 Binom 98, 102, 105, 115, 136
 binomCI 165
 BinomFamily 167
 Binomial distribution 98, 100, 102
 binom.test 196
 binwidth 66
 Bitmap 88
 bmp 88
 boot 12, 209
 box-and-whisker plot 18, 40, 41, 63, 78
 boxplot 40, 92, 97
 breaks 65
 brewer.pal 83

C

c 36, 165, 185, 187, 196
 ceiling 35
 central limit theorem 119, 123, 159, 161, 163, 164, 205,
 207
 cex.points 114
 character 95
 check.names 29

check.names = FALSE 29
 chem 208
 chisq.test 196, 197
 class 12, 130
 cluster 12
 codetools 12
 Coefficient of variation 58
 col 85
 col2rgb 84
 ColorBrewer 80, 83, 85, 92, 93, 212

- diverging color palettes 80, 83, 85, 92, 93, 212
- qualitative color palettes 80, 83, 85, 92, 93, 212
- selection criteria 80, 83, 85, 92, 93, 212
- sequential color palettes 80, 83, 85, 92, 93, 212

 colors 35, 49, 51, 79, 80, 81, 82, 83, 84, 85, 92, 93, 95, 96, 97
 compiler 12
 confidence interval 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 arithmetic mean 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 confidence bounds 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 confidence level 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 CvM-MD estimator 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 MAD 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 maximum estimate error 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 MD estimator 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 median 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 normal approximation 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 one-sided 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 point estimator 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 relative frequency 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 variance 138, 139, 157, 158, 159, 160, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 172, 173, 175, 176, 186, 191, 207
 conf.int 160, 191
 confint 161, 162, 167, 170
 conf.int = TRUE 191
 ContCoef 47
 contingency coefficient 18, 47
 contingency table 44, 196, 208
 Continuity correction 138
 continuous distribution 118
 continuous probability distribution 118
 continuous random variable 117
 contributed packages 13, 88

- Installation 13, 88
- Installation with RStudio 13, 88

 cor 50, 199
 cor.test 199
 covariance 74, 136
 Cramér's V 18, 47, 48
 CramerV 47
 Cramér-von-Mises distance 154
 CramerVonMisesTest 206
 Cross table 18
 cumulative distribution function 42, 43, 72, 99, 103, 105, 117, 118, 135, 150
 curve 69, 118, 120, 128, 130

CvM-MD estimator 138, 154, 170, 171, 175
 confidence interval 138, 154, 170, 171, 175

D

data.frame 24, 26, 29
 Data Import 22, 213
 check 22, 213
 data structure 22, 213
 RStudio 22, 213
 text file 22, 213
 datasets 12, 24, 175, 209
 dbinom 103
 density 18, 67, 68, 69, 78, 118, 119, 120, 123, 125, 126, 128, 130, 133, 134, 135, 144, 147, 148
 density estimation 18, 67, 135
 density plot 78, 147
 descriptive statistics 18, 19, 29, 40, 46, 142
 goal 18, 19, 29, 40, 46, 142
 DescTools 46, 47, 55, 59, 61, 63, 206, 209, 213
 dev.off 89

dexp 127
 dgamma 127
 dhyper 108
 digits 46
 discrete distribution 99, 100, 111, 114, 164
 discrete probability distribution 99, 102
 discrete random variable 99, 100
 display.brewer.all 80
 distr 105, 109, 113, 117, 121, 124, 127, 132, 136, 146, 150, 209
 distribution 13, 22, 29, 42, 43, 53, 54, 57, 60, 61, 62, 63, 65, 66, 67, 72, 73, 78, 98, 99, 100, 101, 102, 103, 105, 106, 107, 108, 109, 111, 112, 113, 114, 115, 117, 118, 119, 120, 121, 123, 124, 125, 126, 127, 128, 129, 130, 132, 133, 134, 135, 136, 137, 138, 141, 142, 143, 145, 146, 147, 150, 152, 153, 154, 157, 158, 159, 161, 163, 164, 170, 172, 175, 176, 182, 184, 185, 187, 188, 190, 193, 196, 199, 201, 205, 206, 207, 208, 212



Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations.

Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
 Visit us at www.skf.com/knowledge

SKF



- left-skewed 13, 22, 29, 42, 43, 53, 54, 57, 60, 61, 62, 63, 65, 66, 67, 72, 73, 78, 98, 99, 100, 101, 102, 103, 105, 106, 107, 108, 109, 111, 112, 113, 114, 115, 117, 118, 119, 120, 121, 123, 124, 125, 126, 127, 128, 129, 130, 132, 133, 134, 135, 136, 137, 138, 141, 142, 143, 145, 146, 147, 150, 152, 153, 154, 157, 158, 159, 161, 163, 164, 170, 172, 175, 176, 182, 184, 185, 187, 188, 190, 193, 196, 199, 201, 205, 206, 207, 208, 212
- leptokurtic 13, 22, 29, 42, 43, 53, 54, 57, 60, 61, 62, 63, 65, 66, 67, 72, 73, 78, 98, 99, 100, 101, 102, 103, 105, 106, 107, 108, 109, 111, 112, 113, 114, 115, 117, 118, 119, 120, 121, 123, 124, 125, 126, 127, 128, 129, 130, 132, 133, 134, 135, 136, 137, 138, 141, 142, 143, 145, 146, 147, 150, 152, 153, 154, 157, 158, 159, 161, 163, 164, 170, 172, 175, 176, 182, 184, 185, 187, 188, 190, 193, 196, 199, 201, 205, 206, 207, 208, 212
- platykurtic 13, 22, 29, 42, 43, 53, 54, 57, 60, 61, 62, 63, 65, 66, 67, 72, 73, 78, 98, 99, 100, 101, 102, 103, 105, 106, 107, 108, 109, 111, 112, 113, 114, 115, 117, 118, 119, 120, 121, 123, 124, 125, 126, 127, 128, 129, 130, 132, 133, 134, 135, 136, 137, 138, 141, 142, 143, 145, 146, 147, 150, 152, 153, 154, 157, 158, 159, 161, 163, 164, 170, 172, 175, 176, 182, 184, 185, 187, 188, 190, 193, 196, 199, 201, 205, 206, 207, 208, 212
- right-skewed 13, 22, 29, 42, 43, 53, 54, 57, 60, 61, 62, 63, 65, 66, 67, 72, 73, 78, 98, 99, 100, 101, 102, 103, 105, 106, 107, 108, 109, 111, 112, 113, 114, 115, 117, 118, 119, 120, 121, 123, 124, 125, 126, 127, 128, 129, 130, 132, 133, 134, 135, 136, 137, 138, 141, 142, 143, 145, 146, 147, 150, 152, 153, 154, 157, 158, 159, 161, 163, 164, 170, 172, 175, 176, 182, 184, 185, 187, 188, 190, 193, 196, 199, 201, 205, 206, 207, 208, 212
- distrMod 145, 146, 147, 154, 162, 167, 170, 209, 212
- distrModOptions 155
- dlnorm 123
- dnbnom 112
- dnorm 120
- do.points = FALSE 44, 72
- dpois 115
- E**
- ecdf 43, 72
- empirical cumulative distribution function 42, 43, 72, 135, 150
- empirical frequency distribution 29
- Encapsulated PostScript 88
- eps 88
- Erlang distribution 99, 126
- Estimation 138, 140
- estimator 138, 140, 141, 142, 144, 145, 146, 153, 154, 155, 157, 158, 163, 164, 168, 170, 171, 172, 175, 176
- bias-free 138, 140, 141, 142, 144, 145, 146, 153, 154, 155, 156, 157, 158, 163, 164, 168, 170, 171, 172, 175, 176
- consistent 138, 140, 141, 142, 144, 145, 146, 153, 154, 155, 156, 157, 158, 163, 164, 168, 170, 171, 172, 175, 176
- efficient 138, 140, 141, 142, 144, 145, 146, 153, 154, 155, 156, 157, 158, 163, 164, 168, 170, 171, 172, 175, 176
- estimator construction 144
- Example 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- body height in Germany 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- diabetes in Andorra 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205

- failure rate of bulbs 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- hospital length of stay 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- intelligence quotient 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- life expectancy of a battery 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- normal range of thyrotropin (TSH) 20, 22, 35, 36, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- opinion poll 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- prevalence of diabetes 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- quality control of bulbs 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- type 1 diabetes in Finland 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- wind speed 20, 22, 35, 36, 37, 41, 53, 56, 89, 100, 101, 103, 106, 108, 111, 112, 114, 115, 120, 124, 127, 130, 142, 145, 158, 164, 168, 172, 178, 179, 180, 182, 187, 191, 193, 196, 199, 201, 203, 205
- Exp 126, 136, 213
- expectation 99, 118, 120, 123, 124, 135, 136, 142, 145, 147, 182
- Exponential distribution 99
- expr 121
- extreme value distribution 130
- F**
- factor 29, 108, 165
- F distribution 99, 134, 193
- fill 92
- finite sample correction 108
- Fisher's exact test 177, 196
- fisher.test 196, 197
- fitdistr 146, 147, 161
- fitdistrplus 145, 210
- for 9, 10, 11, 12, 13, 14, 15, 16, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 34, 35, 36, 38, 40, 41, 46, 47, 50, 52, 54, 55, 56, 59, 61, 62, 63, 66, 69, 77, 78, 79, 80, 81, 82, 84, 85, 87, 88, 89, 92, 95, 97, 99, 100, 101, 103, 104, 106, 108, 109, 111, 112, 114, 115, 117, 118, 120, 123, 124, 126, 128, 129, 130, 135, 136, 138, 139, 140, 142, 143, 144, 145, 146, 147, 150, 151, 153, 154, 156, 157, 158, 159, 160, 161, 162, 164, 165, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 185, 186, 187, 188, 194, 196, 200, 201, 205, 207, 208, 210, 211, 212, 213, 214, 215
- foreign 12, 209
- for loop 151
- formula 92, 190

- from 12, 13, 18, 19, 21, 22, 25, 27, 28, 38, 42, 43, 44, 51, 64, 72, 81, 82, 97, 98, 99, 101, 102, 106, 111, 117, 119, 120, 121, 124, 126, 135, 138, 140, 142, 146, 150, 156, 158, 163, 167, 172, 173, 175, 177, 180, 182, 187, 199, 200, 204, 212
- F test 177, 193
- G**
- Gammad 136
- Gamma distribution 99, 125
- gamma function 125
- Gaussian distribution 119
- geom_bar 33
- geom_boxplot 92
- geom_density 69, 148
- geometric distribution 98, 112, 126
- geometric mean 18, 54, 55, 59
- Geometric standard deviation 59
- geom_histogram 69, 70, 148
- ggplot 33, 69, 72, 148, 204
- ggplot2 32, 33, 41, 44, 51, 66, 69, 72, 78, 85, 92, 94, 96, 97, 148, 209, 214
- ggttitle 33, 69
- Gmean 55
- gold standard 180
- grammar of graphics 32
- graphics 10, 12, 32, 86, 88, 89, 92, 209, 214
- graphic systems 32
- grDevices 12, 88, 209
- grid 12, 89, 121, 209
- Gsd 59
- H**
- handling colors 79, 80
- hexadecimal code 84
- hist 69, 70, 97
- histogram 18, 65, 66, 67, 69, 70, 78, 93, 97, 147, 148, 175, 176
- Hypergeometric distribution 98, 106
- hypothesis 45, 177, 178, 179, 182, 184, 187, 190, 199
- one-sided 45, 177, 178, 179, 182, 184, 187, 190, 199
- two-sided 45, 177, 178, 179, 182, 184, 187, 190, 199

**“I studied English for 16 years but...
...I finally learned to speak it in just six lessons”**

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



I

ICU 25, 28, 29, 37, 44, 50, 53, 55, 59, 63, 66, 78, 97, 142, 143, 152, 154, 159, 165, 167, 169, 170, 171, 175, 176, 187, 188, 191, 196, 201, 207, 208
 ICUData 18, 25, 26, 27
 ICUData.csv 18, 25
 ICU dataset 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 bilirubin 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 Description of variables 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 heart rate 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 import 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 liver failure 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 LOS 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 outcome 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 SAPS II 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 sex 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 surgery 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 temperature 25, 29, 37, 44, 53, 55, 59, 78, 97, 142, 175, 187, 196, 207
 incidence 114
 incidence rate 114
 inferential statistics 18, 19, 56, 133, 135, 138, 139, 179
 goal 18, 19, 56, 133, 135, 138, 139, 179
 integer 27, 28, 35, 110
 interquartile range 18, 37
 Interval estimator 138
 IQR 18, 37, 40, 41

J

jpeg 88

K

Kendall's τ 18, 50, 52, 75
 KernSmooth 12
 knitr 9, 209, 215
 Kolmogorov(-Smirnov) distance 154
 kruskal.test 201, 202
 Kruskal-Wallis test 177, 201, 202, 203, 207, 208
 KS-MD estimator 138, 154, 176
 Kurt 62, 63
 kurtosis 62, 63, 78
 normal distribution 62, 63, 78
L
 lattice 12
 left-skewed 60, 61
 legend 49, 95
 legend.text = TRUE 49
 level 21, 28, 30, 44, 158, 172, 181, 182, 188, 203
 library 32, 33, 208
 likelihood function 144, 145
 LillieTest 206
 lines 26, 69, 148
 load 25, 33, 80, 105, 147, 165
 location and scale model 147
 log-likelihood function 144, 145
 Log-normal distribution 98, 123
 LOS 28
 lower.tail = FALSE 104, 108, 116
 lwd 148
M
 mad 38
 MAD 18, 38, 57, 58, 62, 78, 138, 152, 153, 168, 169, 175
 confidence interval 18, 38, 57, 58, 62, 78, 138, 152, 153, 168, 169, 175
 consistent 18, 38, 57, 58, 62, 78, 138, 152, 153, 168, 169, 175
 standardization 18, 38, 57, 58, 62, 78, 138, 152, 153, 168, 169, 175
 main 20, 31, 88

- MASS 12, 145, 146, 161, 208, 209
 Matrix 12
 maximum likelihood estimator 144
 MD estimator 138, 153, 154, 170, 171, 175, 176
 confidence interval 138, 153, 154, 170, 171, 175,
 176
 consistent 138, 153, 154, 170, 171, 175, 176
 Cramér-von-Mises 138, 153, 154, 170, 171, 175,
 176
 Kolmogorov(-Smirnov), 138, 153, 154, 170, 171,
 175, 176
 MDEstimator 154, 170
 mean 18, 52, 53, 54, 55, 57, 58, 59, 60, 61, 62, 66, 68,
 78, 99, 119, 120, 121, 123, 135, 138, 142, 145,
 146, 152, 158, 159, 160, 162, 170, 175, 182, 188,
 190, 193, 205, 208
 meanlog 123
 median 18, 35, 37, 38, 40, 41, 42, 53, 54, 57, 58, 60, 61,
 62, 78, 114, 127, 133, 136, 137, 138, 152, 153,
 168, 169, 175
 confidence interval 18, 35, 37, 38, 40, 41, 42, 53,
 54, 57, 58, 60, 61, 62, 78, 114, 127, 133, 136,
 137, 138, 152, 153, 168, 169, 175
 consistent 18, 35, 37, 38, 40, 41, 42, 53, 54, 57,
 58, 60, 61, 62, 78, 114, 127, 133, 136, 137,
 138, 152, 153, 168, 169, 175
 median absolute deviation 38
 medianCI 169
 methods 12, 21, 209, 212, 215
 mfrow 151
 mgcv 12
 Minimum-distance estimator 153
 MKmisc 165, 169, 209, 212
 ML estimator 138, 144, 145, 146, 154, 155, 156, 163,
 164, 171, 172, 175, 176
 Bernoulli model 138, 144, 145, 146, 154, 155,
 156, 163, 164, 171, 172, 175, 176
 Exponential model 138, 144, 145, 146, 154, 155,
 156, 163, 164, 171, 172, 175, 176
 normal distribution model 138, 144, 145, 146,
 154, 155, 156, 163, 164, 171, 172, 175, 176
 Poisson model 138, 144, 145, 146, 154, 155, 156,
 163, 164, 171, 172, 175, 176
 MLEstimator 147, 162, 167
 Mode 18
 mu = 188
N
 n 56, 98, 99, 106, 107, 108, 115, 121, 126, 133, 134, 135,
 157, 163, 172, 178, 181, 185, 199
 Negative binomial distribution 98, 111
 nlme 12
 nnet 12
 Normal distribution 98, 119, 145
 NormLocationScaleFamily 147
 nrow 29
 Null hypothesis 178, 179
 NUMERIC 209
O
 one-sample binomial test 196
 asymptotic 196
 exact 196
 one-way ANOVA 201, 202, 203, 207, 208
 Welch 201, 202, 203, 207, 208
 oneway.test 201
 outlier 202
 Kendall's τ 18, 50, 52, 75
 median 202
 quantile 202
 Spearman's ρ 49, 50, 52, 74, 75
P
 package „animation“ 89
 pairwise.t.test 203
 pairwise.wilcox.test 203
 par 151
 parallel 12
 parametric family 140, 141, 168
 parametric model 140, 153, 201
 Pascal distribution 98, 112
 pbinom 103
 pdf 86, 87, 88, 97, 211, 213
 pdfLATEX 9

- Pearson correlation 73, 74, 75
 Pearson's contingency coefficient 18, 47
 percentile 36
 PercTable 46
 pexp 127
 pfmt 46
 pgamma 127
 Phi 47
 phyper 108
 pie 18, 34, 83, 89, 90
 pie chart 83, 89
 drawbacks 83, 89
 plnorm 123
 plot 18, 31, 32, 33, 34, 40, 41, 43, 44, 48, 50, 63, 66, 67,
 69, 72, 75, 77, 78, 83, 87, 89, 92, 95, 97, 105, 120,
 121, 124, 127, 128, 130, 132, 138, 147, 148, 150,
 151, 173, 175, 176, 204, 208
 pnbinom 112
 png 88, 89, 97
 pnorm 120
 Point Estimation 140
 point estimator 140, 141, 157, 158
 points 38, 41, 44, 72, 95, 114, 118, 121, 150, 187
 Poisson distribution 98, 114, 115, 126, 136
 Pólya distribution 98, 112
 population 18, 19, 20, 22, 56, 66, 98, 101, 106, 107, 108,
 109, 122, 142, 143, 165, 172, 175, 182, 211
 Portable document format 88
 Portable network graphics 88
 Post hoc tests 177
 postscript 88
 power. 186, 187
 power.prop.test 186
 power.t.test 183, 186
 ppois 115, 116
 Probability 98, 135, 179
 probability density 118
 probability mass function 99, 100, 102, 103, 105, 107,
 111, 115, 135, 144
 probability model 73, 138, 142, 143, 144, 147, 153, 157,
 159, 164
 probability theory 18, 19, 98, 119, 135, 136



What do you want to do?

No matter what you want out of your future career, an employer with a broad range of operations in a load of countries will always be the ticket. Working within the Volvo Group means more than 100,000 friends and colleagues in more than 185 countries all over the world. We offer graduates great career opportunities – check out the Career section at our web site www.volvologistics.com. We look forward to getting to know you!

VOLVO
 AB Volvo (publ)
www.volvologistics.com

VOLVO TRUCKS | RENAULT TRUCKS | MACK TRUCKS | VOLVO BUSES | VOLVO CONSTRUCTION EQUIPMENT | VOLVO PENTA | VOLVO AERO | VOLVO IT
 VOLVO FINANCIAL SERVICES | VOLVO 3P | VOLVO POWERTRAIN | VOLVO PARTS | VOLVO TECHNOLOGY | VOLVO LOGISTICS | BUSINESS AREA ASIA

Download free eBooks at bookboon.com

prob.table 45
prop.table 48
prop.test 186, 196
ps 88

Q

qbinom 103
qexp 127
qgamma 127
qhyper 108
qlnorm 123
qnbinom 112
qnorm 120
qplot 41, 44, 66, 72
qpois 115
qqline 150, 151, 208
qnorm 150, 151, 208
qq plot 138, 150, 151, 175, 176, 208
qqplot 150
quantile function 99, 103, 104, 105, 112, 116, 118
Quantile-quantile plot 138
quartile 18, 37, 41, 42, 58, 78, 175
quartile coefficient of dispersion 18, 58

R

random numbers 103, 105, 185
random sample 19, 109
random variable 99, 100, 101, 102, 105, 106, 107, 111, 114, 115, 117, 118, 119, 123, 125, 130, 133, 134, 135, 140, 141
rank 49, 50, 52, 73, 77, 156, 177, 191, 201, 202
rank correlation 50
rate 28, 29, 74, 77, 78, 88, 96, 97, 114, 126, 127, 129, 130, 136, 199, 208
rbinom 103, 105
RColorBrewer 80, 83, 97, 209, 212
R Consortium 12, 214
R Core Development Team 12
read.* 23, 24
read.csv 22
read.csv2 22, 26
read.delim 22
read.delim2 22

read.table 22
recommended packages 12, 17, 55, 146
relative frequency 65, 135, 139, 142, 145, 164, 166
approximative confidence interval 65, 135, 139, 142, 145, 164, 166
confidence interval 65, 135, 139, 142, 145, 164, 166
cross table 65, 135, 139, 142, 145, 164, 166
efficient 65, 135, 139, 142, 145, 164, 166
exact confidence interval 65, 135, 139, 142, 145, 164, 166
unbiased 65, 135, 139, 142, 145, 164, 166

rep 93
representative sample 19, 140, 141, 176
rev 93
rexp 127
rgamma 127
rgb 84
RGB code 84
rhyper 108
right-skewed 54, 60, 62, 67
right = TRUE 70
R Installation 13, 213
 Linux 13, 213
 Mac OS X 13, 213
 Windows 13, 213

rlnorm 123
rnbinom 112
rnorm 120, 151, 185
round 45, 46, 53, 70, 183
rpart 12
rpois 115
R script 16, 17, 18, 23, 25, 31, 79, 99, 138, 177
 new 16, 17, 18, 23, 25, 31, 79, 99, 138, 177
 open 16, 17, 18, 23, 25, 31, 79, 99, 138, 177

RStudio 10, 14, 15, 16, 17, 18, 23, 24, 25, 26, 27, 28, 31, 32, 46, 80, 86, 87, 105, 147, 165
interactive help 10, 14, 15, 16, 17, 18, 23, 24, 25, 26, 27, 28, 31, 32, 46, 80, 86, 87, 105, 147, 165
panes 10, 14, 15, 16, 17, 18, 23, 24, 25, 26, 27, 28, 31, 32, 46, 80, 86, 87, 105, 147, 165

window Environment 10, 14, 15, 16, 17, 18, 23, 24, 25, 26, 27, 28, 31, 32, 46, 80, 86, 87, 105, 147, 165
 window Help 10, 14, 15, 16, 17, 18, 23, 24, 25, 26, 27, 28, 31, 32, 46, 80, 86, 87, 105, 147, 165
 window History 10, 14, 15, 16, 17, 18, 23, 24, 25, 26, 27, 28, 31, 32, 46, 80, 86, 87, 105, 147, 165
 window Packages 10, 14, 15, 16, 17, 18, 23, 24, 25, 26, 27, 28, 31, 32, 46, 80, 86, 87, 105, 147, 165
 RStudio IDE 15, 16

S

S 10, 11, 210, 211, 214
 sample size calculation 139, 181
 SAPS II 37, 38, 40, 41, 43, 50, 51, 52, 92, 208, 215
 save 9, 18, 25, 87
 save.image 25
 Scalable vector graphics 88
 scale_colour_manual 96
 scale of measurement 21
 interval scaled 21
 nominal 21
 ordinal 21
 ratio scaled 21
 scan 22
 scatter diagram 75, 96, 97
 scatter plot 18, 50, 77
 sd 56, 120, 146
 sdlog 123
 seq 36
 shape measure 59, 62
 ShapiroFranciaTest 206
 shapiro.test 206
 show.details = „minimal“ 155
 Six Sigma 120
 size = 1 167
 Skew 60, 61
 skewness 59, 60, 61, 62, 63, 78
 spatial 12
 Spearman's ρ 49, 50, 52, 74, 75
 splines 12
 standard deviation 18, 38, 55, 56, 57, 59, 60, 62, 78, 100, 119, 120, 121, 123, 136, 138, 152, 156, 157, 162, 163, 168, 169, 170, 172, 173, 175
 standardization 18, 38, 55, 56, 57, 59, 60, 62, 78, 100, 119, 120, 121, 123, 136, 138, 152, 156, 157, 162, 163, 168, 169, 170, 172, 173, 175
 standard error 158
 standard normal distribution 120, 158, 161, 163, 164
 stat_function 148
 statistical programming language S 10
 statistical test 139, 181, 182
 acceptance region 139, 181, 182
 Ansari-Bradley test 139, 181, 182
 correlation test 139, 181, 182
 distribution test 139, 181, 182
 extremely significant 139, 181, 182
 Fisher's exact test 139, 181, 182
 F test, 139, 181, 182
 highly significant 139, 181, 182
 Kruskal-Wallis test 139, 181, 182
 one-way ANOVA 139, 181, 182
 post hoc 139, 181, 182
 power 139, 181, 182
 rejection region 139, 181, 182
 relevance 139, 181, 182
 sample size calculation 139, 181, 182
 sensitivity 139, 181, 182
 significant 139, 181, 182
 specificity 139, 181, 182
 steps 139, 181, 182
 test of normality 139, 181, 182
 t test 139, 181, 182
 type I error 139, 181, 182
 type II error 139, 181, 182
 very significant 139, 181, 182
 Wilcoxon-Mann-Whitney U test 139, 181, 182
 Wilcoxon signed rank test 139, 181, 182
 stats 12, 209
 stats4 12, 145, 209
 str 26
 survival 12, 126, 137
 svg 88

T

table 18, 22, 24, 29, 44, 45, 46, 48, 196, 208
 Tagged image file format 88
 tcltk 12
 t distribution 99, 133, 158, 163, 182, 184, 199
 tests of normality 205, 208
 Cramér-von Mises test 205, 208
 Lilliefors (Kolmogorov-Smirnov) test 205, 208
 Shapiro-Francia test 205, 208
 Shapiro-Wilk test 205, 208
 tiff 88
 to 9, 10, 12, 13, 14, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26,
 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41,
 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54,
 55, 57, 58, 59, 60, 62, 63, 65, 66, 67, 69, 70, 73,
 74, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88,
 89, 90, 91, 92, 94, 95, 96, 97, 98, 99, 100, 101,
 102, 103, 105, 106, 108, 111, 112, 113, 114, 115,
 116, 118, 120, 121, 123, 124, 126, 127, 128, 129,
 130, 132, 136, 137, 138, 139, 140, 141, 142, 143,
 144, 145, 146, 147, 148, 150, 151, 152, 153, 154,

155, 157, 158, 159, 164, 165, 167, 168, 169, 170,
 171, 172, 174, 175, 176, 177, 178, 179, 181, 182,
 183, 184, 185, 187, 188, 190, 191, 193, 194, 196,
 199, 200, 201, 202, 203, 204, 205, 206, 207, 208,
 212, 214

tools 12, 209

t test 177, 182, 187, 188, 190, 191, 192, 193, 203, 207,
 208, 213

one-sample 177, 182, 187, 188, 190, 191, 192,
 193, 203, 207, 208, 213

paired 177, 182, 187, 188, 190, 191, 192, 193,
 203, 207, 208, 213

pairwise 177, 182, 187, 188, 190, 191, 192, 193,
 203, 207, 208, 213

two-sample 177, 182, 187, 188, 190, 191, 192,
 193, 203, 207, 208, 213

Welch 177, 182, 187, 188, 190, 191, 192, 193,
 203, 207, 208, 213

t.test 160, 183, 185, 186, 187, 188, 203

Types of attributes 18, 21

U

universe 18
utils 12, 209

V

var 56, 187, 193
var.equal = TRUE 187
variable 21, 24, 27, 28, 29, 53, 65, 78, 95, 97, 99, 100, 101, 102, 105, 106, 107, 111, 114, 115, 117, 118, 119, 123, 125, 130, 133, 134, 135, 140, 141
metric 21, 24, 27, 28, 29, 53, 65, 78, 95, 97, 99, 100, 101, 102, 105, 106, 107, 111, 114, 115, 117, 118, 119, 123, 125, 130, 133, 134, 135, 140, 141

variable names 29

variance 55, 56, 62, 99, 100, 102, 108, 112, 115, 118, 120, 123, 126, 130, 135, 136, 142, 145, 146, 147, 158, 182, 183, 184, 187, 194
confidence interval 55, 56, 62, 99, 100, 102, 108, 112, 115, 118, 120, 123, 126, 130, 135, 136, 142, 145, 146, 147, 158, 182, 183, 184, 187, 194
standardization 55, 56, 62, 99, 100, 102, 108, 112, 115, 118, 120, 123, 126, 130, 135, 136, 142, 145, 146, 147, 158, 182, 183, 184, 187, 194
unbiased 55, 56, 62, 99, 100, 102, 108, 112, 115, 118, 120, 123, 126, 130, 135, 136, 142, 145, 146, 147, 158, 182, 183, 184, 187, 194

var.test 193

view 26, 65, 123, 179, 207

W

waiting time distribution 112
Weibull distribution 99, 129, 130, 132, 137, 176
Wilcoxon-Mann-Whitney U test 177, 191, 207
pairwise 177, 191, 207
Wilcoxon signed rank test 177, 191
wilcox.test 191, 192, 203
working directory 15, 25, 26, 89

change 15, 25, 26, 89
check 15, 25, 26, 89

write.csv 24

write.csv2 24

write.table 24

X

xlab 69
xlim 41, 65

Y

ylab 31, 33, 69

Z

z-transformation 60